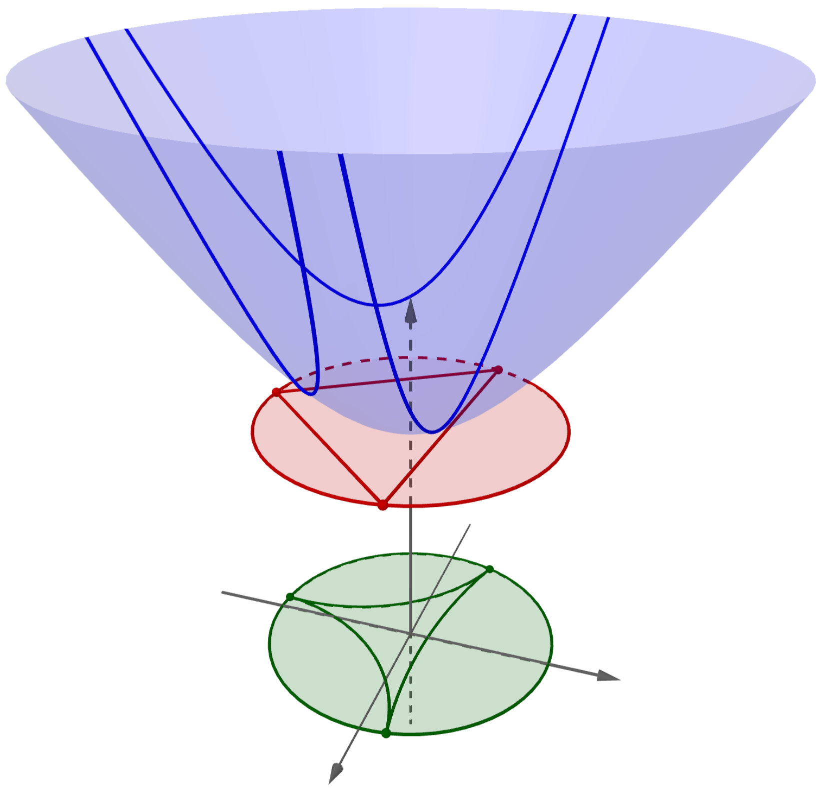
Hyperbolic geometry

Brice Loustau



A book, even if it is written with complete honesty, is always worthless from one *standpoint: for no one really needs to write a book, since there are many other things to do in the world.*

– Ludwig Wittgenstein[1]

[1] Letter to Ludwig von Ficker, 4 December 1919. English translation in [Luc96, pp. 82–98].

DISCLAIMER

This document is a **preliminary** version of a textbook to be published in 2022.

Several chapters have yet to be written, and many others are still **drafts**.
Please reserve your judgement until the final version!

Brice Loustau
19 September 2021

This document was compiled with pdfLaTeX on 19/09/2021 at 21:17

todo

This page summarizes the remaining tasks before the book is ready.

Remaining chapters to write or complete:

- Chapter 6
- Chapter 15, Chapter 16
- Chapter 17, Chapter 18
- Appendix A
- Exercises hints and solutions
- Index of notations, Index of terminology

Remaining chapters to rewrite or review:

- Chapter 1, Chapter 2
- Chapter 4, Chapter 5
- Chapter 8
- Chapter 9, Chapter 10
- Chapter 11, Chapter 12
- Chapter 14

Remaining chapters to proofread:

- Preface
- Chapter 3
- Chapter 7
- Chapter 13
- References

Details: ./Source/TODO.tex

The estimated end date is: **29/10/2022**

Preface

Hyperbolic geometry is a very special subject: it is the star of geometries, and geometry is the star of mathematics! Well, perhaps this is a bit of an exaggeration, yet a useful one to have in mind—few topics have such historical and conceptual weight.

The history of mathematics and science, indeed, speaks for the importance of hyperbolic geometry. The names of several of the greatest mathematicians are attached to it, such as Gauss and Poincaré, and its incredibly fertile development is related to the (re)birth of projective spaces, Fuchsian groups, Minkowski spacetime, among other decisive notions for modern mathematics and physics, including the mathematical framework for Einstein's theory of relativity!

While the revolutionary discovery of hyperbolic geometry mainly took place in the 19th century, it continued to play a leading role in the mathematics of the 20th, culminating with Thurston's geometrization program and its completion in the early 21st century by Perelman, which solved the famous Poincaré conjecture. To this day, hyperbolic geometry and its avatars remain an intensely active field of research, both in mathematics and in applied sciences—it shows promise, for instance, in the emerging field of data science and machine learning.

* * *

Why did I write this textbook? It started as a set of lecture notes that I wrote for a Master course I taught at TU Darmstadt in the winter 2019–2020, building on notes for a similar course I held at Rutgers University in 2017. After a push of some colleagues and friends, I decided to upgrade them into a proper book. I hope that it fills a gap in the literature, as an ambitious first course on hyperbolic geometry—more details below.

Goal and intended audience. The goal of the book is to provide a first course on hyperbolic geometry with little or no prerequisites of differential geometry. It intends to be fairly thorough while staying self-contained and not too advanced. The book is suitable for a course at the early graduate (Master) level or advanced undergraduate (Bachelor) level in a mathematics curriculum. More broadly, it is meant to be useful to anyone looking to properly learn the basics (and more) of hyperbolic geometry, whether to pursue higher level education or research in a related area, or to apply it to other fields.

Contents. All the standard features of hyperbolic spaces are rigorously introduced and studied: the different models of any dimension (the hyperboloid model, Cayley–Klein and Beltrami–Klein models, Poincaré ball and half-space models); hyperbolic geodesics, distance, curvature, and isometries; the ideal boundary and the classification of isometries; hyperbolic triangles and trigonometry; tessellations of the hyperbolic plane, and more. Beyond hyperbolic geometry, readers will have the opportunity to learn many essential notions of geometry: the concept of curvature, Minkowski space and the Lorentz group, projective geometry and quadrics, Möbius transformations and conformal geometry, metric geometry and Gromov hyperbolicity… In addition, the book features a couple of "bonus chapters": on Einstein's theory of relativity and its connection to hyperbolic geometry, and on the applications of hyperbolic geometry to data science and machine learning.

Approach and style. The approach aims to be clean and rigorous, using the framework and style of modern mathematics, although historical aspects are occasionally mentioned. (For other subjects usually covered in mathematics textbooks, this would go without saying, but hyperbolic geometry is special given the historical weight of the "synthetic" approach.) After the first two introductory chapters, the book develops the formal concepts that allow the most effective definitions of hyperbolic spaces, such as pseudo-Euclidean spaces and projective spaces. Readers are therefore expected to be able to handle a certain level of abstraction.

Why learn hyperbolic geometry? I see at least three excellent reasons for students in mathematical sciences to learn hyperbolic geometry: (1) Since the 19th century, non-Euclidean geometry has become a standard framework in mathematics and physics. Hyperbolic geometry is the star of non-Euclidean geometries, and gives fundamental insight on all phenomena related to negative curvature. (2) A course in hyperbolic geometry is a great opportunity to learn a diversity of classical geometric notions that are useful across many areas of mathematics. In this book, you will (re-)discover bits of Riemannian geometry, relativity theory, real and complex projective geometry, and more. (3) In contemporary mathematical research, hyperbolic geometry is at the intersection of several important fields: geometry and topology, group theory, complex geometry, and others. I refer to [Can+97, §15] for a discussion of this. (4) Bonus reason! Hyperbolic geometry shows promising possibilities for data science and machine learning: this is the content of Part VII. It could therefore appeal to students who aspire to be data scientists or engineers.

What does this book not cover? This remains to be written.

Prerequisites. No prerequisites are assumed beyond a (solid) standard undergraduate curriculum in mathematics, including linear algebra and multivariable calculus. Students who have an additional background in geometry (such as differential geometry of curves and surfaces, Riemannian geometry, projective geometry, etc.) will nevertheless be able to put their prior knowledge to excellent use. In contrast to the course that I taught at Rutgers University, I elected not to include a mini-course on Riemannian geometry beyond the introductory Chapter 2. As a consequence, the book focuses more on classical geometric aspects than differential ones, despite my personal inclination for the latter.

Exercises. Each chapter is concluded with a list of exercises. Some solutions and hints are included at the end of the book, but I recommend resisting the temptation to look at them as long as possible. It is essential that students really work on the exercises. No serious mathematics can truly be learned without asking and answering many questions and solving problems. Spending this time and effort cannot be spared, but it makes it more fun and often more effective to work with other students or a teacher.

Other references. This remains to be written.

Historical disclaimer. This remains to be written.

How did I write this book? This book was written with the LaTeX typesetting system. Technically, I run a `TeX Live` installation on the operating system `Debian GNU/Linux`, and I use the editor `Kile`. To create figures, I have worked with the software `GeoGebra`, the vector graphics editor `Inkscape`, the `Python` package `matplotlib`, and the LaTeX package `Tikz`. To keep track of files, I use the version control system `Git`. (There is nothing original about these choices!) Most if not all the sofware that I use is *free* and open-source, not only because I support it philosophically, but also because it is often the best. I am indebted to Donald Knuth, Richard Stallman, Linus Torvalds, and all other free-spirited enthusiasts and talented programmers who have made and continue to make free sofware a reality.

Acknowledgments. This remains to be written.

To all readers: learn and enjoy; write me! I sincerely hope that you find this book instructive and that you find joy and beauty in learning hyperbolic geometry. I absolutely appreciate all feedback: please contact me with any mathematical or non-mathematical questions, reports of mistakes and typos, suggestions for improvement, criticism, and more. The best way to contact me is by e-mail: `brice@loustau.eu`. Thank you!

Brice Loustau
Heidelberg, September 2021

Contents

Part I

Preliminaries

You must not attempt this approach to parallels. I know this way to its very end. I have traversed this bottomless night, which extinguished all light and joy of my life. I entreat you, leave the science of the parallels alone…I thought I would sacrifice myself for the sake of truth. I was ready to become a martyr who would remove the flaw from geometry and return it purified to mankind. I accomplished monstrous, enormous labors […] I have traveled past all reefs of this infernal Dead Sea and have always come back with broken mast and torn sail. The ruin of my disposition and my fall date back to this time.

– Farkas Bolyai to his son János in 1820, on Euclid's parallel postulate[2]

I have discovered such wonderful things that I was amazed, and it would be an everlasting piece of bad fortune if they were lost. When you, my dear Father, see them, you will understand; at present I can say nothing except this: that out of nothing I have created a strange new universe.

– János Bolyai's response to his father in 1823[3]

[2]Quoted from [Gra04].
[3]Quoted from [Gra04].

CHAPTER 1

From Euclid to hyperbolic geometry

Disclaimer: This chapter is a draft.

In this first preliminary chapter, we propose a brief introduction to hyperbolic geometry that puts the emphasis on the axiomatic approach going back to Euclid's *Elements*.

Note that in the remainder of the book, we will neglect the historical (and the "synthetic") approach to hyperbolic geometry in favor of a modern and effective mathematical treatment. It will therefore be harmless to forget the contents of this chapter for the most part, although we will sporadically refer back to it for insight.

* * *

The discovery of non-Euclidean geometries in the 19th century was one of the most significant developments in the history of mathematics and had a profound impact on science and philosophy. This is well put by Marvin Greenberg [Gre93]:

> *Most people are unaware that in the early nineteenth century a revolution took place in the field of geometry that was as scientifically profound as the Copernican revolution in astronomy and, in its impact, as philosophically important as the Darwinian theory of evolution*[1]. *"The effect of the discovery of hyperbolic geometry on our ideas of truth and reality has been so profound," wrote the great Canadian geometer H. S. M. Coxeter, "that we can hardly imagine how shocking the possibility of a geometry different from Euclid's must have seemed in 1820." Today, however, we have all heard of the space-time geometry in Einstein's theory of relativity. [...]*

[1]While these parallels are compelling, they should perhaps not be taken too literally. The Copernican revolution (1543), a part of the "Scientific Renaissance", was still a classical—arguably "Euclidean"—theory. On the other hand, hyperbolic geometry bloomed in the romantic 19th century, which broke classical rules in art and science. A comparison with Einstein's theory of relativity would seem more pertinent to me.

There are many excellent references discussing the fascinating discovery of non-Euclidean geometry and its historical, mathematical, and philosophical implications. Although I am not an expert, I can confidently recommend the monographs of Harold Coxeter [Cox98], Jeremy Gray [Gra11], Marvin Greenberg [Gre93], John Milnor [Mil82], Boris Rosenfeld [Ros88], John Stillwell [Sti96], and Richard Trudeau [Tru08]. For those who can read French, I recommend the phenomenal book of Jean-Daniel Voelke [Voe05]; let me also mention the essays of Jean Dieudonné [Die96] and Henri Poincaré [Poi02, Chap. 3].

The goal of this chapter is of course not to compete with these books. Instead, we will content with just enough background to convey a decent sense of the "origin story" of hyperbolic geometry, without attempting to be historically precise or thorough. That said, I absolutely encourage readers to explore the above references and more; the subject is fascinating and great insight is to be gained from learning the history of geometry.

1.1 Euclid's postulates

The long history leading up to the discovery of hyperbolic geometry originates in the *Elements* of Euclid[2]. This treatise of mathematics divided in 13 books was written in ca. 300 BC by the mathematician Euclid of Alexandria[3]. The *Elements* were the used as the main reference for geometry and more generally for mathematics until the 19th century. For this reason, it can hardly be disputed that it is the most important work of mathematics ever written.

Euclid's method is axiomatic and constructive. This approach is far from outdated, on the contrary: the foundation of contemporary mathematics, as it has been formalized since the first half of the 20th century with mathematical logic, is strikingly similar to Euclid's Elements. We discuss this more in § 1.3.

Euclid introduces 5 ***postulates*** (axioms):

(E1) There exists a line through any two points.
(E2) Any line may be extended indefinitely.
(E3) Any center and radius determines a unique circle.
(E4) All right angles are congruent. (See Figure 1.1.)
(E5) If a straight line falling on two straight lines make the interior angles on the same side less than two right angles, the two straight lines, if produced indefinitely, meet on that side on which the angles are less than two right angles. (See Figure 1.2.)

Remark 1.1. By ***line***, Euclid means *straight line segment*. He does not directly consider infinite lines, which is a very reasonable position. By two lines being ***parallel***, one must understand: they do not intersect, even when extended indefinitely.

Remark 1.2. Note that Euclid does not state uniqueness in the first postulate (E1). In particular, it does not exclude spherical geometry.

Remark 1.3. The fourth postulate (E4) may be interpreted as follows: For any two configurations of two straight lines intersecting at a right angle, there exists a *rigid motion* (i.e. an orientation-preserving isometry) which takes the first configuration to the second: see Figure 1.1.

Based on these five postulates (and five "common notions"), Euclid develops an extensive treatise of mathematics (geometry and number theory). It is divided in thirteen books, consisting of a collection of definitions, constructions, theorems, and proofs. For instance, the first book contains the Pythagorean theorem and the sum of the angles in a triangle, as well as many other constructions of plane geometry.

[2] For a translation of the *Elements* in English, Thomas L. Heath's 1908 translation is the main reference. A second edition was published by Dover in 1956 [Euc56].

[3] Despite his epithet, Euclid was Greek, like his predecessors Thales and Pythagoras; he lived in Egypt under the reign of Ptolemy I, shortly after it was conquered by Alexander the Great. That being said, the contribution of the ancient Egyptians to Greek mathematics is often underestimated, as is shown in the wonderful article [Her18], which I recommend reading for a taste of a serious historical approach!

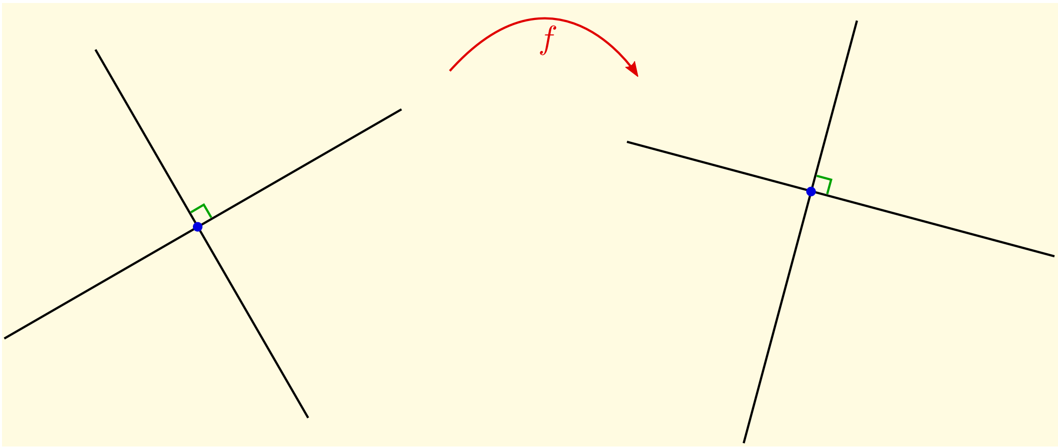


Figure 1.1: Euclid's fourth postulate: Any two right angles are congruent.

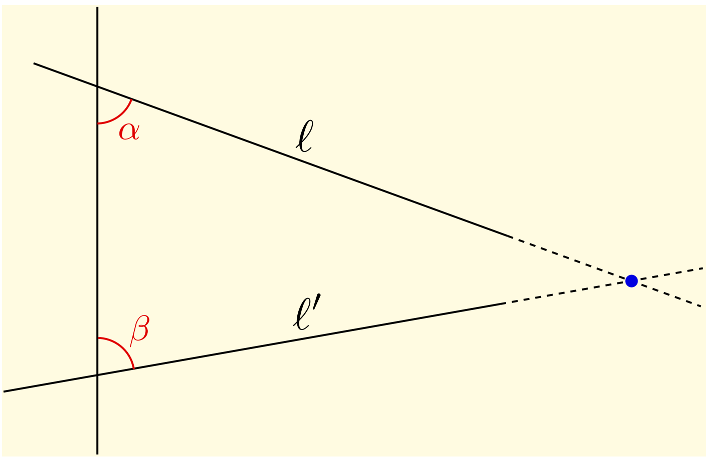


Figure 1.2: Euclid's fifth postulate: If $\alpha + \beta < \pi$, then l and l' intersect on the side of α and β.

1.2 Discovery of non-Euclidean geometry

For centuries, mathematicians have questioned the fifth postulate, often called the ***Parallel postulate***. See Figure 1.2 for an illustration. This postulate sounds more convoluted than the first four. Could it not simply be derived from them?

Today, the parallel postulate often stated into the equivalent form:

(E5') Given a line and a point not on it, there exists a unique parallel through the point.

There are many other equivalent formulations of the fifth postulate, such as: the sum of the angles of any triangle is equal to two right angles.

Until the 19th century, mathematicians were unable to prove whether the fifth postulate was a consequence of the first four or not. The 19th century was the century of romanticism, which decided that classical rules should be broken. A breakthrough was achieved by Lobachevsky (and Gauss, Bolyai, Taurinus, Cayley, and others)[4]. He constructed a complete alternative to Euclidean geometry, starting from the assumption that the first four postulates are true, but the fifth postulate is false. Initially, the goal of this strategy was to reach a contradiction, which would prove that the fifth postulate does derive from the first four. However, it eventually became clear that this new geometry was as respectable and beautiful as Euclid's.

Remark 1.4. Spherical geometry also offers an alternative to non-Euclidean geometry. This is the geometry on a sphere, where straight lines are arcs of great circles. Note that it does not satisfy the first axiom of Euclid if we add uniqueness to straight lines through two points: consider antipodal points. In fact, antipodal points are especially problematic because there is an infinity of straight lines between them. One can remedy this issue by identifying any two antipodal points. The resulting surface is known as the real projective plane, equipped with the geometry inherited from the sphere. This geometry is called ***elliptic geometry***, and is the only other non-Euclidean geometry besides hyperbolic geometry. The fifth postulate for elliptic geometry reads: Any two lines intersect (i.e., there are no parallels). This case must be excluded to obtain hyperbolic geometry, therefore the fifth postulate for hyperbolic geometry reads:

(H5) Given a line and a point not on it, there exists at least two parallels through the point.

As an example, Lobachevsky developed the notion of ***angle of parallelism***: given a line l and a point A at distance a from l, the angle of parallelism α is the least angle such that the line l' as in Figure 1.3 is parallel to l (i.e. does not intersect l).

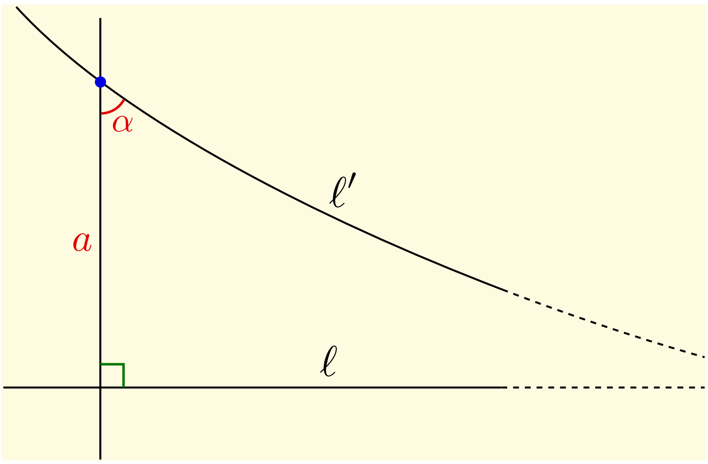


Figure 1.3: Angle of parallelism.

[4]Hyperbolic geometry is still occasionally called ***Lobachevsky geometry***. Lobachevsky was the first to publish his extensive work in 1829, but by then other mathematicians had also discovered in part this new geometry: Gauss, Schweikart and Taurinus, Farkas and Janos Bolyai.

It is important to note that Lobachevsky, despite writing a considerable treatise of hyperbolic geometry *à la Euclid*, did not answer the question of whether Euclid's fifth postulate is independent of the first four. Indeed, it was still possible that Lobachevsky's geometry was inconsistent, and that he simply did not yet find a contradiction. The same can be said of the work of Gauss, Taurinus, and Bolyai.

The question was definitively settled by Beltrami in 1868 [Bel68a; Bel68b], who found a *model*—in fact several models—for the hyperbolic plane, in other words a "universe" where the axioms of hyperbolic geometry are satisfied.

The first model proposed by Beltrami is now known as the Beltrami–Klein model (or Cayley–Klein model, or simply Klein model[5]). We will study it in detail in Chapter 8, but its description is surprisingly simple: the hyperbolic plane is an open Euclidean disk; points in this model are points inside the disk, and lines are chords, i.e. straight line segments with (imaginary) endpoints on the boundary circle. See Figure 1.4. However, angles and distances are not as they appear to our Euclidean eyes. In particular, until we define angles, distances, and isometries, we cannot verify Euclid's axioms 3. and 4.

Beltrami also proposed a second pair of models, which are now known as the Poincaré disk and half-space models[6]. The disk model is again an open Euclidean disk, but this time lines are defined as circles of arcs that are orthogonal to the boundary circle. See Figure 1.5. Distances are also distorted in this model with respect to our Euclidean eyes, but not angles: it is a ***conformal*** model. We will study this model in Chapter 10.

1.3 Notions of mathematical logic

Let us reconsider the previous historical discussion in the eyes of mathematical logic. I warn the reader that what follows is a naive presentation: Euclid's system does not actually meet the requirements of a theory as it is defined by first-order logic. The axiomatic foundation of geometry has generated considerable work since the late 19th century; notable modern axiomatizations of Euclid's theory were proposed by Hilbert (1899), Birkhoff (1932), Tarski (1959).

A mathematical theory is based on a syntax, axioms, and rules of inference, from which theorems are derived (also cosmetically called lemmas, propositions, corollaries, etc). The majority of contemporary mathematics implicitly uses the theory of sets of Zermelo–Fraenkel, but other setups can also be relevant. Regardless, Euclid's treatise and its axiomatic approach appears singularly modern.

Ideally, the axioms that one chooses to base a mathematical theory should have the

[5] Klein ([Kle71; Kle73]) showed the projective nature of Beltrami's model and gave the formula for the metric in terms of cross-ratios, inspired by work of Cayley [Cay59]. For a more detailed historical account, refer to [AP15].

[6] Poincaré rediscovered the disk and half-plane models in 1882 and revealed the connection between 2-dimensional hyperbolic geometry and complex geometry [Poi82].

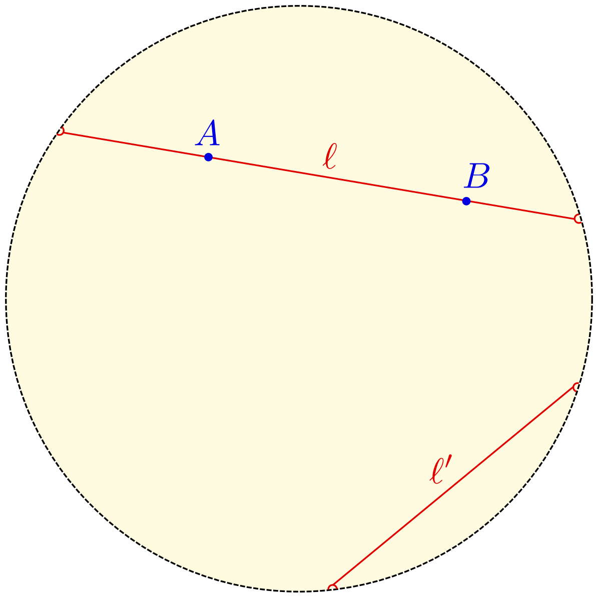


Figure 1.4: Points and lines in the Beltrami–Klein model.

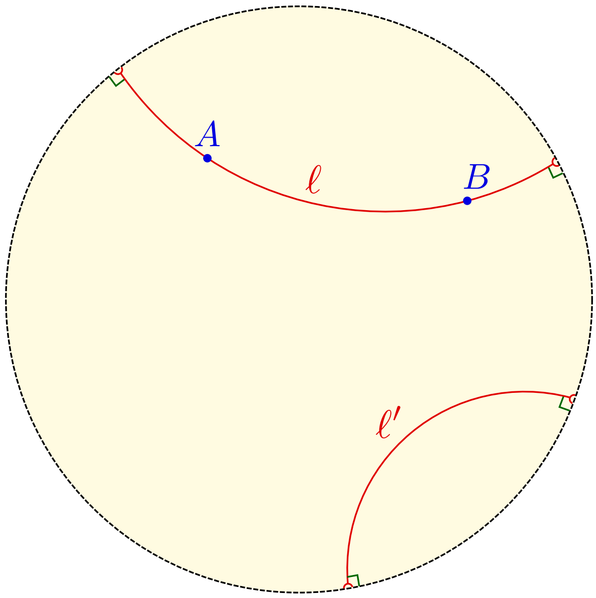


Figure 1.5: Points and lines in the Poincaré disk model.

following qualities:

(1) ***Consistency***: No two axioms are incompatible; more generally, no contradiction can be derived from the axioms.
(2) ***Completeness***: Any mathematical statement that makes sense in the theory should be either provable or disprovable.

(3) ***Independence***: No axiom should be a consequence of the others.

Clearly, the most important quality is consistency: an inconsistent theory is worthless. Completeness is less essential, but a theory feels imperfect without it. In theory, independence is not an important quality, but it is the question of the independence of Euclid's axioms that led to the discovery of hyperbolic geometry!

Let us go back to discussing the independence of Euclid's fifth postulate. Beltrami's work shows that, assuming there exists a model for Euclidean geometry (the Euclidean plane!), one can construct a model where the axioms of hyperbolic geometry, namely (E1)–(E4) and (H5), are satisfied. In particular, he created a model for Lobachevsky's geometry. From the point of view of logic, Beltrami's model shows that hyperbolic geometry is consistent if Euclidean geometry is. This is a direct consequence of Gödel's completeness theorem:

Theorem 1.5 (Gödel's completeness theorem)**.** *A theory is consistent if and only if it has a model.*

Consequently, assuming Euclidean geometry is consistent, Euclid's fifth postulate cannot be a consequence of the first four.

Let us mention that in general, achieving/proving consistency and completeness of a theory in first-order logic is tragically elusive. The celebrated incompleteness theorems of Gödel say that

(1) It is impossible to prove the consistency of a theory; not unless one includes it in a larger theory that is assumed consistent.
(2) A theory is never complete.

However, Gödel's incompleteness theorems only apply to theories containing arithmetic. In particular, they do not apply to Euclidean geometry. It has been proven that (a modern axiomatization of) Euclidean geometry is in fact consistent and complete: refer to [Mat17] as a starting point to seek more information about this.

1.4 Exercises

Exercise 1.1.

Beltrami–Klein disk and Poincaré disk

(1) Prove that Euclid's postulate (E1) holds in the Beltrami–Klein disk.
For now, we cannot really discuss postulates (E2), (E3), and (E4), because we have yet to define distances, angles, and isometries in this model, but we will see that they also hold.

(2) Show that Euclid's postulate (E5) does not hold in the Beltrami–Klein disk.

(3) Repeat the exercise with the Poincaré disk.

Exercise 1.2.

Triangles in the Poincaré disk

We recall that the Poincaré disk model is *conformal*: the angles between two lines (or curves) from the point of view of hyperbolic geometry is the same as their Euclidean angle (i.e., the angle between the tangents).

(1) Draw a right-angled triangle in the Poincaré disk.

(2) Show that in the Poincaré disk, the sum of angles in a triangle is always less than π. Argue that over all nondegenerate triangles, the sum ranges in the interval $(0, \pi)$.

Exercise 1.3.

Independence of Euclid's fifth postulate

Using Gödel's theorem, explain carefully why Beltrami's models for the hyperbolic plane show that hyperbolic geometry is no less consistent than Euclidean geometry. Conclude that if Euclidean geometry is consistent, then Euclid's fifth postulate is independent from the first four.

Remark: This exercise, just like the presentation of § 1.3, is naive: it implicitly assumes that Euclid's system meets the requirements of a theory as defined by first-order logic, where Gödel's theorem applies. This is not quite the case.

CHAPTER 2

Curvature

Disclaimer: This chapter is a draft.

In this second chapter, we propose a semi-formal introduction to the concept of curvature. This is another preliminary chapter, but it is important to study it because it includes definitions and propositions that will be used in subsequent chapters. In particular, a few useful notions of Riemannian geometry are introduced, without assuming any knowledge of differentiable manifolds.

Roughly speaking, curvature measures how much a geometric object—such as a curve, a surface, or a higher-dimensional object—deviates from being flat, in other words Euclidean. Exploring and developing this idea unveils substantial and beautiful mathematics. This is what this chapter attempts to do, although superficially in order to avoid getting overwhelmed by technical details or theoretical obstacles.

The main protagonist of this book, hyperbolic space, is the model geometric object of constant negative curvature. *One* of the goals of this course is to prove this fact in several ways, to derive some consequences, and beyond: to acquire a fairly deep understanding of the features of a negatively curved "world".

* * *

The flow of the chapter is quite straigthforward: we begin with the curvature of space curves in § 2.1, then surfaces § 2.2, and work our way towards more generality in § 2.3. We conclude with the model spaces of constant curvature in § 2.4 and a brief mention of curvature in metric spaces in § 2.5.

Prerequisites and references. Having some knowledge of differential geometry will make it far easier to read this chapter, but it is not a prerequisite. We only assume a solid background in Euclidean vector spaces and multivariable calculus, although we will recall many basic definitions. For readers looking to properly learn Riemannian geometry, I can recommend the great books of Jack Lee ([Lee18], preceded by [Lee11] and [Lee13]), I also like [GHL04] preceded by [Laf15]. For "elementary differential geometry" (curves and surfaces in Euclidean space), do Carmo [Car16] is a good option, but there are many others. Spivak's books [Spi99] are a thorough reference for differential geometry, though more advanced.

2.1 Curvature of space curves

Curves are the one-dimensional objects of differential geometry. A curve typically "lives" in an ambient space, such as the Euclidean plane, three-dimensional Euclidean space, or more general geometric objects (surfaces, higher dimensional manifolds, metric spaces, etc.). It is even possible to consider "abstract" curves, that do not live in any ambient space.

However, the curvature of a curve is a notion relative to the space in which it lives; abstract curves do not any have curvature in any reasonable sense[1]. In this section, we discuss the curvature of "space curves", i.e. curves in three-dimensional Euclidean space.

2.1.1 Basic definitions

Consider three-dimensional Euclidean space $E = \mathbb{R}^3$. More generally, we could take for E any Euclidean vector or affine space. Let us recall that a ***Euclidean vector space*** is a finite-dimensional vector space equipped with an inner product, and a ***Euclidean (affine) space*** is an affine space modelled on a Euclidean vector space.

Let γ be a ***smooth (parametrized) curve*** in E. By definition, this is a smooth map $\gamma\colon I \longrightarrow E$, where $I \subseteq \mathbb{R}$ is an interval of the real line (that is nonempty and not reduced to a point). In this book, we shall use ***smooth*** as an alias for "of class $\mathcal{C}^\infty$". The fact that we will always assume our curves and other differential geometric objects to be of class $\mathcal{C}^\infty$ is mainly a lazy habit; for instance it would just fine to work with curves of class $\mathcal{C}^2$ or $\mathcal{C}^3$.

The ***velocity*** of the curve γ at a time $t \in I$ is the derivative $\overrightarrow{\gamma'(t)} \in \mathbb{R}^3$. Note that when E is a Euclidean affine space, the velocity $\overrightarrow{\gamma'(t)}$ is not an element of E, but of the associated vector space. This being understood, we shall drop the arrow over $\gamma'(t)$, which is a useful notation but a bit cumbersome. Using the inner product of our vector space, one can measure the norm $\|\gamma'(t)\|$, called the ***speed*** of γ.

In many situations, one is not really interested in the parametrized curve γ itself, which is a map $I \longrightarrow E$, but only in its image, which is a subset of E. It is common, although not very rigorous, to say "the curve γ" to refer to either. One can always ***reparametrize*** a curve without changing its image: this consists in putting $\tilde{\gamma}(s) := \gamma(t)$ where $t = \varphi(s)$ is a change of variables given by a function $\varphi\colon J \longrightarrow I$ (which is assumed smooth, increasing, and bijective). Note that changing the parametrization of a curve does not change its image, but it does change its velocity and speed.

An example of quantity that is independent of parametrization is the ***length*** of a curve: by definition, this is the integral of its speed $l(\gamma) := \int_I \|\gamma'(t)\| \, \mathrm{d}t$. The fact that $l(\gamma)$ is left unchanged by a reparametrization is an immediate application of the change of variables theorem for integrals.

[1] Rather, any sensible definition of curvature should imply that curves have vanishing intrinsic curvature. One way to convince oneself of this is to realize that any metric curve is locally isometric to a Euclidean line (the arclength parameter provides a local isometry to $\mathbb{R}$), which is flat by definition.

The parametrized curve γ is called ***regular*** if its velocity (equivalently its speed) never vanishes. It is always possible to parametrize a regular curve ***by arclength***: this means that γ has unit speed (constant speed equal to 1). In this situation, the parameter is usually denoted s and called ***arclength parameter*** . This name comes from the fact that the arclength parameter is unique up to addition of a constant, and given by $s = \int_{t_0}^{t} \|\gamma'(u)\| \,\mathrm{d}u$, in other words s is the length of the curve γ between a fixed time t_0 and the time t.

Example 2.1. A ***circle*** in the Euclidean plane $E = \mathbb{R}^2$ is the set of points at a distance $R > 0$ (the ***radius***) from some fixed point $\Omega = (x_0, y_0)$ (the ***center***). It can be parametrized by $\gamma(t) = \big(x_0 + R\cos(\omega(t - t_0)), y_0 + R\sin(\omega(t - t_0))\big)$ where ω and t_0 are real constants. This parametrization has constant speed $v = R\omega$, in fact any constant speed parametrization is of this form (up to reversing time).

More generally, in a Euclidean space E, the set of points at distance $R > 0$ from a point $\Omega \in E$ is a ***sphere***, and its intersection with any affine plane P going through Ω is a circle of center Ω and radius R. This circle is parametrized by $\gamma(t) = \Omega + R\cos(\omega(t - t_0))\vec{i} + \sin(\omega(t - t_0))\vec{j}$, where $(\vec{i}, \vec{j})$ is an orthonormal basis of the vector space underlying P.

2.1.2 Curvature

Let γ be a smooth regular curve. Without loss of generality, one can assume that γ is parametrized by arclength, in other words has unit speed. The following proposition is elementary but conceptually important:

Proposition 2.2. *A curve γ parametrized by arclength is a straight line if and only $\gamma'' = 0$.*

In the language of differential geometry, a straight line parametrized by arclength (or more generally by constant speed) is called a ***geodesic***. The lemma above thus says that geodesics are curves with vanishing acceleration, a characterization that holds in great generality (for Riemannian manifolds).

When γ is any regular curve, still parametrized by constant speed, its acceleration is always normal to the curve. Indeed, taking the derivative of the identity $\|\gamma'(s)\|^2 = 1$ yields $2\langle\gamma'(s), \gamma''(s)\rangle = 0$, and since $\gamma'(s)$ is non-null and tangent to the curve, $\gamma''(s)$ is normal.

Informally speaking, the direction of $\gamma''(s)$ shows in which direction the curve is turning, and its norm indicates how fast it is turning (this is consistent with the preceding lemma, at least). The next definition thus sounds reasonable:

Definition 2.3. Let γ be a smooth space curve parametrized by arclength. The ***curvature*** of γ is the function $\kappa(s) := \|\gamma''(s)\|$.

Remark 2.4. This notion of curvature is *extrinsic*, in the sense that depends on how the curve is embedded in Euclidean space. If one moves the curve in space, even without stretching it, its curvature will change. We shall see that for surfaces, it is possible to define an *intrinsic* curvature. One observation that holds in great generality is that curvature has to do with second-order derivatives.

While this definition of curvature is extremely simple analytically, it is possible to give a more geometric interpretation of it. First observe that, when γ is a circle, its curvature is equal to the inverse of the radius (this is an easy calculation, given the parametrization given in Example 2.1). This makes sense: the smaller the radius, the sharper the turn.

It is possible to extend this interpretation to any regular curve γ by introducing the notion of ***osculating circle***. By definition, the osculating circle at $\gamma(s_0)$ is the circle having best contact with γ at s_0. More precisely, it is the unique circle parametrized by arclength such that $c(s_0) = \gamma(s_0)$, $c'(s_0) = \gamma'(s_0)$, and $c''(s_0) = \gamma''(s_0)$. We leave as an easy exercise to the reader to check that this circle is indeed uniquely defined. The radius $R(s_0)$ of the osculating circle is called the ***radius of curvature***. See Figure 2.1.

Remark 2.5. The osculating circle at $\gamma(s_0)$ is not well-defined when $\gamma''(s_0) = 0$, strictly speaking. Note however that the tangent line to the curve has second-order contact with the curve in this situation. One can therefore consider that this line is the osculating "circle", and that it has infinite radius $R(s_0) = +\infty$.

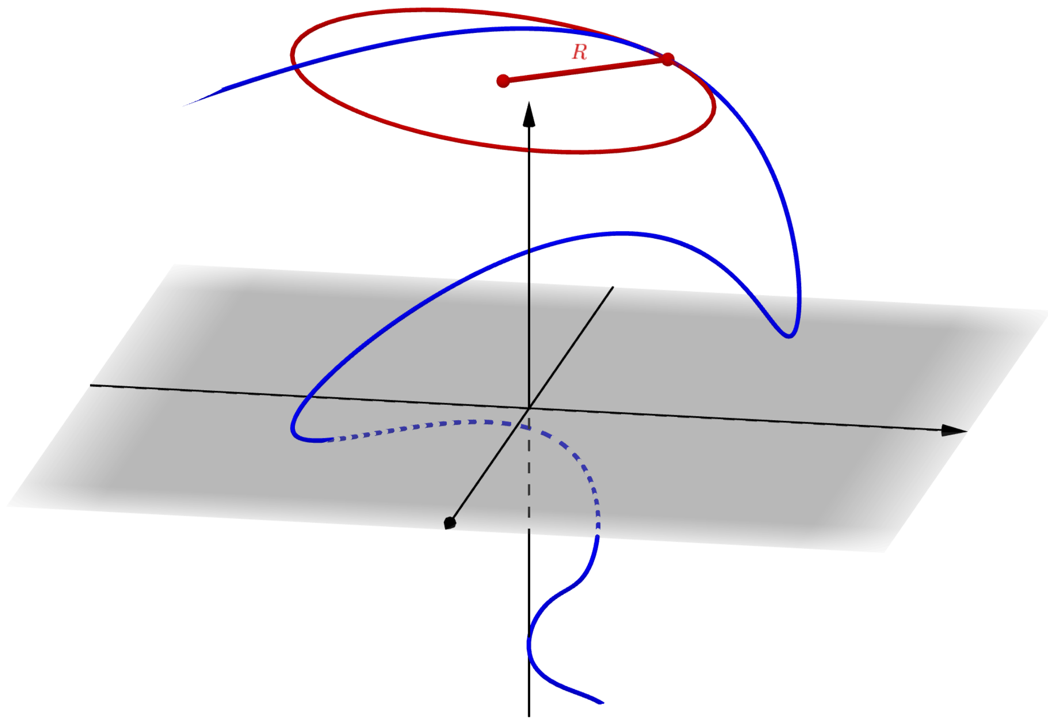


Figure 2.1: Osculating circle and radius of curvature.

Note: This figure and many others were created with the free software `GeoGebra` [Hoh+18].

Remark 2.6. If γ is not parametrized by arclength, one can always reparametrize it by arclength to define its curvature $\kappa(t)$ at any point. Alternatively, $\kappa(t)$ can be directly computed by

$$\kappa(t) = \frac{\|T'(t)\|}{\|\gamma'(t)\|} = \frac{\|\gamma'(t) \times \gamma''(t)\|}{\|\gamma'(t)\|^3}$$

where $T(t) = \frac{\gamma'(t)}{\|\gamma'(t)\|}$ is the ***unit tangent vector*** and $\gamma'(t) \times \gamma''(t)$ denotes the cross-product of vectors in $\mathbb{R}^3$. This formula is an elementary exercise: see Exercise 2.1.

Having set everything up, the following proposition should be fairly clear:

Proposition 2.7. *Let γ be a smooth regular curve. The curvature of γ is the inverse of its radius of curvature:*

$$\kappa(t) = \frac{1}{R(t)}$$

Proof. Since the circle of curvature does not "see" the parametrization of γ, we can assume that γ is parametrized by arclength. Let $s_0 \in I$. By definition, $\kappa(s_0) = \|\gamma''(s_0)\|$. Let $c(s)$ be the osculating circle to γ at s_0. We know that the curvature of a circle is equal to the inverse of its radius, therefore $\|c''(s_0)\| = \frac{1}{R(s_0)}$. Since $c''(s_0) = \gamma''(s_0)$, we conclude that $\kappa(s_0) = \frac{1}{R(s_0)}$. ∎

For completeness, we could further introduce the ***Frenet–Serret frame*** (T, N, B) , the notion of ***torsion*** , and mention the "fundamental theorem of space curves" (curves are determined by their curvature and torsion). This would not be require much more work but we shall not need these notions; out of interest, they are discussed in Exercise 2.3.

2.2 Curvature of surfaces

Let now $S \subseteq \mathbb{R}^3$ be surface. To be accurate, this means that S is a smooth 2-dimensional embedded submanifold of $\mathbb{R}^3$; there are several equivalent definitions of what this means, but let us not worry about these details.

Tangent plane. At any point $p \in S$, there is a tangent plane to the surface $\mathrm{T}_p S \subseteq \mathbb{R}^3$, which is an affine plane. There are many equivalent definitions of it; one possible way to think about tangent vectors $\vec{u} \in \mathrm{T}_p S$ is that they are the velocities of smooth curves $\gamma\colon I \to S$ (i.e. smooth curves $\gamma\colon I \to \mathbb{R}^3$ whose image is in S).

Unit normal. One can also define a ***unit normal*** $\vec{N}_p$ to the surface at p: it is a unit vector that is orthogonal to $\mathrm{T}_p S$. There are two choices for the unit vector $\vec{N}_p$. Locally, one can always make a consistent choice at all points near p (so that the map $p \mapsto \vec{N}_p$ is continuous). Globally, one can make a consistent choice for $\vec{N}$ if and only if the surface is orientable.

Geodesics. Among curves in S, the most special are ***geodesics***. Intuitively, a geodesic is easy to define: imagine that you have a little car toy, whose wheels are powered by a battery that never runs out. You can initially set the speed of rotation of the wheels, and never change it afterwards. Also, the car wheels are always straight, it never turns. If you release this car on a plane, its trajectory will be a straight line, parametrized by the car. More generally, if you release this car on a surface, it will define a geodesic. This is really the right way to think about geodesics: they are parametrized curves with constant intrinsic velocity, i.e. zero intrinsic acceleration. Of course, formalizing all this requires some work, which we skip. This description makes the following proposition intuitively obvious:

Proposition 2.8. *For any $\vec{v} \in \mathrm{T}_p S$, there exists a unique geodesic in S through p with initial tangent vector $\vec{v}$.*

We will denote this geodesic $\gamma_{\vec{v}}$. The following proposition gives a possible alternative definition for geodesics:

Proposition 2.9. *A curve γ on S is a geodesic if and only if it has constant speed and is locally length minimizing.*

Precisely, being locally length minimizing means that every $t_0 \in I$ and for every t_1 sufficiently close to t_0, the length of γ between t_0 and t_1 is minimal along all curves from t_0 to t_1.

We will make good use of the following proposition:

Proposition 2.10. *Let $f : S \longrightarrow S$ be an isometry (e.g., induced by an isometry of $\mathbb{R}^3$). Let $F \subseteq S$ denote a connected component of the fixed point set of f. If $v \in \mathrm{T}_p S$ is tangent to F, then the (image of) whole geodesic γ_v is contained in F.*

Proof. Consider the curve $\hat{\gamma}_v := f \circ \gamma_v$. Since f is an isometry, $\hat{\gamma}_v$ is also a geodesic. Moreover, since F is fixed by f, tangent vectors to F are fixed by $\mathrm{d}f$. It follows that $\hat{\gamma}_v'(0) = \mathrm{d}f(\gamma_v'(0)) = \mathrm{d}f(v) = v$. By uniqueness of the geodesic with initial velocity v, we conclude that $\hat{\gamma}_v = \gamma_v$. This shows that γ_v is contained in the fixed set of F, and one concludes by connectedness of γ_v. ∎

Remark 2.11. Proposition 2.10 holds more generally in any Riemannian manifold: the proof is the same.

Length of curves. Note that on S, one can measure the length of any curve $\gamma : I \longrightarrow S$: it is simply its length as a curve in $\mathbb{R}^3$.

Intrinsic metric (first fundamental form). Note that since the velocity of γ is always tangent to S, the length of curves in S only depends on the restriction of the inner product of $\mathbb{R}^3$ to the tangent planes to S. This data, the assignment $p \in S \mapsto g_p$ where $g_p x$ is the inner product on $\mathrm{T}_p S$, is called the ***intrinsic Riemannian metric*** on S, or ***first fundamental form***.

Remark 2.12 (Comment on the word "intrinsic"). Let $f : S \longrightarrow S'$ be an (Riemannian) isometry between surfaces in $\mathbb{R}^3$. By definition, this means that at any $p \in S$, the differential $\mathrm{d}f$ is a linear isometry between the Euclidean planes $(\mathrm{T}_p S, g_p)$ and $(\mathrm{T}_{f(p)} S, g'_{f(p)})$. This easily implies that f (locally) preserves lengths of curves, in particular f is a (local) metric isometry. A notion relative to surfaces is called ***intrinsic*** if, for any isometry $f : S \longrightarrow S'$, the notion on S' coincide with its transport from S to S' using f. One quickly see that the first fundamental form being intrinsic is essentially a tautology.

Extrinsic curvatures. We would like to define the ***extrinsic curvature*** of S at p in the direction $\vec{v} \in \mathrm{T}_p S$ as the curvature at p of the geodesic $\gamma_{\vec{v}}$ (see Remark 2.14). The only problem is that this does not have a sign, or rather it is always nonnegative. However, given a choice of unit normal $\vec{N}$, one can choose the sign as follows: we decide that the extrinsic curvature is positive if $\gamma''(0)$ and N have same direction ($\vec{N}$ points towards the center of the osculating circle), and is negative if they have opposite directions ($\vec{N}$ points away from the center of the osculating circle). For example, Figure 2.2 illustrates an negative extrinsic curvature. NB: It is a consequence of γ being a geodesic that its acceleration is always normal to the surface (by definition, a geodesic has vanishing intrinsic acceleration, which means that the orthogonal projection of the acceleration to the tangent space of the surface vanishes).

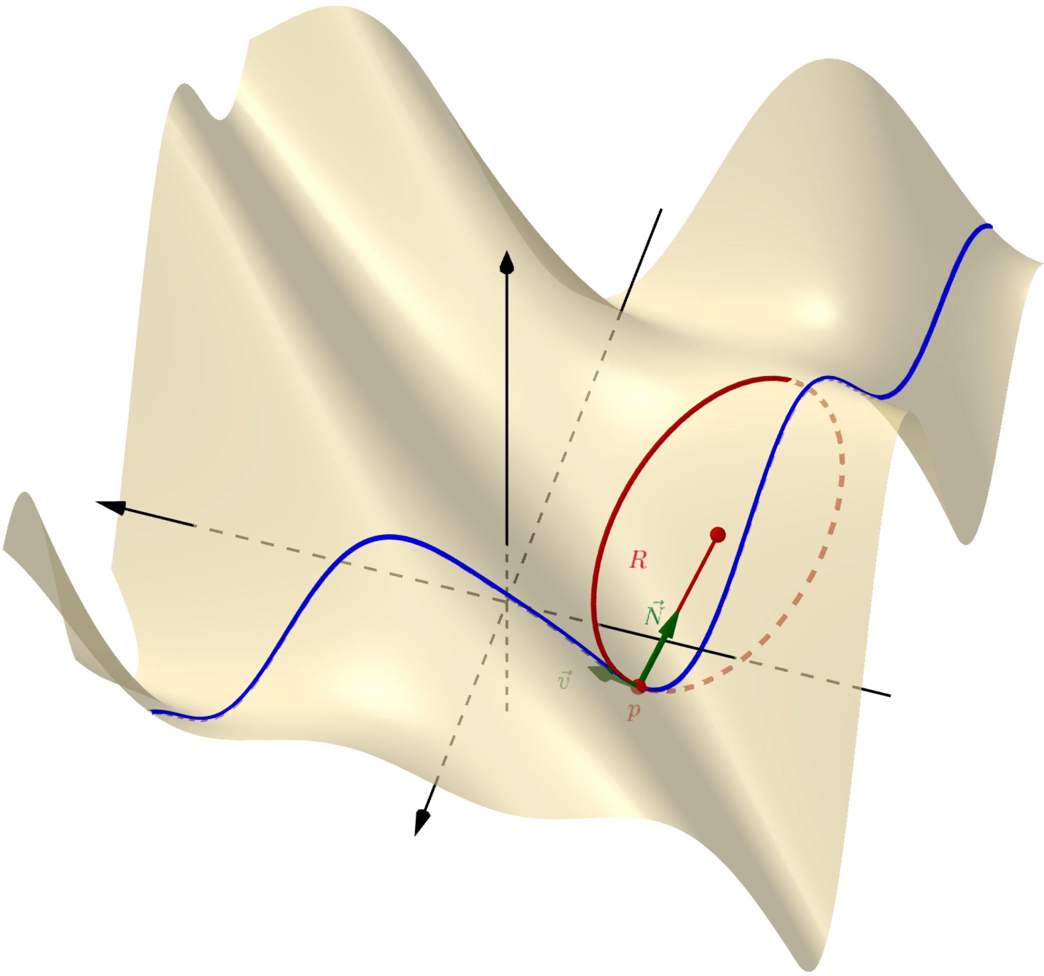


Figure 2.2: Any point $p \in S$ and tangent vector $\vec{v} \in \mathrm{T}_p S$ define a unique geodesic $\gamma_{\vec{v}}$. In this example, the acceleration $\gamma''_{\vec{v}}(0)$ and the chosen normal $\vec{N}$ have same direction: $\vec{N}$ points towards the center of the osculating circle. Therefore the extrinsic curvature $\rho(\vec{v})$ is positive, given by $\rho(\vec{v}) = +\frac{1}{R}$ where R is the radius of the osculating circle.

This definition of the extrinsic curvature, while theoretically right (see Remark 2.14), is not very practical because it is generally not easy to find the geodesic $\gamma_{\vec{v}}$ explicitly. Thankfully there is a variation of this definition that allows straightforward calculations. Let γ be any curve in S with $\gamma'(0) = \vec{v}$. We cannot just take the curvature of γ, because that is not independent of the choice of γ. However, the quantity $\langle \vec{N}, \gamma''(0)\rangle$ is independent of γ (see Proposition 2.13) That quantity, usually called ***normal curvature of*** γ, clearly coincides with the extrinsic curvature for the geodesic $\gamma_{\vec{v}}$ since $\gamma''_{\vec{v}}(0)$ is collinear to $\vec{N}$. (Another special curve having this property is the curve $\gamma_{\vec{v},N}$ obtained by intersecting the affine plane through p spanned by $\vec{v}$ and $\vec{N}$ with S, parametrized by arclength.)

Proposition 2.13. *Let $p \in S$ and $v \in \mathrm{T}_p\, S$. The normal curvature $\left\langle \gamma''(0), \vec{N} \right\rangle$ is independent of the choice of the curve γ such that $\gamma(0) = p$ and $\gamma'(0) = \vec{v}$. We call it* ***extrinsic curvature*** *of S at p in the direction $\vec{v}$ and denote it $\rho_p(\vec{v})$. Moreover, it can be written:*

$$\rho_p(\vec{v}) = \left\langle \gamma''(0), \vec{N} \right\rangle = -\left\langle \nabla_{\vec{v}}\vec{N}, \vec{v} \right\rangle .$$

Proof. Since the curve γ is always in S, its velocity is always orthogonal to $\vec{N}$:

$$\langle \vec{N}_{\gamma(t)}, \gamma'(t)\rangle = 0 .$$

Differentiating this identity gives

$$\langle (\nabla_{\gamma'(t)}\vec{N})_{\gamma(t)}, \gamma'(t)\rangle + \langle \vec{N}_{\gamma(t)}, \gamma''(t)\rangle = 0 .$$

At $t = 0$, this reads $\left\langle \nabla_{\vec{v}}\vec{N}, \vec{v} \right\rangle + \langle \vec{N}_{\gamma(t)}, \gamma''(0)\rangle = 0$. ∎

Second fundamental form. Proposition 2.13 shows that $\rho_p(\vec{v})$ is a quadratic form of $\vec{v}$: there exists a symmetric bilinear form $B_p\colon\, \mathrm{T}_p\, S \times \mathrm{T}_p\, S \longrightarrow \mathbb{R}$ such that $\rho_p(\vec{v}) = B_p(\vec{v}, \vec{v})$, namely:

$$B_p(\vec{u}, \vec{v}) = -\left\langle \nabla_{\vec{u}}\vec{N}, \vec{v} \right\rangle .$$

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

Remark 2.14. For advanced readers, let us mention that more generally, the second fundamental form can elegantly be defined as the (Riemannian) Hessian of the inclusion of a submanifold. This amounts to the definition above in terms of acceleration of geodesics. In general, the second fundamental form takes values in the normal bundle of the submanifold. Amusingly, the mean curvature is the trace of the Hessian (divided by the dimension), i.e. the "Laplacian" (also known as *tension field*) of the inclusion.

Principal curvatures, mean curvature, Gaussian curvature. By definition, the ***principal curvatures*** at p are the minimal and maximal values of the extrinsic curvatures at p in the directions of all unit vectors, attained in the respective ***principal directions of curvature***. The ***mean curvature*** $H_p \in \mathbb{R}$ (also sometimes called ***extrinsic curvature***) at p is defined as the average (half-sum) of principal curvatures, and the ***Gaussian curvature*** $K_p \in \mathbb{R}$ is the product of the principal curvatures.

A nice and immediate consequence of the spectral theorem is:

Theorem 2.15. *The principal curvatures are the eigenvalues of the second fundamental form B (or rather, of the matrix of B taken in any orthonormal basis). The principal directions of curvature are orthogonal, and eigenvectors of B. The mean curvature is the trace of B, and the Gaussian curvature is the determinant of B.*

A very important theorem is the *Theorema Egregium* (which roughly means "very important theorem"):

Theorem 2.16 (Theorema Egregium)**.** *The Gaussian curvature is intrinsic.*

In other words, if $f: S \to S'$ is an isometry, then $K = K' \circ f$. We will not prove the Theorema Egregium. The other most important theorem of the theory of surfaces in $\mathbb{R}^3$ is the theorem of Gauss–Bonnet:

Theorem 2.17 (Gauss–Bonnet Theorem)**.** *If S is a closed surface without boundary, then*

$$\int_S K \, \mathrm{d}A = 2\pi\chi(S) \, .$$

The integer $\chi(S)$ is the Euler characteristic of S, a topological invariant. It is remarkable that the Gauss–Bonnet theorem relates the geometry and the topology of the surface.

Remark 2.18. The Gauss–Bonnet theorem holds more generally for abstract Riemannian surfaces, possibly with boundary. We will briefly see this general version in Chapter 14.

2.3 Curvature of Riemannian manifolds

2.3.1 Riemannian surfaces

Let S be a surface in $\mathbb{R}^3$. More generally, S can be any "abstract surface" (2-dimensional manifold), whatever that means—at the very least, S has a well-defined tangent space $\mathrm{T}_p S$ at any point $p \in S$. Suppose that, instead of taking the restriction of the inner product of $\mathbb{R}^3$ in $\mathrm{T}_p S$, we take any other inner product product. In other words, we choose a map g which assigns to a point p an inner product in $\mathrm{T}_p S$, and we require that g depends smoothly on p, whatever that means. (It means, for instance, that $g_p(X, Y)$ is a smooth function of $p \in S$ whenever X and Y are smooth vector fields.) Such a family of inner products on the tangent spaces of S is called a ***Riemannian metric*** on S.

An important class of examples is when $S = \Omega$ is an open subset of $\mathbb{R}^2$. In this case, $\mathrm{T}_p S$ is canonically identified to $\mathbb{R}^2$ for every $p \in S$. Therefore a Riemannian metric on S is simply a $\mathcal{C}^\infty$ map $g: \Omega \to \mathrm{S}_2^+(\mathbb{R})$ where $\mathrm{S}_2^+(\mathbb{R})$ indicates the set of symmetric positive definite 2×2 matrices.

As an example, let us look at the Poincaré half-plane (which we will study in Chapter 10). We take $\Omega = \{(x, y) \in \mathbb{R}^2, y > 0\}$ and $g_{(x,y)} = \frac{g_0}{y^2}$, where $g_0 = \langle \cdot, \cdot \rangle$ is the standard inner product on $\mathbb{R}^2$. In usual Riemannian geometry notations, this Riemannian metric is written

$$\mathrm{d}s^2 = \frac{\mathrm{d}x^2 + \mathrm{d}y^2}{y^2}\,.$$

Remark 2.19 (Riemannian geometry notations). Professors or textbooks rarely take the time to carefully explain notations such as $\mathrm{d}s^2 = \frac{\mathrm{d}x^2+\mathrm{d}y^2}{y^2}$; let me try to correct this bad habit.

The notation $g = \mathrm{d}s^2$ for the Riemannian metric is a customary abuse of notations. It relates to the fact that when s is an arclength parameter, $\mathrm{d}s$ is called the "line element", because it gives the length of the curve when integrated. Given any parametrization $t \mapsto \gamma(t)$ of a curve, the line element can be computed as $\mathrm{d}s = \|\vec{v}\|\,\mathrm{d}t$, where $\vec{v} = \gamma'(t)$ is the velocity. In other words, we have $\mathrm{d}s = \sqrt{g(\vec{v},\vec{v})}\,\mathrm{d}t$. By abuse of notation, the quadratic form $\vec{v} \mapsto g(\vec{v},\vec{v})$ and the associated symmetric bilinear form g are both denoted $\mathrm{d}s^2$.

Now let us explain the notations $\mathrm{d}x$, $\mathrm{d}y$, $\mathrm{d}x^2$, etc. Basically, the inner product g on $\mathbb{R}^2$ with matrix $\begin{pmatrix} \alpha & \beta \\ \beta & \gamma \end{pmatrix}$ is denoted $g = \alpha\,\mathrm{d}x^2 + 2\beta\,\mathrm{d}x\,\mathrm{d}y + \gamma\,\mathrm{d}y^2$. But why?

Technically, $\mathrm{d}x$ is the derivative of the function $\Omega \subseteq \mathbb{R}^2 \to \mathbb{R}$, $(x,y) \mapsto x$. It is the constant map on Ω with values in $\mathcal{L}(\mathbb{R}^2,\mathbb{R})$ which is always equal to e_1^*. Similarly, $\mathrm{d}y$ is the constant map equal to e_2^*. Recall that (e_1^*, e_2^*) denotes the dual basis of the canonical basis of $\mathbb{R}^2$: e_i^* is the linear form $u \mapsto u_i$.

Finally, $\mathrm{d}x\,\mathrm{d}y$ stands for the *symmetric product* of $\mathrm{d}x$ and $\mathrm{d}y$, while $\mathrm{d}x^2$ (resp. $\mathrm{d}y^2$) is the symmetric product of $\mathrm{d}x$ (resp. $\mathrm{d}y$) with itself. In general, the *symmetric product* of two linear forms α and β is the symmetric bilinear form defined by $(u,v) \mapsto \frac{\alpha(u)\beta(v)+\alpha(v)\beta(u)}{2}$.

When (S,g) is a surface equipped with a Riemannian metric, one can seamlessly develop all the same notions as before: the velocity of curves on S still make sense, so does their speed (using g), etc. In particular, geodesics are well-defined. Note however that unless we have an isometric embedding of (S,g) in a Euclidean space $\mathbb{R}^N$[2], we cannot define the Gaussian curvature like before. Nevertheless, it is possible to define the Gaussian curvature in a consistent way, so that whenever $S \to S'$ is an isometry, $K = K' \circ f$. This can be done explicitly with formulas, but they are not very insightful.

One particular case of interest is when g is *conformally flat metric*, i.e. g is pointwise proportional to a Euclidean metric. On $\Omega \subseteq \mathbb{R}^2$, this means that $g = f g_0$ where $f = e^{2\varphi} \colon \Omega \to \mathbb{R}$ is some smooth positive function, and $g_0 = \mathrm{d}x^2 + \mathrm{d}y^2$ is the Euclidean metric. In this case, the Gaussian curvature is computed by:

$$K = -e^{-2\varphi}\Delta\varphi\,.$$

(This is a particular case of the more general formula $K = e^{-2\varphi}\left(K_0 + \Delta_0\varphi\right)$ relating the Gaussian curvatures of any two conformal metrics.)

[2]Such an isometric embedding always exists: this is the *Nash embedding theorem.*

In particular, for the Poincaré half-plane, we have $\varphi = -\log(y)$, from which we find $\Delta\varphi = \frac{1}{y^2}$, and $K = -1$. Thus we already have a "proof" that the Poincaré half-plane has constant curvature -1. We shall later see several other proofs of this fundamental feature of the hyperbolic plane.

2.3.2 Higher dimensional Riemannian manifolds

Sectional curvature

When M is a higher dimensional Riemannian manifold, the Gaussian curvature is generalized as the ***sectional curvature***. This depends on the choice of a point $p \in S$ and a 2-dimensional subspace $P \subseteq \mathrm{T}_p M$.

For instance, assume M is a submanifold of R^n. One can still equip M with a Riemannian metric by restricting the Euclidean inner product to each tangent space of M. Hence we can still measure lengths of curves, talk about geodesics, etc.

The definition of the sectional curvature is as follows. Consider the surface S_P obtained by taking all geodesics in M are tangent to P at p (their initial velocity belongs to P). Then S is a surface (for connoisseurs of Riemannian geometry: S is just $\exp_p(P)$. Just take its Gaussian curvature.

Riemann curvature tensor

The sectional curvature is very geometric, but as an mathematical object it is a bit complicated: one could say it is a real-valued function on the Grassmannian of 2-planes $\mathrm{Gr}_2 M$. It turns out that all sectional curvatures can be encoded in an object that has a concise definition and is easier to calculate: the Riemann curvature tensor. This object R is a quadrilinear map on the tangent space: given 4 tangent vectors $v_1, v_2, v_3, v_4 \in \mathrm{T}_p M$, it assigns a number $R(v_1, v_2, v_3, v_4)$. This is a ***tensor***, i.e. is linear in $v_1, v_2, v_3, v_4 \in \mathrm{T}_p M$, moreover it has several symmetries, but let us not go into details.

The relation between the Riemann curvature tensor and the sectional curvature is that for any two vectors u and v at p, the sectional curvature of the plane spanned by u and v is

$$K(u,v) = \frac{\langle R(u,v)v, u\rangle}{\|u \wedge v\|^2}$$

where we denote $\|u \wedge v\|^2 = \|u\|^2\|v\|^2 - \langle u, v\rangle^2$ (and $\langle\cdot,\cdot\rangle = g$ is the Riemannian metric).

It is "just" linear algebra to show that K determines R and conversely. For a differential geometer, R has an incredibly pleasant definition: it is exactly the lack of commutation of second derivatives in M. More precisely, it is the curvature of the Levi-Civita connection:

$$R(X,Y) = \nabla^2_{X,Y} - \nabla^2_{Y,X}$$

But explaining this more precisely would take us beyond the scope of this course.

2.3.3 Taylor expansion of the metric

In a way, the curvature of a Riemannian manifold is precisely the measurement of how the Riemannian metric locally differs from the Euclidean metric to second order. This point of view is in fact faithful to Bernhard Riemann's original approach: he defines the curvature tensor in his 1854 habilitation [Rie13] via the formula:

$$g_{ij} = \delta_{ij} - \frac{1}{3} R_{ikjl} x^k x^l + O(r^3)$$

in normal coordinates. Let us give a more geometric characterization (we refer to [GLM19, Appendix A] for details).

Let $p \in M$ and consider two tangent vectors $u, v \in \mathrm{T}_p M$. Denote by γ_u and γ_v the geodesics from p with initial velocities u and v respectively. Then

$$d\left(\gamma_u(t), \gamma_v(t)\right)^2 = \|u - v\|^2\, t^2 - \frac{1}{3}\langle R(u,v)v, u\rangle\, t^4 + O(t^5)\,.$$

as $t \longrightarrow 0$. In other words, with the sectional curvature:

$$d\left(\gamma_u(t), \gamma_v(t)\right)^2 = \|u - v\|^2\, t^2 - \frac{1}{3} K(u,v) \|u \wedge v\|^2\, t^4 + O(t^5)\,.$$

The important thing to note is that $d_{\mathrm{E}}\left(\gamma_u(t), \gamma_v(t)\right) := \|u - v\|^2\, t^2$ is exact in a Euclidean space, therefore the next order term gives the deviation from the Euclidean distance. In particular, observe that if $K > 0$, then $d < d_{\mathrm{E}}$: the distance between geodesics is closer than in a Euclidean space; on the contrary, if $K < 0$ then $d > d_{\mathrm{E}}$: geodesics diverge faster than in a Euclidean space. See Figure 2.3 for an illustration.

2.4 Model spaces of constant curvature

Using classical techniques of Riemannian geometry, one can show:

Theorem 2.20. *Any two Riemannian manifolds of the same dimension and with same constant sectional curvature are locally isometric.*

In other words, n-dimensional metrics of constant sectional curvature $k \in \mathbb{R}$ are locally unique. In a nutshell, the proof goes as follows: using Jacobi fields, one sees that for a Riemannian manifold of constant sectional curvature, the Riemannian metric's expression is forced to a have a fixed expression in normal coordinates. For more details, see [Lee18, Theorem 10.14, Corollary 10.15].

By definition, a ***model space*** or a ***space form*** of constant sectional curvature is a complete, simply-connected manifold of constant sectional curvature. We recall that a Riemannian manifold is called ***complete*** if lines (geodesics) can be extended indefinitely. Equivalently, it is complete as a metric space (Hopf-Rinow theorem).

Using the Cartan–Hadamard theorem, one can derive from the previous theorem:

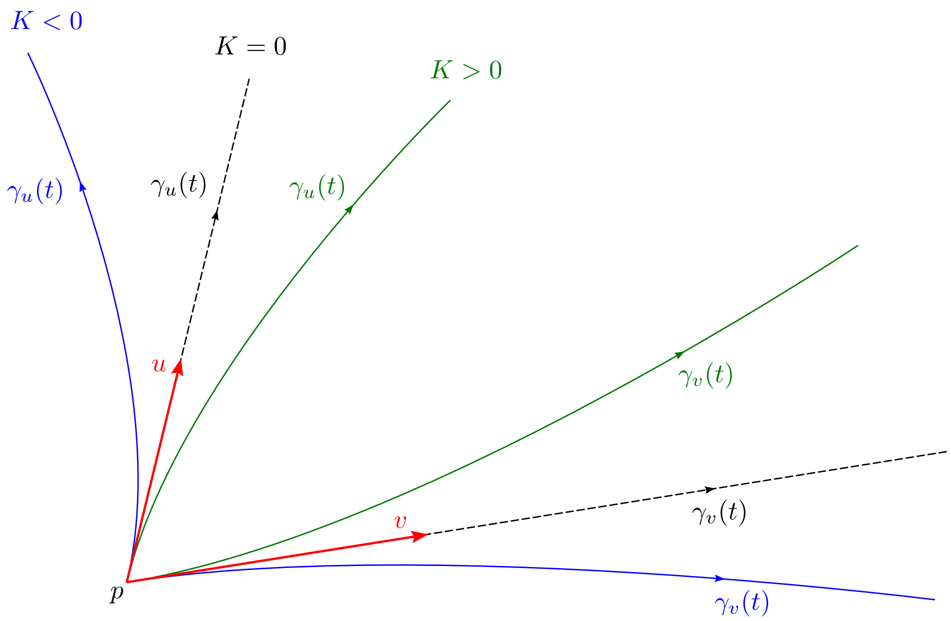


Figure 2.3: Geodesic deviation: the distance between geodesics $\gamma_u(t)$ and $\gamma_v(t)$ is controlled by the sectional curvature $K(u, v)$.

Theorem 2.21. *In any dimension, the space form of constant curvature $k \in \mathbb{R}$ is unique up to isometry.*

Space forms are thus essentially unique; also, they exist! Depending on the sign of $k \in \mathbb{R}$, the space form $\mathbb{M}_k^n$ takes three different forms:

- For $k > 0$, the space form of constant curvature k is denoted $\mathbb{S}_R^n$ where $k = \frac{1}{R^2}$. The usual model for it is the Euclidean sphere of squared radius R in $\mathbb{R}^{n+1}$.
- For $k = 0$, the space form of constant curvature k is Euclidean space $\mathbb{E}^n$. The usual model for it is $\mathbb{R}^n$ with its standard Euclidean structure.
- For $k < 0$, the space form of constant curvature k is hyperbolic space $\mathbb{H}_R^n$ where $k = -\frac{1}{R^2}$. One model for it is the pseudo-Euclidean sphere of "imaginary radius" $R\sqrt{-1}$ in Minkowski space $\mathbb{R}^{n,1}$, as we shall see in Chapter 5. However, we shall also see other useful models: the Beltrami–Klein model (Chapter 8), the Poincaré ball and half-space models (Chapter 10).

Combining the two previous theorems, we can state:

Theorem 2.22. *Let M be a complete Riemannian manifold of constant sectional curvature $k \in \mathbb{R}$. Then M is covered by the space form $\mathbb{M}_k^n$. In other words, M is isometric to a quotient of the space form $\mathbb{M}_k^n$ by a free and wandering action of a discrete group of isometries.*

Note that, after scaling the metric, one can assume M has constant sectional curvature 1, 0, or -1. In many ways, the latter case is the most interesting. A manifold with constant sectional curvature -1 is called a ***hyperbolic manifold***.

The previous theorem implies that any complete hyperbolic manifold is a quotient of hyperbolic space $\mathbb{H}^n$. This is remarkable because it shows that while Riemannian metrics are rather flexible objects, hyperbolic metrics are quite rigid. A consequence of this is that the study of hyperbolic manifolds is more algebraic, and less differential, that one could expect. This explains why a course in hyperbolic geometry belongs in the realm of classical geometry more than differential geometry, much like a course in Euclidean geometry.

2.5 Curvature of metric spaces

Metric spaces are clearly much more general than Riemannian manifolds. Can we extend the notion of curvature to metric spaces? In a nutshell, yes, there are several slightly different definitions of curvature in metric spaces that coincide for Riemannian manifolds. However:

- All such definitions build on the Riemannian case, or at least the model spaces of constant sectional curvature (called ***space forms***). Therefore, one should start by understanding curvature in Riemannian manifolds, or at least in space forms.
- There is a trade-off: the notion of curvature in metric spaces is not as precise as in Riemannian manifolds.

Despite these nuances, the notion of curvature in metric spaces is very useful. In particular, Gromov hyperbolic spaces offer the right frame to classify the isometries of hyperbolic space. We postpone this discussion until Chapter 12.

2.6 Exercises

Exercise 2.1.

Formula for the curvature of a space curve

Let $\gamma\colon I \longrightarrow E$ be a smooth regular curve in a 3-dimensional Euclidean space. We assume that $\gamma(t)$ is not necessarily parametrized by arclength and we call s an arclength parameter.

(1) Let $T(t) := \frac{\gamma'(t)}{\|\gamma'(t)\|}$ (*unit tangent*). Using the chain rule, show that $T(s) = \frac{\mathrm{d}\gamma}{\mathrm{d}s}$.

(2) Show that $\kappa(s) = \left\|\frac{\mathrm{d}T}{\mathrm{d}s}\right\|$ and $\kappa(t) = \frac{\|T'(t)\|}{\|\gamma'(t)\|} = \frac{\|\gamma'(t)\times\gamma''(t)\|}{\|\gamma'(t)\|^3}$.

(3) Compute the curvature of the ellipse with equation $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$ in the xy-plane.

Exercise 2.2.

Osculating ellipse

Let γ_1 and γ_2 be two smooth regular curves in a Euclidean space E. Assume that γ_1 and γ_2 meet at some point $p = \gamma_1(t_1) = \gamma_2(t_2)$.

(1) One could say that *γ_1 and γ_2 have order k contact at p* if there is an equality of derivatives $\gamma^{(i)}(t_1) = \gamma^{(i)}(t_2)$ for all $i \in \{0, \dots, k\}$. Is this notion invariant under reparametrization of γ_1 or γ_2? Is it invariant under post-composition of γ_1 and γ_2 by a diffeo $\varphi\colon \mathbb{R}^3 \longrightarrow \mathbb{R}^3$?

(2) We say instead that *γ_1 and γ_2 have order k contact at p* if there is an equality of derivatives $\gamma^{(i)}(s_1) = \gamma^{(i)}(s_2)$ for all $i \in \{0, \dots, k\}$ after γ_1 and γ_2 have been reparametrized by arclength. Answer the same questions asked in (1).

(3) Let us call γ_1 and γ_2 ***osculating*** at p if they have second-order contact at p in the sense of (2). Is this consistent with the definition of the osculating circle? Is it true that γ_1 and γ_2 are osculating at p if and only if they are tangent and have same curvature at p?

(4) (*) Let C denote the unit circle in $\mathbb{R}^2$ (with equation $x^2 + y^2 = 1$). Show that for any $p_0 \in C$ and for any $p \neq p_0 \in \mathbb{R}^2$, there exists a unique ellipse through p that has order 4 contact with C at p_0. *Hint: Start with the case $p_0 = (1, 0)$ and $p = (x_p, 0)$.*

Exercise 2.3.

Frenet–Serret frame and torsion

Let $\gamma\colon I \longrightarrow E$ be a smooth regular curve in a 3-dimensional Euclidean space, parametrized by arclength. Assume that γ'' does not vanish.

(1) Let $T(s) := \gamma'(s)$ (***unit tangent***), $N(s) := \frac{\gamma''(s)}{\|\gamma''(s)\|}$ (***principal normal***), and $B(s) := T(s) \times N(s)$ (***unit binormal***). Show that $(T(s), N(s), B(s))$ is an orthonormal basis.

(2) The frame of E with origin $\gamma(s)$ and basis $(T(s), N(s), B(s))$ is called ***Frenet–Serret frame*** (see Figure 2.4). What is the equation of the osculating circle in this frame?

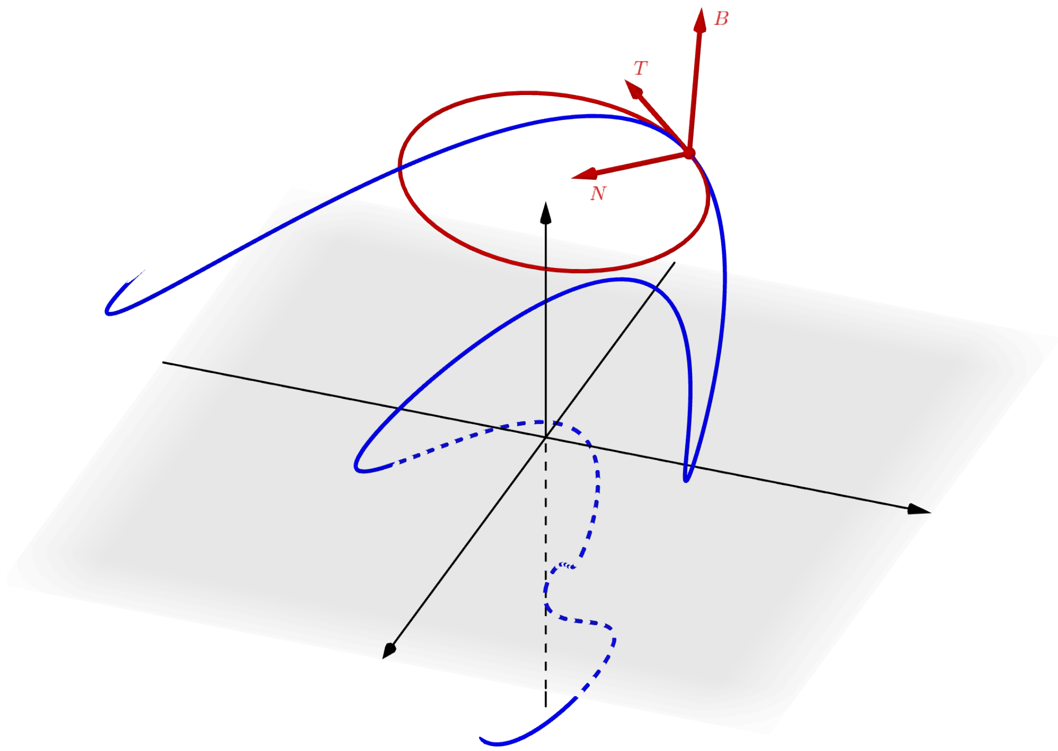


Figure 2.4: Frenet–Serret frame.

(3) We pick a fixed orthonormal frame of E. Let $Q(s)$ be the matrix whose rows are given by the coordinates of $T(s)$, $N(s)$, and $B(s)$ respectively.
  (a) Argue that $Q \in \mathrm{O}(3, \mathbb{R})$, i.e. $Q(s)Q(s)^\mathsf{T} = I_3$.
  (b) Derive from the previous question that $Q'(s)Q(s)^\mathsf{T}$ is antisymmetric.
  (c) On the other hand, show that the first row of $Q'(s)Q(s)^\mathsf{T}$ is $\begin{bmatrix} 0 & \kappa(s) & 0 \end{bmatrix}$.
  (d) Derive from the two previous question that there exists a number $\tau(s)$, called the ***torsion*** of γ, such that $Q'(s)Q(s)^\mathsf{T} = \begin{bmatrix} 0 & \kappa(s) & 0 \\ -\kappa(s) & 0 & \tau(s) \\ 0 & -\tau(s) & 0 \end{bmatrix}$.
  (e) Conclude that the ***Frenet–Serret formulas*** hold:

$$\begin{aligned} T' &= \kappa N \\ N' &= -\kappa N + \tau B \\ B' &= -\tau B \end{aligned}$$

(4) Check that the ***helix*** $\gamma(t) = (a\cos t, a\sin t, bt)$ has constant curvature and torsion.

(5) (*) Using the Picard–Lindelöf (aka Cauchy–Lipschitz) theorem, show that the curvature and torsion of a space curve determine it uniquely up to an affine isometry of E. (This is the "fundamental theorem of space curves".)

(6) Show that any curve with constant nonzero curvature and torsion is a helix.

Exercise 2.4.

Mean curvature

Recall that for a surface $S \subseteq \mathbb{R}^3$, we defined the mean curvature H_p at a point $p \in S$ as the half-sum of the principal curvatures. Show that H_p could instead be defined as, quite literally, the mean extrinsinc curvature at p. (First you'll have to make sense of this statement!)

Exercise 2.5.

The sphere

Let S^n_R denote the sphere of radius $R > 0$ centered at the origin in $\mathbb{R}^{n+1}$. We would like to understand geodesics and curvature in S^n_R. This exercise may seem basic, but it is very important: we will follow the same strategy for the hyperboloid in Minkowsi space.

(1) Show that any linear isometry of $\mathbb{R}^{n+1}$ induces a Riemannian isometry of S^n_R. Optional: show that the group of isometries of S^n_R is $\mathrm{O}(n+1)$.

(2) For now, we consider the sphere $S = S^2_R$ in $\mathbb{R}^3$.
   (a) Show that for any $p \in S$ and $v \in \mathrm{T}_p S$, there exists a plane $H \subseteq \mathbb{R}^3$ such that the reflection s_H through H leaves p and v invariant.
   (b) Show that geodesics on S are exactly the great circles (intersection of S with planes through the origin), parametrized with constant speed.
   (c) Show that we have the explicit expression:
$$\gamma_v(t) = \cos(\|v\| t)\, p + R \sin(\|v\| t) \frac{v}{\|v\|}\,.$$
   (d) Let $p, q \in S$. Show that their distance on S is given by $d(p,q) = R\, \measuredangle(p,q)$ where $\measuredangle(p,q)$ denotes the unoriented angle between p and q seen as vectors in $\mathbb{R}^3$.

(3) What is the exterior unit normal N at p? Show that the extrinsic curvature $\rho_p(v)$ is equal to $-\frac{1}{R}$ for any unit vector v. Conclude that the Gaussian curvature is $\frac{1}{R^2}$ at p, and hence everywhere.

(4) Let $n \geqslant 2$.
   (a) Show that (2) remains true with S^n_R instead of S and $\mathbb{R}^{n+1}$ instead of $\mathbb{R}^3$, as long as by *plane* we mean a 2-dimensional subspace.
   (b) Let $P \subseteq \mathrm{T}_p S^n_R$ be a 2-plane. Denote $E_P \subseteq \mathbb{R}^{n+1}$ the subspace spanned by p and P. Show that the union of geodesics in S^n_R with initial velocity in P is the sphere S_P of radius R in E_P. *In the terminology of Riemannian geometry:* $\exp_p(P) = S_P$.

(c) Conclude that S^n_R has constant sectional curvature $\frac{1}{R^2}$.

Exercise 2.6.

The tractricoid

One of the obstacles to the discovery of the hyperbolic plane is that it cannot be smoothly completely embedded as a surface in $\mathbb{R}^3$. [3] However, it is possible to smoothly embed a piece of the hyperbolic plane in $\mathbb{R}^3$, as this exercise illustrates.

(1) Consider the ***tractrix*** curve in the xz-plane parametrized by:

$$\begin{aligned} \gamma\colon\ [0,+\infty) &\longrightarrow \mathbb{R}^3 \\ t &\longmapsto (x(t) = \operatorname{sech} t, y(t) = 0, z(t) = t - \tanh t) \end{aligned}$$

where $\operatorname{sech} = \frac{1}{\cosh}$ is the hyperbolic secant and $\tanh = \frac{\sinh}{\cosh}$ is the hyperbolic tangent. Draw the tractrix in the plane. Optional: Show that the tractrix is the path followed by a reluctant dog on a leash (in German, a tractrix is a ***Hundekurve***).

(2) The ***tractricoid*** (sometimes called ***pseudosphere***[4]) is the surface S in $\mathbb{R}^3$ obtained by rotating the tractrix defined above around the z-axis. Show that it has parametric equations:

$$\begin{aligned} x &= \operatorname{sech} t \cos\theta \\ y &= \operatorname{sech} t \sin\theta \\ z &= t - \tanh t\,. \end{aligned}$$

Show that rotations around the z-axis and reflections through vertical planes containing the z-axis are isometries of S. Draw a sketch of S.

(3) We denote $f(\theta,t) := (x(\theta,t), y(\theta,t), z(\theta,t))$. Consider the curves $c_t(\theta) = f(\theta,t)$ when t is fixed ("parallels") and $\gamma_\theta(t) = f(\theta,t)$ when θ is fixed ("meridians"). Draw such curves on S. Using a symmetry argument, show that the curves $\gamma_\theta(t)$ are geodesics up to parametrization.

(4) Consider a point $p = f(\theta_0, t_0)$ on the tractricoid. Our goal is to show that the Gaussian curvature of S at p is -1.

(a) Explain why it is enough to show it when $\theta_0 = 0$.

[3] There are no complete surfaces of constant Gaussian curvature -1 of class C^2 in $\mathbb{R}^3$ (Efimov's theorem, 1964 [Efi64], also see [Mil72]). Hilbert first proved it for class C^4 in 1901. Surprisingly, there are C^1 embeddings of the hyperbolic plane in $\mathbb{R}^3$. This is a corollary of the Nash-Kuiper C^1 embedding theorem. See `http://www.math.cornell.edu/~dwh/papers/crochet/crochet.html` for illustrations of crocheted hyperbolic planes.

[4] Depending on authors, *pseudosphere* may refer to the tractricoid specifically, or to any surface in $\mathbb{R}^3$ of Gaussian curvature -1. I find the term more appropriate for level sets of the quadratic form in a pseudo-Euclidean vector space (this includes the hyperboloid model of the hyperbolic plane).

   (b) Compute the velocities of c_{t_0} and γ_0 at p. Derive an expression of the unit normal vectors at p.
   (c) Compute the extrinsic curvatures of S at p in the unit directions tangent to c_{t_0} and γ_0.
   (d) Using a symmetry argument, explain why the principal directions of curvatures of S at p must be tangent to c_{t_0} or γ_0. Derive the value of the principal curvatures at p, conclude that S has Gaussian curvature -1 at p, and hence everywhere.

(5) Compute the arclength parameter of $\gamma(t)$. Show that the tractricoid is incomplete.

Exercise 2.7.

The Poincaré disk

The ***Poincaré disk*** $\mathbb{D}$ is defined as the unit disk equipped with the Riemannian metric:

$$\mathrm{d}s^2 = \frac{4\left(\mathrm{d}x^2 + \mathrm{d}y^2\right)}{(1 - x^2 - y^2)^2}$$

In this exercise, we denote $O \in \mathbb{D}$ the point which is at the origin in $\mathbb{R}^2$.

(1) Show that the Poincaré metric on $\mathbb{D}$ is conformal to the Euclidean metric. Is the Euclidean metric complete on $\mathbb{D}$?
(2) Show that any $f \in \mathrm{O}(2)$ induces an isometry of $\mathbb{D}$ that fixes O. Optional: show the converse.
(3) Show that any diameter of $\mathbb{D}$ (straight chord through the origin) is a geodesic. *Hint: consider the fixed points of a reflection $f \in \mathrm{O}(2)$.*
(4) Find a parametrization of geodesics through the origin. Find an expression of the distance between O and an arbitrary point in $\mathbb{D}$.
(5) Show that $\mathbb{D}$ is complete. *Use the Hopf-Rinow theorem.*
(6) Compute the curvature of $\mathbb{D}$.

Exercise 2.8.

Euclid's postulates for Riemannian surfaces (*)

Give an interpretation of Euclid's postulates for Riemannian surfaces and discuss their implications.

This exercise is not easy, and best suited to students with a solid background of Riemannian geometry. Regardless, I recommend that you read the solution.

Part II

Pseudo-Euclidean geometry and the hyperboloid model

A four-dimensional continuum described by the "co-ordinates" x_1, x_2, x_3, x_4, was called "world" by Minkowski, who also termed a point-event a "world-point". [...] We can regard Minkowski's "world" in a formal manner as a four-dimensional Euclidean space (with an imaginary time coordinate) ; the Lorentz transformation corresponds to a "rotation" of the co-ordinate system in the four-dimensional "world."

– Albert Einstein[5]

The mathematical education of the young physicist [Albert Einstein] was not very solid, which I am in a good position to evaluate since he obtained it from me in Zurich some time ago.

– Hermann Minkowski[6]

[5] Einstein, *Relativity: The Special and the General Theory* [Ein15].

[6] Quoted from [New00].

CHAPTER 3

Pseudo-Euclidean spaces

In this chapter, we introduce pseudo-Euclidean spaces and, a special case, the Minkowski space of any dimension. Our main motivation for studying Minkowski space is that it is the stage for the hyperboloid model introduced in Chapter 5. Remarkably, Minkowski space also plays a central role in the theory of relativity presented in the “bonus” Chapter 6.

In modern mathematics, a *Euclidean space* is defined as a finite-dimensional real vector space equipped with an inner product, that is a positive definite symmetric bilinear form. (More generally, it can be an affine space modelled on such a vector space.) *Pseudo-Euclidean spaces* are the analog when the inner product is indefinite, though still nondegenerate. They share many similarities with Euclidean spaces but also have important differences. Minkowski spaces[1] are a special case, having index $q = 1$, and offer a few specific features. (The *index* is the maximal dimension of a negative definite subspace.)

Historically, Minkowskian and pseudo-Euclidean geometry (and their differential extensions, Lorentzian and pseudo-Riemannian geometry) rose after the discovery of the theory of relativity in the late 19th and early 20th century by Hendrik Lorentz, Henri Poincaré, and Albert Einstein. In special relativity, Minkowski space is the model for “spacetime”; it is the solution of Einstein’s equations in a vacuum.

For students in mathematics, it is not necessary to understand—or even be aware of—the physics side of the story in order to learn Minkowski spaces and hyperbolic geometry, although it is palpable in some of the terminology introduced in this chapter: *spacelike* and *timelike* vectors, *light cone*, etc. Moreover, in Chapter 6 (which can safely be skipped), we will see direct connections between hyperbolic geometry and the theory of relativity.

For presentation purposes, I decided to include a separate chapter on Minkowski spaces (Chapter 4), but it is little more than a summary of definitions and results of this chapter specialized to $q = 1$. A benefit of this presentation is that the reader in a hurry may skip this chapter and refer back to it whenever needed, although I do recommend reading it throughout.

As a prerequisite for this chapter, we assume that the reader is familiar with abstract linear algebra and Euclidean vector spaces: orthogonality, linear isometries, etc.

[1]Some authors such as A. Thompson [Tho96] call *Minkowski space* a finite-dimensional real vector space equipped with a norm. I am under the impression that this choice of terminology is marginal and misleading.

3.1 Symmetric bilinear forms

Readers are expected to be partly familiar with the contents of this section; for this reason we keep it fairly condensed. (A good reference for a more detailed treatment is Marcel Berger's *Geometry*[2].)

3.1.1 Symmetric bilinear forms

Let V be a vector space over a field $\mathbb{K}$. Soon we will fix $\mathbb{K} = \mathbb{R}$, but in later chapters we will occasionally encounter complex vector spaces.

A ***bilinear form*** on V is a function $b\colon V \times V \longrightarrow \mathbb{K}$ that is linear in each entry: the two functions $b(\cdot, v_0)$ and $b(u_0, \cdot)$ must be linear forms $V \longrightarrow \mathbb{K}$ for any fixed u_0 and v_0 in V. It is called ***symmetric*** if $b(v,u) = b(u,v)$ for all $u, v \in V$.

The ***quadratic form*** associated to b is the function $q\colon V \longrightarrow \mathbb{K}$ defined by $q(v) = b(v,v)$. Clearly, q is uniquely determined by b; the converse is also true by the ***polarization formula:***

$$b(u,v) = \frac{1}{2}\left(q(u+v) - q(u) - q(v)\right) .$$

This identity is simply obtained by writing $b(u+v, u+v) = b(u,u) + 2b(u,v) + b(v,v)$. Note that to conclude, we need to assume that 2 is invertible, i.e. $\mathbb{K}$ is not a field of characteristic 2.

If V is finite-dimensional and equipped with a basis $(e_1, \dots, e_n)$, a vector $u = \sum_{k=1}^{n} u_k e_k$ is represented by the column vector $U = [u_1, \dots, u_n]^\mathsf{T}$. The bilinear form b has a ***matrix representation*** $B = \left[b_{ij}\right]_{1 \leqslant i,j \leqslant n} \in \mathrm{M}(n, \mathbb{K})$ defined by $b_{ij} = b(e_i, e_j)$. If $u, v \in V$ are represented by column vectors U, V, then $b(u,v)$ can be computed as the matrix product $b(u,v) = U^\mathsf{T} B V$. The bilinear form b is symmetric if and only if B is a symmetric matrix i.e. $B^\mathsf{T} = B$.

Remark 3.1. The choice of a basis $(e_1, \dots, e_n)$ is equivalent to an isomorphism $V \xrightarrow{\sim} \mathbb{K}^n$: the vector with coordinates $(u_1, \dots, u_n)$ is identified to the corresponding element of $\mathbb{K}^n$. The matrix representation of b identifies it as a symmetric bilinear form on $\mathbb{K}^n$.

Example 3.2. The determinant on $V = \mathbb{R}^2$ is defined by $\det((x_1, y_1), (x_2, y_2)) = x_1 y_2 - x_2 y_1$. Is it a bilinear form? Write its matrix representation in the canonical basis of $\mathbb{R}^2$. Is it symmetric?

If $(e'_1, \dots, e'_n)$ is a new basis, the change of basis is encoded by the ***transition matrix*** $P \in \mathrm{GL}(n, \mathbb{K})$, whose k-th column is e'_k represented in the old basis. If U and U' are the representations of $u \in V$ in the old and new bases, then $U = PU'$. It follows that the representation of b in the new basis is $B' = P^\mathsf{T} B P$. The matrices B and B' are called ***congruent***.

Example 3.3. Let $q(x,y) = xy$. Show that q is a quadratic form on $\mathbb{R}^2$ and give the associated symmetric bilinear form. Compute its matrix representation in the canonical basis of $\mathbb{R}^2$, and then in the basis (e'_1, e'_2) with $e'_1 = (1,1)$ and $e'_2 = (-1,1)$.

[2]Chapter 13 in vol. II of [Ber09] (English) or [Ber77] (French). The French edition was re-edited in 2016 [Ber16] but it seems scarcely available.

One can associate to a bilinear form b two maps $b_{\mathrm{L}}, b_{\mathrm{R}} \colon V \longrightarrow V^*$ defined by $b_{\mathrm{L}}(u) = b(u,\cdot)$ and $b_{\mathrm{R}}(v) = b(\cdot, v)$. (This technique is called ***currying***, especially in computer science.) The bilinear form b is symmetric if and only if $b_{\mathrm{L}} = b_{\mathrm{R}}$, in which case their common kernel is called the ***radical*** or ***kernel*** of b, denoted $\ker b$. It is furthermore called ***nondegenerate*** if it has trivial radical, equivalently b_{L} is injective, and a ***perfect pairing*** when b_{L} is an isomorphism $V \xrightarrow{\sim} V^*$. When V is finite-dimensional, the ***rank*** of b is defined as the rank of b_{L} (or b_{R}), and is equal to the codimension of $\ker b$ by the rank-nullity theorem; in particular, b is nondegenerate if and only if it is a perfect pairing.

Example 3.4. On $V = \mathbb{K}^2$, the (bilinear form with) quadratic form $q(x,y) = 2xy$ is nondegenerate: its polarization is $b((x_1,y_1),(x_2,y_2)) = x_1y_2 + x_2y_1$, which is zero for all (x_2,y_2) only if $(x_1,y_1) = 0$. On the other hand, $q(x,y) = (x-y)^2$ is degenerate: the polarization is $b((x_1,y_1),(x_2,y_2)) = (x_1-y_1)(x_2-y_2)$, and its radical is the line $x = y$.

Example 3.5. A sophisticated example, for readers familiar with de Rham cohomology: On a (compact, connected, oriented) manifold, the exterior product of differential forms induces a perfect pairing $\mathrm{H}^k \times \mathrm{H}^{n-k} \longrightarrow \mathrm{H}^n \approx \mathbb{R}$. This is a difficult result that can be proven with Hodge theory or the de Rham isomorphism. It yields the *Poincaré duality* $\mathrm{H}^{n-k} \xrightarrow{\sim} (\mathrm{H}^k)^* \approx \mathrm{H}_k$.

If $f \colon V \longrightarrow W$ is a linear map and b is a bilinear form on W, the ***pullback*** of b by f is the bilinear form on V defined by $(f^*b)(v_1,v_2) = b(f(v_1), f(v_2))$. When $W = V$ and $f \in \mathrm{GL}(V)$, b and f^*b are called ***equivalent***. In a basis of V, if the matrix representations of b and f are B and P respectively, then the matrix of f^*B is $P^{\mathsf{T}} B P$. In particular, two bilinear forms are equivalent if and only if their matrix representations are congruent. (See Exercise 3.3.)

3.1.2 Orthogonality

Assume V is equipped with a symmetric bilinear form b, which we also denote $\langle\cdot,\cdot\rangle$. Two vectors u, v are called ***(b-)orthogonal*** if $\langle u, v\rangle = 0$. When this happens, we write $u \perp v$.

Remark 3.6. This terminology and notations are typically reserved for b nondegenerate (and maybe $\mathbb{K} = \mathbb{R}$), which we will assume soon enough.

More generally, when A, B are subsets of V, we say that A and B are orthogonal and write $A \perp B$ when $u \perp v$ for all $u \in A$ and $v \in B$. The largest subset of V orthogonal to A is denoted $A^\perp$, it consists of all $v \in V$ such that $v \perp A$, and is a vector subspace.

Remark 3.7. Check that the radical of b is $V^\perp$.

A vector v is called ***isotropic*** if $v \perp v$, in other words $q(v) = 0$. The set of isotropic vectors is the ***isotropic cone*** (of b, or q). The radical of b is always contained in the isotropic cone, but the converse is false in general. If b has a trivial isotropic cone (only contains 0), it is called ***definite*** (and ***indefinite*** otherwise). This is a very strong property: it is equivalent to b being nondegenerate in restriction to any subspace. See Exercise 3.2.

Remark 3.8. The isotropic cone is a ***(linear) cone*** in the sense that it is invariant by scalar multiplication, i.e. it is a union of vector lines. It is also a ***quadric***, as the set of solutions of the quadratic equation $q = 0$. This will be discussed in more detail in § 7.5.

Example 3.9. Let $V = \mathbb{R}^3$. The isotropic cone of $q(x, y, z) := x^2 + y^2 + z^2$ is trivial; the isotropic cone of $q(x, y, z) := x^2 + y^2 - z^2$ is the cone $x^2 + y^2 = z^2$ shown in Figure 3.1; the isotropic cone of $q(x, y, z) := xy$ is the union of the two vector planes $x = 0$ and $z = 0$.

Proposition 3.10. *Assume V is finite-dimensional. For any subspace W, $\dim W^\perp \geqslant \operatorname{codim} W$ with equality if b is nondegenerate, and $W \oplus W^\perp = V$ if and only if of $b_{\restriction W}$ is nondegenerate.*

Remark 3.11. It is possible that b is degenerate but not its restriction $b_{\restriction W}$, and vice-versa.

Proof. Consider the map $f : V \to W^*$ defined by $f(v)(w) = b(v, w)$. By the rank–nullity theorem, $\dim\ker f + \dim\operatorname{Im} f = \dim V$. Since $\ker f = W^\perp$ and $\dim\operatorname{Im} f \leqslant W$, we get $\dim W^\perp \geqslant \dim V - \dim W$, as desired. If b is nondegenerate, then b_{L} is surjective, hence f is surjective (any linear form on W can be extended to V), therefore we have equality in the previous argument. (Alternatively, $W^\perp$ is the preimage by b_{L} of the annihilator (also called polar) $W^\circ \subseteq W^*$, and it is a basic fact of linear algebra that $\dim W^\circ = \operatorname{codim} W$.)

For the second statement, observe that $\ker(b_{\restriction W}) = W \cap W^\perp$. Therefore $W \cap W^\perp = \{0\}$ if and only if $b_{\restriction W}$ is nondegenerate. The conclusion then follows from the first statement. ∎

Recall that in a vector space V, any decomposition $V = W_1 \oplus W_2$ allows one to define the ***projection onto*** W_1 ***along*** W_2 by $p(w) = w_1$ where $w = w_1 + w_2$. (It has the properties: $W_1 = \ker p$, $W_2 = \operatorname{Im} p$, and p is ***idempotent***: $p^2 = p$. Conversely, any idempotent linear map is a projection.) If $W_2 = W_1^\perp$, then p is called the ***orthogonal projection*** onto W.

Corollary 3.12. *Assume V is finite-dimensional. The orthogonal projection $p : V \to W$ on a subspace W is well-defined if and only if $b_{\restriction W}$ is nondegenerate. If b is definite, then the orthogonal projection onto any subspace W is well-defined.*

Proof. If b is definite, the restriction of b to any subspace is definite hence nondegenerate. ∎

Theorem 3.13. *Assume V is finite-dimensional and b is a symmetric bilinear form. Then V admits an* ***orthogonal basis****, i.e. a basis whose elements are pairwise orthogonal.*

Proof. If $q = 0$, any basis works. Otherwise, let $e_1 \in V$ such that $q(e_1) \neq 0$. By Proposition 3.10, $e_1^\perp$ is a subspace of codimension 1 such that $V = e_1 \oplus e_1^\perp$. Conclude by induction. ∎

Remark 3.14. There are two famous algorithms to explicitly construct an orthogonal basis:

- The ***Lagrange method*** (also called ***Gauss reduction*** in French, strangely): Write the quadratic form q as a sum of squares of linear forms by eliminating mixed terms. For instance, $q(x, y) = ax^2 + 2bxy + cy^2$ can be rewritten (assuming $a \neq 0$):

$$q(x, y) = a\left[x + \frac{b}{a}y\right]^2 + \left(c - \frac{b^2}{a}\right)y^2 .$$

  Hence the change of coordinates: $x' = x + \frac{b}{a}y$ and $y' = y$.

- The ***Gram–Schmidt process***, which only applies when b is definite. Start with any basis $(v_1, \dots, v_n)$, put $e_1 = v_1$, and project v_{k+1} orthogonally on $\operatorname{span}(e_1, \dots, e_k)$ to obtain e_{k+1}. For instance, if $q(x, y) = ax^2 + 2bxy + cy^2$ in the basis (v_1, v_2), then $e_2 = v_2 - \frac{b}{a} v_1$.

(If you are not acquainted with these methods, it is a good idea to look them up.) The example above shows that when both methods apply, they are dual to each other: both produce the transition matrix $P = \begin{bmatrix} 1 & \frac{-b}{a} \\ 0 & 1 \end{bmatrix}$, but Lagrange yields the new coordinates while Gram–Schmidt yields the new basis. I am surprised not to find this fact spelled out anywhere!

Corollary 3.15. *Assume V is finite-dimensional and let q be a quadratic form. In suitable linear coordinates $(x_1, \dots, x_n)$, q can be written $q(x) = \lambda_1 x_1^2 + \cdots + \lambda_r x_r^2$ where r is the rank of q and $\lambda_1, \dots, \lambda_r \in \mathbb{K}^\times$ are nonzero scalars.*

Proof. Take the coordinate system associated to an orthogonal basis. The matrix representation of b is diagonal, and the number of nonzero entries is the rank of the matrix. ∎

Corollary 3.16. *If V is finite-dimensional and $\mathbb{K} = \mathbb{C}$, any quadratic form can be written $q(x) = x_1^2 + \cdots + x_r^2$ in a suitable coordinate system, and r is the rank of q.*

Corollary 3.17. *If V is finite-dimensional and $\mathbb{K} = \mathbb{C}$, any quadratic form can be written $\underline{q}(x) = x_1^2 + \cdots + x_p^2 - x_{p+1}^2 - \cdots - x_{p+q}^2$ in suitable coordinates, and $p + q$ is the rank of $\underline{q}$.*

Proof of Corollary 3.16 and Corollary 3.17. Let $(e_1, \dots, e_n)$ and $\lambda_1, \dots, \lambda_r$ be the basis and the scalars given by Corollary 3.15. If $\mathbb{K} = \mathbb{C}$, any scalar is a square, so we can put $\lambda_k = \mu_k^2$ and $e'_k = e_k / \mu_k$. If $\mathbb{K} = \mathbb{R}$, put $\lambda_k = \pm \mu_k^2$ depending on the sign of λ_k: call p the number of positive values and q the number of negative ones. Conclude. ∎

3.1.3 Positivity and signature

Henceforth, and for the remainder of the chapter, we assume that $\mathbb{K} = \mathbb{R}$.

A symmetric bilinear form b is ***positive (semi)definite*** [resp. ***negative (semi)definite***] if the associated quadratic form q takes (semi)positive [resp. (semi)negative] values on $V - \{0\}$. (We use *semipositive* as a synonym for *nonnegative*, i.e. positive or zero.) A vector subspace $W \subseteq V$ is called *(semi)positive* [resp. *(semi)negative*] if the restriction $b_{\restriction W}$ has that quality.

Proposition 3.18. *Assume that b is positive [resp. negative] semidefinite. Then the isotropic cone of b is equal to its radical. Furthermore, it is equivalent for b to be positive or negative definite, or to be definite, or to be nondegenerate.*

Proof. This is a good exercise: try to do it yourself (also see Exercise 3.2). ∎

Theorem 3.19 (Sylvester's law of inertia)**.** *Let V be real vector space of finite dimension n and b be a symmetric bilinear form.*

(i) The dimension p [resp. q] of any maximal positive [resp. negative] subspace is the same, called ***positive index*** *[resp.* ***(negative) index****]. The pair (p, q) is the* ***signature*** *of b.*

(ii) One has $p + q + r = n$ where r is the dimension of the radical of b.

(iii) There exists a basis of V for which the matrix representation of b is the diagonal matrix:

$$I_{p,q} := \left[\begin{array}{c|c|c} I_p & 0 & 0 \\ \hline 0 & -I_q & 0 \\ \hline 0 & 0 & 0 \end{array}\right]$$

(iv) There exists a system of coordinates in which the quadratic form associated to b is:

$$\underline{q}(x) = x_1^2 + \cdots + x_p^2 - x_{p+1}^2 - \cdots - x_{p+q}^2 \,.$$

Conversely, any quadratic form with this property has signature (p, q).

(v) Two symmetric bilinear forms are equivalent if and only if they have same signature.

Proof. (i) Let W_0 be a positive subspace whose dimension p_0 is maximal among all positive subspaces. If W is a maximal positive subspace of dimension p, then $p \leqslant p_0$. On the other hand, $V = W \oplus W^\perp$ (by Proposition 3.10), and $W^\perp$ is seminegative, otherwise W would not be maximal. Since W_0 is positive, it intersects $W^\perp$ trivially, so that $\dim(W_0 + W^\perp) = \dim W_0 + \dim W^\perp = p_0 + (n - p)$. However $\dim(W_0 + W^\perp) \leqslant n$, that is $p_0 \leqslant p$. We conclude that $p = p_0$. Of course, the negative version of this argument also works.

(ii) Let W_+ be a maximal positive subspace. We saw that $W_+^\perp$ is seminegative, and it clearly contains $\ker b$. Consider the restriction $b_{\restriction W_+^\perp}$. It is seminegative, and its kernel is still $\ker b$. By Proposition 3.18, $\ker b$ is also its cone, therefore b is negative on any complementary subspace W_- of $\ker b$ in $W_+^\perp$. We obtain the decomposition

$$V = W_+ \oplus W_- \oplus \ker b$$

where W_+ is maximal positive and W_- is negative. W_- is in fact maximal negative because b is semipositive on the orthogonal complement $W_+ \oplus \ker b$. This proves that $p + q + r = n$.

(iii) Construct a basis of V by concatenating an orthonormal basis of (W_+, b), an orthonormal basis of $(W_-, -b)$, and any basis of $\ker b$. (We assume it known that positive spaces admit orthonormal bases, either by Corollary 3.17, or by the Gram–Schmidt process.)

(iv) It is evident that in the coordinate system associated to that basis, $\underline{q}(x) = x_1^2 + \cdots + x_p^2 - x_{p+1}^2 - \cdots - x_{p+q}^2$. Conversely, let $\underline{q}$ be a quadratic form that can be written like so relative to some basis $(e_1, \ldots, e_n)$. The subspace W_+ [resp. W_-] spanned by $(e_1, \ldots, e_p)$ [resp. $(e_{p+1}, \ldots, e_{p+q})$] is positive [resp. negative] and has dimension p [resp. q], therefore $p \leqslant p_0$ and $q \leqslant q_0$. On the other hand, $p + q$ and $p_0 + q_0$ are both equal to the rank of b. In conclusion we must have $p = p_0$ and $q = q_0$.

(v) Being equivalent is a transitive relation, and (iv) shows that a symmetric bilinear form has signature (p, q) if and only if it is equivalent to $\underline{q}$. ■

We call ***(pseudo-)orthonormal*** (with respect to b, or q) any basis of V as in (iii), as well as the associated system of coordinates as in (iv).

We conclude this section with a proposition that will be key for the hyperboloid model (rendezvous in Chapter 5):

Proposition 3.20. *If W is a positive [resp. negative] subspace of dimension k, then $W \oplus W^{\perp} = V$, and the signature of b on $W^{\perp}$ is $(p-k, q)$ [resp. $(p, q-k)$].*

Proof. We have already seen that $W \oplus W^{\perp} = V$ in Proposition 3.10. It is enough to do the case where W is positive: for the other case, just take $-b$. The same argument as in the proof of Theorem 3.19 shows that we can write $W^{\perp} = U_{+} \oplus U_{-} \oplus \ker b$ with U_{+} maximal positive and U_{-} maximal negative in $W^{\perp}$. Denote by (p_2, q_2) the signature of b on $W^{\perp}$, so that $p_2 = \dim U_{+}$ and $q_2 = \dim U_{-}$. The decomposition $V = W \oplus U_{+} \oplus U_{-} \oplus \ker b$ shows that $k + p_2 + q_2 + r = n$. On the other hand, we know that $p + q + r = n$, moreover $k + p_2 \leqslant p$ since $W \oplus U_{+}$ is positive and $q_2 \leqslant q$ since U_{-} is negative. We conclude that $p = k + p_2$ and $q = q_2$. ∎

3.2 Pseudo-Euclidean spaces

3.2.1 Definition

Let V be a real vector space. Recall that an ***inner product*** on V is a positive definite symmetric bilinear form; when V is finite-dimensional it is then called ***Euclidean***. Similarly:

Definition 3.21. Let V be a finite-dimensional real vector space. A ***pseudo-inner product*** on V is a nondegenerate symmetric bilinear form $b = \langle \cdot, \cdot \rangle$. Equipped with a pseudo-inner product, V is called a ***pseudo-Euclidean vector space***.

Remark 3.22. We do not rule out Euclidean spaces, but one could. We say that V (or b) has ***mixed signature*** when b is indefinite, i.e. has signature (p, q) with $p > 0$ and $q > 0$.

Remark 3.23. A ***Euclidean space*** is either a Euclidean vector space or an affine space modelled on one. We retain the same convention for a ***pseudo-Euclidean space***. By definition, a pseudo-Euclidean affine (sub)space has the same signature as its underlying vector space.

Example 3.24. Let $V = \mathbb{R}^n$. Choose two integers $p, q \geqslant 0$ such that $n = p + q$. Consider the symmetric bilinear form:

$$\langle x, y \rangle = x_1 y_1 + \cdots + x_p y_p - x_{p+1} y_{p+1} - \cdots - x_n y_n \,. \tag{3.1}$$

By Theorem 3.19 (iv), it has signature (p, q), and by (ii) it is nondegenerate. The space $(\mathbb{R}^n, \langle \cdot, \cdot \rangle)$ is the ***canonical pseudo-Euclidean space of signature*** (p, q), denoted $\mathbb{R}^{p,q}$.

Example 3.25. A pseudo-Euclidean space of negative index $q = 1$ is called a ***Minkowski space***. The ***canonical Minkowski space*** of dimension $n+1$ is $\mathbb{R}^{n,1}$. We will study Minkowski spaces specifically in Chapter 4.

3.2.2 Spacelike, timelike, lightlike

This terminology originates in physics, of course (rendezvous in Chapter 6). Instead of using the adjectives *positive*, *negative*, and *isotropic* for nonzero vectors, it is customary to say ***spacelike***, ***timelike***, and ***lightlike*** (or ***null***).

The isotropic cone is also called the ***light cone***. See Figure 3.1 for an illustration of the light cone in the Minkowski space $\mathbb{R}^{2,1}$.

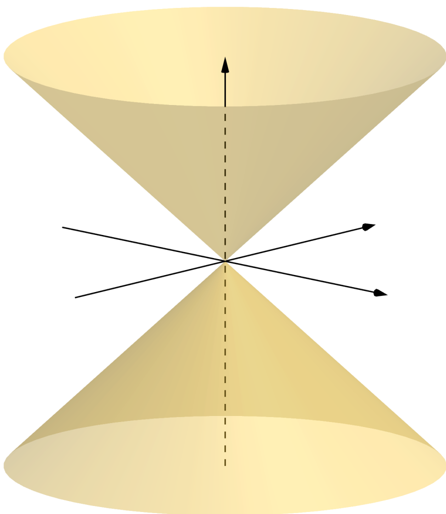


Figure 3.1: The light cone in Minkowski space $\mathbb{R}^{2,1}$.

Remark 3.26. For linear (or affine) subspaces, the most common convention is: ***spacelike*** = positive, ***timelike*** = contains a timelike vector, and ***lightlike*** = semipositive but not positive. In addition, ***null*** = isotropic, and ***causal*** = seminegative. The spacelike/timelike/lightlike quality of a vector or subspace is called its ***causal character***. Note that a vector has the same causal character as its linear span; for this reason the zero vector is considered spacelike. (See for instance [SW77], [ONe83], [Chr19].)

Remark 3.27. This terminology is especially relevant for a Minkowski space, but it is harmless to use it for any pseudo-Euclidean space. In physics, a Minkowski space is usually called a ***Minkowski spacetime***, and a pseudo-Euclidean space of signature (p, q) a ***(p+q)-spacetime***.

3.2.3 Classification

Recall that a basis $(e_1, \ldots, e_n)$ of a pseudo-Euclidean space of signature (p, q) is ***orthonormal*** if $\langle e_k, e_l\rangle = 0$ for $k \neq l$, $\langle e_k, e_k\rangle = 1$ for $1 \leqslant k \leqslant p$, and $\langle e_k, e_k\rangle = -1$ for $p + 1 \leqslant k \leqslant p + q$.

Theorem 3.28. *Any pseudo-Euclidean vector space V admits an orthonormal basis. Its number of spacelike [resp. timelike] elements is the positive [resp. negative] index of V.*

Proof. This is a rephrasing of Theorem 3.19 (iii). ∎

We will study pseudo-Euclidean isometries in § 3.7, but let us give a definition now. A map $f\colon (V,b) \to (V',b')$ between two pseudo-Euclidean vector spaces is called a ***linear isometry*** if f is a linear isomorphism and $f^*b' = b$ (see § 3.1.1 for the definition of pullback). Concretely, $\langle f(u), f(v)\rangle_{V'} = \langle u,v\rangle_V$ for any $u, v \in V$. When such an isometry exists, V and V' are called ***isomorphic*** (as pseudo-Euclidean vector spaces) or simply ***isometric***.

Theorem 3.29. *Two pseudo-Euclidean vector spaces are isometric if and only if they have the same signature. Any pseudo-Euclidean space of signature (p,q) is isometric to $\mathbb{R}^{p,q}$.*

Proof. This is again essentially a rephrasing of Sylvester. Let V be a pseudo-Euclidean vector space of signature (p,q). By Theorem 3.19 (iv), there exist coordinates in which the pseudo-inner product is like (3.1). This coordinate system defines an isomorphism $V \xrightarrow{\sim} \mathbb{R}^{p,q}$. The first statement follows from the second and the uniqueness of the signature. ■

Remark 3.30. Theorem 3.29 is easily extended to pseudo-Euclidean affine spaces.

3.3 Pseudo-Euclidean spheres

Pseudo-Euclidean spheres are the analog of Euclidean spheres. The first notable difference is that a of negative "square radius" is generally not empty.

Definition 3.31. Let V be a pseudo-Euclidean vector space. Let $m \in \mathbb{R}$. The subset $S_m \subseteq V$ defined by $S := \{v \in V \mid \langle v,v\rangle = m\}$ is called a ***(pseudo-)sphere*** and m is its ***square radius***.

Remark 3.32. In other words, S_m is the m-level set of the quadratic form $v \mapsto \langle v,v\rangle$.

Remark 3.33. More generally, in a pseudo-Euclidean space E, the ***sphere with center*** $O \in M$ ***and square radius*** $m \in \mathbb{R}$ is the set $S_{O,m} := \{M \in E \mid \langle \overrightarrow{OM}, \overrightarrow{OM}\rangle = m\}$. Using affine notation, this is simply put: $S_{O,m} = O + S_m$.

Example 3.34. Let $V = \mathbb{R}^{2,1}$. The m-pseudosphere in V is the subset of $\mathbb{R}^3$ with equation $x^2 + y^2 - z^2 = m$. For $m > 0$, this is a connected surface called ***hyperboloid of one sheet***, and for $m < 0$ it is a disconnected surface called ***hyperboloid of two sheets***. For $m = 0$, it is the isotropic cone of $\mathbb{R}^{2,1}$, also called ***light cone***. See Figure 3.2.

We list the main features of pseudo-Euclidean spheres in two theorems:

Theorem 3.35. *Let V be a pseudo-Euclidean vector space of mixed signature (p,q). Let $S = S_m$ be the sphere with square radius m in V where $m \in \mathbb{R}^\times$ is any nonzero real number.*

(i) S is nonempty.

(ii) S is closed in V but not compact. It is connected unless ($q = 1$ and $m < 0$) or ($p = 1$ and $m > 0$), in which case it has two connected components.

(iii) For any subspace $W \subseteq V$, the sphere of square radius m in W is $S \cap W$.

(iv) S is preserved by any linear isometry $f\colon V \to V$.

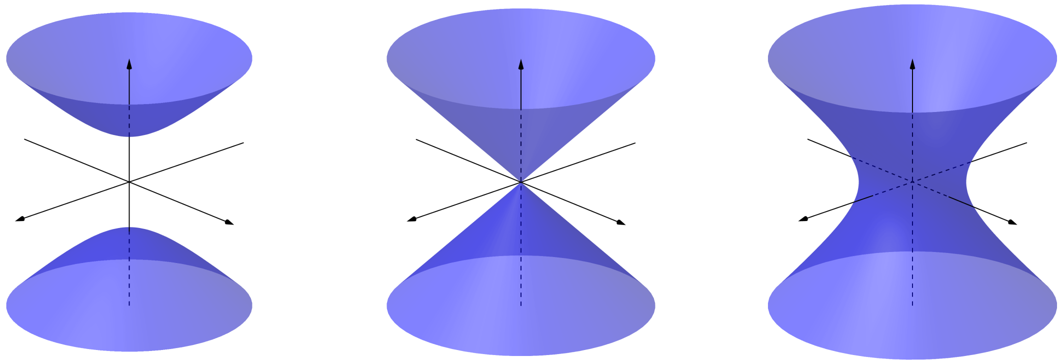


Figure 3.2: Spheres in $\mathbb{R}^{2,1}$: a two-sheeted hyperboloid ($m < 0$), the light cone ($m = 0$), a one-sheeted hyperboloid ($m > 0$).

Theorem 3.36. *In the same setup as the previous theorem:*

(i) S is a proper quadric.
(ii) S is a smooth hypersurface.
(iii) The linear tangent space to S at a point $v \in S$ is the hyperplane $\mathrm{T}_v S = v^\perp$.
(iv) The signature of any tangent space to S is $(p-1, q)$ if $m > 0$ and $(p, q-1)$ if $m < 0$.

Remark 3.37. Both theorems are easily extended to any sphere $S_{O,m}$ in a pseudo-Euclidean space (see Remark 3.33). The linear tangent space at $M \in S$ is then written $\mathrm{T}_M S = \big(\overrightarrow{OM}\big)^\perp$.

Remark 3.38. Both theorems also apply for a Euclidean sphere, with the same proof. The only problem with allowing $p = 0$ or $q = 0$ is that S can be empty if m has the wrong sign.

Remark 3.39. Theorem 3.36 (i) (and its proof) can be ignored until after reading Chapter 7.

Proof of Theorem 3.35. (i) If $m > 0$, take any spacelike vector v, which exists because $p > 0$, and scale it so that $\langle v, v\rangle = m$. Proceed similarly if $m < 0$.

(ii) S is closed because it is the preimage of $\{m\}$ by the quadratic function $v \mapsto \langle v, v\rangle$. It is not compact because it is unbounded with respect to a Euclidean norm on V. To see this, let e_+ [resp. e_-] be a unit spacelike [resp. timelike] vector. The intersection of S with the plane $\mathrm{span}(e_+, e_-)$ is the hyperbola with equation $x^2 - y^2 = m$.

The fact that S has one or two connected components as stated is fairly easy to see but lengthy to argue properly. I suggest you try to figure it out on your own instead of reading further (start by looking at Figure 3.2).

Let us first reduce the number of dimensions. By Theorem 3.19 (iv), in suitable coordinates S is given by $x_1^2 + \cdots + x_p^2 - x_{p+1}^2 \cdots - x_n^2 = m$. For any $j \in \{2, \ldots, p\}$, any Euclidean rotation (centered at the origin) in the x_1x_j-plane preserves $x_1^2 + x_j^2$. It follows that this rotation, while leaving all other coordinates unchanged, preserves S. If $x \in S$ is a point on the sphere with $x_j \neq 0$, one can always apply such a rotation r_θ to obtain $x_j = 0$. The path $t \in [0, \theta] \mapsto r_t(x)$

is a continuous path in S that takes x to the intersection of S with the hyperplane $x_j = 0$. This proves that S does not have more connected components than $S \cap \{x_j = 0\}$. By applying the same argument repeatedly, we can kill off all the positive coordinates except x_1, and similarly kill all the negative coordinates except x_{p+1}. We are left with the hyperbola $x_1^2 - x_{p+1}^2 = m$, which has two branches, therefore S has at most two connected components.

If $p \geqslant 2$ and $m > 0$, one can take one step back (put back the x_2 dimension) to realize that S is actually connected: the two branches of the hyperbola $x_1^2 - x_{p+1}^2 = m$ are connected by the rotations r_t with $t = 0 \to \pi$ inside $x_1^2 + x_2^2 - x_{p+1}^2$. Visually: the hyperboloid of one sheet shown in Figure 3.2 is clearly connected. Similarly, S is connected if $q \geqslant 2$ and $m < 0$.

Finally, let us prove that S has at least two connected components if $q = 1$ and $m < 0$ [resp. $p = 1$ and $m > 0$]. Let v_0 be any vector $\in S$ (we have seen this exists), then $-v_0$ is also in S. The hyperplane $H := v_0^\perp$ is positive [resp. negative] by Proposition 3.20; since $m < 0$ [resp. $m > 0$], H does not intersect S. The complement $V - H$ has two connected components: $H_\pm = \{v \in V : \pm \langle v, v_0 \rangle > 0\}$; one contains v_0 and the other $-v_0$. In conclusion, S is a subset of the disjoint union of two open sets, H_+ and H_-, but is not contained in either: this is the definition of a disconnected set.

(iv) By definition, the pseudo-inner product in W is the restriction of V's. Conclude.

(iii) If $f : V \to V$ is an isometry, then $\langle f(v), f(v) \rangle = \langle v, v \rangle$ for any $v \in V$, therefore $f(v) \in S_m$ if and only if $v \in S_m$. ∎

Proof of Theorem 3.36. Let $f : V \to \mathbb{R}$ be the function defined by $f(v) = \langle v, v \rangle - m$, so that S is the zero set of f.

(i) In coordinates, f is a polynomial function of degree 2, therefore S_m is a quadric. We have seen that it is nonempty in Theorem 3.35 (i). It remains to show that S is nondegenerate, i.e. its projective completion is nondegenerate. The homogenization of f on $V \times \mathbb{R}$ is easy to figure out: $\hat{f}(v, t) = \langle v, v \rangle - mt^2$. This is a quadratic form of signature $(p, q+1)$ on $V \times \mathbb{R}$, therefore it is nondegenerate by Theorem 3.19 (ii).

(ii) The function $f(v) = \langle v, v \rangle - m$ is differentiable at any $v \in V$ with $(\mathrm{d}f)_v = 2\langle v, \cdot \rangle$. If $v \in S$, then $(\mathrm{d}f)_v$ is not the zero linear form, since $(\mathrm{d}f)_v(v) = 2m \neq 0$. This shows that f is a *submersion* on S by definition. Recall moreover that S is a level set of f (the zero level set). It is a classical theorem of differential geometry that in this scenario, S is a smooth hypersurface (a smooth submanifold of codimension one), and the linear tangent space to S at v is $\ker(\mathrm{d}f)_v$.

(iii) Since $\mathrm{T}_v S = \ker(\mathrm{d}f)_v$ and $(\mathrm{d}f)_v = 2\langle v, \cdot \rangle$, we find $\mathrm{T}_v S = \ker\langle v, \cdot \rangle = v^\perp$.

(iv) We just saw that $\mathrm{T}_v S = v^\perp$. If $m > 0$, then v is spacelike, therefore $v^\perp$ has signature $(p-1, q)$ by Proposition 3.20. Conclude similarly if $m < 0$.

∎

3.4 Distances and lengths

3.4.1 Norm and distance

There is no sensible definition of a genuine norm in a pseudo-Euclidean vector space. We can try $\|v\| := \sqrt{|\langle v, v\rangle|}$, but this "norm" has problems: $\|v\| = 0$ does not imply $v = 0$; worse, the triangle inequality $\|u+v\| \leqslant \|u\| + \|v\|$ is not always satisfied. For this reason, it is not even a ***seminorm***. Bearing this in mind, we still call $\|v\| = \sqrt{|\langle v, v\rangle|}$ the ***(pseudo-)length*** of v.

For the same reasons, in a pseudo-Euclidean affine space, one can try define a "distance" $d(A, B) := \|B - A\|$, but it runs into the problems that $d(A, B) = 0$ does not imply $A = B$, and it does not verify the triangle inequality. The latter defect says that it is not even a ***pseudometric***. Let us mention that, from the physical point of view of "spacetime":

- When $u = B - A$ is spacelike, then $d(A, B)$ can be interpreted as a spatial distance (although the perceived distance between A and B differs for most observers).
- When $u = B - A$ is timelike, then $d(A, B)$ represents a *(proper) time interval.*

This will be explained in Chapter 6.

3.4.2 Lengths of curves

As in the Euclidean case, one can extend the notion of length to curves:

Definition 3.40. Let E be a pseudo-Euclidean space. The ***(pseudo-)length*** of a curve $\gamma\colon I \to E$ of class $\mathcal{C}^1$ is the nonnegative real number $\ell(\gamma) := \int_I \|\gamma'(t)\|\,\mathrm{d}t$.

Remark 3.41. In the language of differential geometry, one can write $\ell(\gamma) = \int_\gamma \mathrm{d}s$ where $\mathrm{d}s$ is the 1-density on E defined by $\mathrm{d}s(v) = \|v\|$. (It is the ***line element*** induced by the pseudo-Riemannian metric $g = \langle\cdot,\cdot\rangle$ on E.) Indeed, by definition $\int_\gamma \mathrm{d}s = \int_I \gamma^* \,\mathrm{d}s = \int_I \mathrm{d}s(\gamma'(t))\,\mathrm{d}t$.

A differentiable curve γ is called ***spacelike*** [resp. ***timelike***, resp. ***lightlike***] if $\gamma'(t)$ is spacelike [resp. timelike, resp. lightlike] for all t. Notice that if γ is lightlike, then $\ell(\gamma) = 0$. As in the Euclidean setting, the length of a spacelike or timelike curve is invariant by reparametrization, and one can always reparametrize a spacelike or timelike curve so that its speed $\|\gamma'(t)\| = \frac{\mathrm{d}s}{\mathrm{d}t}$ is constant or even equal to 1. (In physics, a timelike curve is also called a ***world line*** and the "arclength parameter" s is called ***proper time*** and often denoted τ.)

Recall that γ is called ***regular*** if γ' is nonvanishing, and we call γ a ***geodesic*** if $\gamma'' = 0$, equivalently $\gamma(t) = A + ut$ for some $A \in E$ and $u \in V = \vec{E}$, in other words γ is an affine parametrization of a straight line. Similarly to the Euclidean setting (§ A.3.3), one shows:

Theorem 3.42. *Let $\gamma\colon I \to E$ be a smooth regular curve in a pseudo-Euclidean space E. The following are equivalent:*

(i) γ is a geodesic.

(ii) γ is a critical point of the ***energy functional*** $\mathcal{E}(\gamma) := \dfrac{1}{2}\displaystyle\int_I \langle\gamma'(t), \gamma'(t)\rangle\,\mathrm{d}t.$

Furthermore, if γ is assumed spacelike or timelike:

(iii) γ has constant speed and is a crit. point of the ***length functional*** $\ell(\gamma) = \displaystyle\int_I \|\gamma'(t)\| \, \mathrm{d}t$.

Proof. The proof is the same as in the Euclidean setting (see § A.3.3). The only modification we need is: how to conclude that $\gamma'' = 0$ knowing that $\int_I \langle \gamma''(t), X(t) \rangle \, \mathrm{d}t = 0$ for any infinitesimal variation $X(t)$? In the Euclidean case, we easily win with $X(t) = \gamma''(t)$, but in the pseudo-Euclidean setting this argument fails: what if $\gamma''(t)$ is isotropic? Instead, we conclude with:

Lemma 3.43. *If $\alpha(t)$ is a nonvanishing continuous curve in $\vec{E}$, there exists another continuous curve $\beta(t)$ such that $\langle \alpha(t), \beta(t) \rangle = 1$ for all t.*

To prove this lemma, first observe that for any nonzero $\alpha(t) \in \vec{E}$, there exists some $\beta(t) \in \vec{E}$ such that $\langle \alpha(t), \beta(t) \rangle \neq 0$ (by nondegeneracy of the inner product). After scaling $\beta(t)$, we can get $\langle \alpha(t), \beta(t) \rangle = 1$. The fact that $\beta(t)$ can be chosen continuously when $\alpha(t)$ is continuous is an annoying exercise of topology which we leave to the most diligent readers.

To conclude the proof of the theorem, assume that γ'' does not vanish identically. By continuity, there exists a small interval $J \subseteq I$ where γ'' is nonvanishing. Apply the lemma to the restriction $\alpha''_{\restriction J}$ to obtain a curve $\beta(t)$. Put $X(t) = \rho(t)\beta(t)$ where ρ is a ***bump function*** supported in J (i.e. a smooth function such that $0 < \rho \leqslant 1$ inside J and $\rho = 0$ outside of J; this always exists.) We obtain $\int_I \langle \gamma''(t), X(t) \rangle \, \mathrm{d}t = \int_J \rho(t) \, \mathrm{d}t > 0$: contradiction. ∎

Remark 3.44. The assumption that γ is spacelike or timelike for (iii) is necessary to compute the first variation of the length as in § A.3.3.

Corollary 3.45. *For any $A, B \in E$, the line segment $[A, B]$ is the unique geodesic from A to B (up to parametrization) and is length-minimizing among all C^1 curves from A to B.*

Proof. Let $u := B - A$. Since geodesics are affine parametrizations of straight lines, it is clear that $\gamma_0(t) = A + ut$ is the unique geodesic from A to B up to reparametrization. Its length is $\ell(\gamma_0) = \int_0^1 \|u\| \, \mathrm{d}t = \|u\| = d(A, B)$. To conclude that this is the minimum of the lengths of all curves from A to B is a bit more subtle than in the Euclidean case (§ A.3.3).

If u is lightlike, then $d(A, B) = 0$, so there is nothing to show. If u is spacelike or timelike, we would love to conclude by saying that a minimum of the length (parametrized by arclength) is a critical point of the energy functional, which must be a geodesic by Theorem 3.42. However, this proof is not complete: we need to rule out the possibility that the energy has no minimum. Let us leave this last point for the most advanced readers to reflect on. ∎

Remark 3.46. While a spacelike geodesic is a minimizer of the energy functional, a timelike geodesic has negative energy and is a maximizer! A lightlike geodesic is also a critical point of the energy yet neither a minimum nor maximum (call it a ***saddle point***). In all three cases, a geodesic is a minimizer of the length functional by Corollary 3.45. However it is not true that a curve is a geodesic if and only if it is length minimizing, do you see why? (Exercise 3.4.)

3.5 Angles

The notion of angle is also tricky to extend to pseudo-Euclidean spaces. To understand this, let us first remember how angles are defined in a Euclidean space.

3.5.1 Euclidean angles

Geometrically, an angle (measured in radians) represents an arclength on a unit circle. More precisely, let u and v be two nonzero vectors in a Euclidean vector space V. One can always divide u and v by their length to obtain unit vectors u_0 and v_0 which lie on the unit circle S in the plane $P = \mathrm{span}(u, v)$. The (unoriented) angle $\theta := \measuredangle(u, v)$ is, by definition, the length of the arc between $\vec{u_0}$ and $\vec{v_0}$ on S: see Figure 3.3.

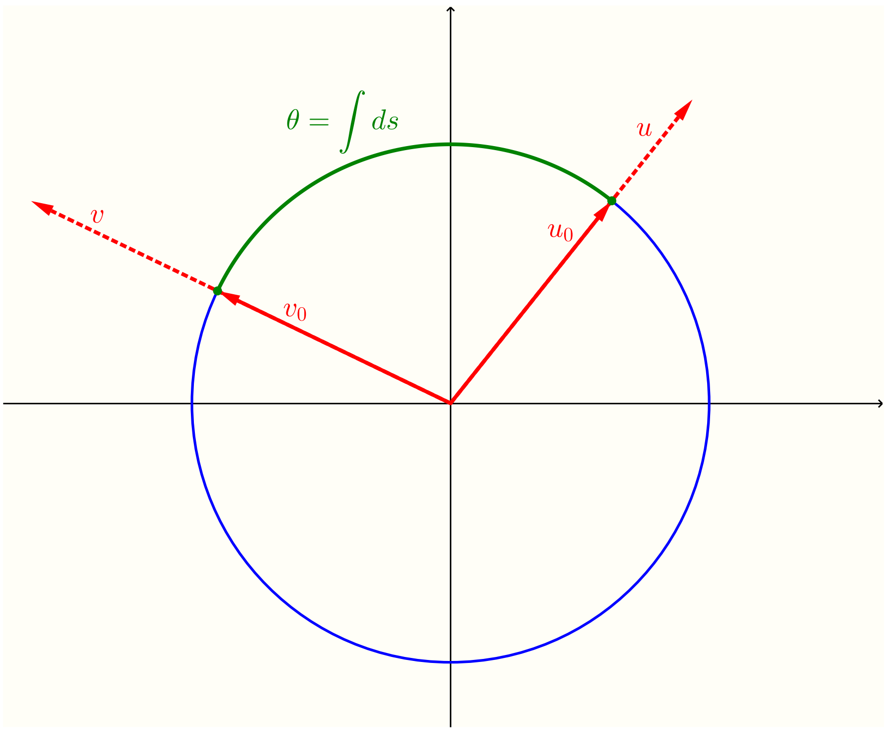


Figure 3.3: Angle between two vectors in a Euclidean plane. *Here* $\mathrm{d}s$ *is the Euclidean line element:* $\mathrm{d}s = \sqrt{\mathrm{d}x^2 + \mathrm{d}y^2}$ *in orthonormal coordinates.*

Alternatively, one can define the angle $\measuredangle(u, v)$ algebraically (we could also say analytically) by observing that it measures the deficit of equality in the ***Cauchy–Schwarz inequality***:

$$|\langle u, v\rangle| \leqslant \|u\| \, \|v\| \, .$$

Indeed, $\theta := \measuredangle(u, v)$ can be defined as the unique real number in $[0, \pi]$ (modulo 2π) such that:

$$\langle u, v\rangle = \|u\| \, \|v\| \cos\theta \, .$$

Note that the Cauchy–Schwarz inequality derives directly from the positive definiteness of the inner product. Do you remember the proof? It is very pretty: the function $p(t) = \langle tu+v, tu+v\rangle = \|u\|^2 t^2 + 2t\langle u,v\rangle + \|v\|^2$ is a second degree polynomial, moreover it is always nonnegative, so its (reduced) discriminant $\Delta/4 = \langle u,v\rangle^2 - \|u\|^2\|v\|^2$ must be nonpositive.

Remark 3.47. Of course, the geometric and the algebraic definitions of angles are equivalent. This coincidence can be interpreted either as a definition of the cosine function, or as a consequence of the celebrated ***Euler's formula*** $e^{i\theta} = \cos\theta + i\sin\theta$. (Can you see why?)

3.5.2 Pseudo-Euclidean angles

Let u and v be two vectors in a pseudo-Euclidean vector space V contained in a vector plane $P \subseteq V$. Similarly to the Euclidean case, we want to define the angle between u and v as the arc length between the matching points on the unit circle in P, but we need to be a bit careful.

First of all, there is *a priori* two unit "circles" in P, namely $S^{\pm} = \{w \in P \mid \langle w,w\rangle = \pm 1\}$. (These are one-dimensional pseudo-Euclidean spheres, see § 3.3.) Henceforth we assume that u and v are both spacelike or both timelike, so that u_0 and v_0 both lie on S^+ or on S^-. In other cases, e.g. when u or v is isotropic, we consider that the angle $\measuredangle(u,v)$ is undefined.

Let us examine the different cases for the signature of the plane P:

- If P is positive definite [resp. negative definite], there is no issue: u_0 and v_0 both lie on the circle S^+ [resp. S^-] (while the other is empty), and the angle $\measuredangle(u,v) = \int_\gamma \mathrm{d}s$ is well-defined similarly to the Euclidean case (see Figure 3.3). We call it a ***circular angle***.
- If P has signature $(1,1)$ (i.e. is a Minkowski plane, see § 4.1.5), then S^+ and S^- are both hyperbolas: $x^2 - y^2 = \pm 1$ in orthonormal coordinates. To define $\measuredangle(u,v)$, we assume that u_0 and v_0 both lie on either S^+ or on S^-, i.e. are either both spacelike or both timelike. Furthermore, u_0 and v_0 must be on the same branch of the hyperbola for there to be an arc joining them. Algebraically, the latter condition says that $\langle u,v\rangle > 0$ if u and v are spacelike, and $\langle u,v\rangle < 0$ if u and v are timelike. In summary, if $\langle u,u\rangle$, $\langle v,v\rangle$, and $\langle u,v\rangle$ are all positive or all negative, then one can define $\measuredangle(u,v) = \ell(\gamma) = \int_\gamma \mathrm{d}s$ as in Figure 3.4. We call it a ***hyperbolic angle***.
- If P is degenerate: this is not the most important case but it is insightful to understand it. Say P has signature $(1,0)$ (the $(0,1)$ case is similar). Then S^- is empty, and S^+ is the union of two lines: in orthonormal coordinates (x,y), the equation of S^+ is $x^2 = 1$, so it is the union of the lines $x = \pm 1$. The unit vectors u_0 and v_0 lie on the same branch provided that $\langle u,v\rangle > 0$, in which case we define $\measuredangle(u,v) = \int_\gamma \mathrm{d}s$ as in Figure 3.5 and call it a ***degenerate angle***. However, since the line element $\mathrm{d}s = |\mathrm{d}x|$ vanishes along $x = \pm 1$, one finds that $\measuredangle(u,v) = 0$: the measure of any degenerate angle is zero!

Remark 3.48. If u and v are linearly independent, then the plane P is of course their linear span. Otherwise, there could be several choices of P with different signatures. Check that, depending on this choice, $\measuredangle(u,v)$ could be either be equal to 0, π, or undefined.

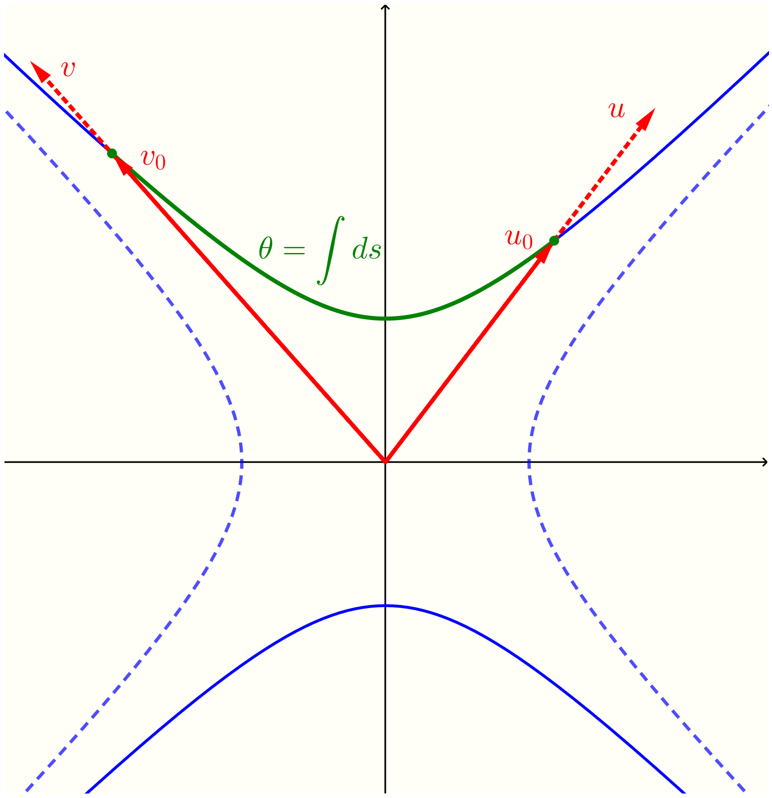


Figure 3.4: Hyperbolic angle between two vectors in a Minkowski plane. *Mind that* $\mathrm{d}s$ *is the pseudo-Euclidean line element:* $\mathrm{d}s = \sqrt{|\mathrm{d}x^2 - \mathrm{d}y^2|}$ *in orthonormal coordinates.*

What about the algebraic definition of angles with Cauchy–Schwarz? The following theorem could be taken as a definition:

Theorem 3.49. *Let $u, v \in V$ and let $P \subseteq V$ be a vector plane containing them. Assume that u and v are both spacelike or both timelike and let $\varepsilon = \pm 1$ indicate their common sign.*

- *If P is positive or negative definite, then Cauchy–Schwarz holds: $|\langle u, v\rangle| \leqslant \|u\|\,\|v\|$. The angle $\measuredangle(u, v)$ is circular and equals the unique real number $\theta \in [0, \pi]$ such that:*

$$\langle u, v\rangle = \varepsilon\,\|u\|\,\|v\|\cos\theta\,.$$

- *If P has mixed signature, then reversed Cauchy–Schwarz holds: $|\langle u, v\rangle| \geqslant \|u\|\,\|v\|$. The angle $\measuredangle(u, v)$ is well-defined when the sign of $\langle u, v\rangle$ is ε, in which case it is a hyperbolic angle and equals the unique real number in $\theta \in [0, +\infty)$ such that:*

$$\langle u, v\rangle = \varepsilon\,\|u\|\,\|v\|\cosh\theta\,.$$

- *If P is degenerate, then the Cauchy–Schwarz equality holds: $|\langle u, v\rangle| = \|u\|\,\|v\|$. The angle $\measuredangle(u, v)$ is well-defined when that the sign of $\langle u, v\rangle$ is ε, in which case it is a degenerate angle and equals $\theta = 0$. Note that we still have:*

$$\langle u, v\rangle = \varepsilon\,\|u\|\,\|v\|\,\mathrm{cos[h]}\,\theta\,.$$

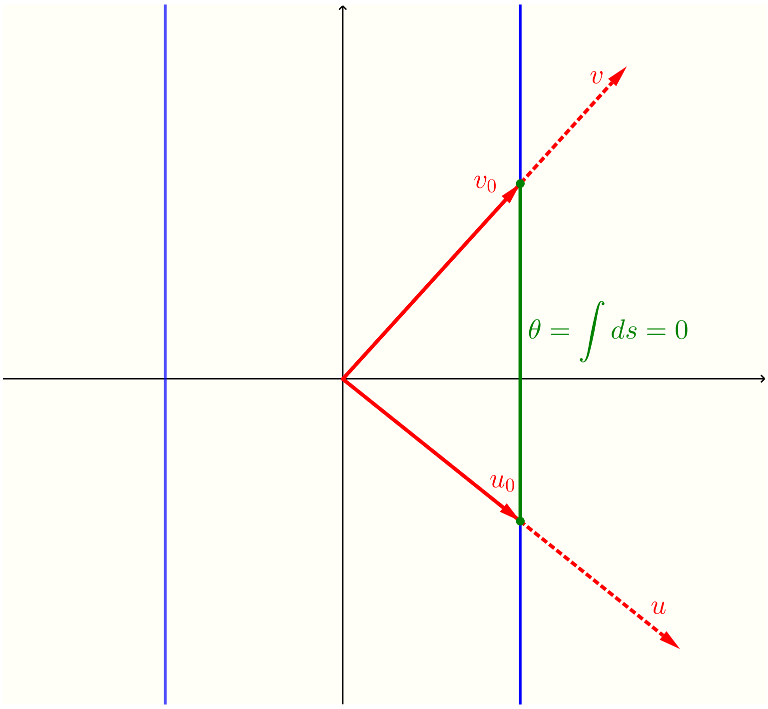


Figure 3.5: Degenerate angle between two vectors in a $(1,0)$ plane. *Here* $\mathrm{d}s$ *is the degenerate line element:* $\mathrm{d}s = |\mathrm{d}x|$ *in orthonormal coordinates. It follows that* $\measuredangle(u,v) = 0$*!*

Proof. Let us assume that u and v are linearly independent, so that the vector plane P is their linear span. (The theorem is quickly checked when u and v are collinear.)

Let $p\colon \mathbb{R} \longrightarrow \mathbb{R}$ be defined by $p(t) = \langle tu + v, tu + v\rangle$. This is a second degree polynomial: $p(t) = \|u\|^2 t^2 + 2t\langle u, v\rangle + \|v\|^2$, and its (reduced) discriminant is $\Delta/4 = \langle u, v\rangle^2 - \|u\|^2\|v\|^2$.

If P is positive definite or negative definite, then the function p never changes sign, so Δ must be nonpositive, i.e. $\langle u, v\rangle^2 \leqslant \|u\|^2\|v\|^2$: this is Cauchy–Schwarz. On the contrary, if P has mixed signature, then $p(t)$ must change sign, so that $\Delta > 0$, and we get reversed Cauchy–Schwarz. If P is degenerate, i.e. of signature $(1,0)$ or $(0,1)$, then p never changes sign as in the first case, but it admits a real root since there exists isotropic vectors. This means that $\Delta = 0$, and we obtain the Cauchy–Schwarz equality.

It remains to compute the angle $\theta = \measuredangle(u,v)$ in the first two cases. First observe that we can always consider the opposite inner product $-\langle\cdot,\cdot\rangle$, so without loss of generality we can assume that u and v are spacelike ($\varepsilon = 1$). We can also harmlessly scale u and v to get unit vectors. Under these assumptions, it remains to show that $\langle u, v\rangle = \cos\theta$ when P is positive definite and $\langle u, v\rangle = \cosh\theta$ when P has mixed signature.

If P is positive definite, then in orthonormal coordinates the inner product is written $\langle (x,y), (x,y)\rangle = x^2 + y^2$. The unit circle $x^2 + y^2 = 1$ can be parametrized by $\gamma(t) = (\cos t, \sin t)$. This is an arclength parametrization: $\|\gamma'(t)\| = \sqrt{\sin^2 t + \cos^2 t} = 1$. Let $u = \gamma(t_1)$ and

$v = \gamma(t_2)$. Without loss of generality, assume that $t_1 < t_2$. By definition, the angle between u and v is $\theta = \int_{t_1}^{t_2} \|\gamma'(t)\| \, \mathrm{d}t = t_2 - t_1$. On the other hand, we have $\langle u, v\rangle = \langle \gamma(t_1), \gamma(t_2)\rangle$, that is $\langle u, v\rangle = \cos(t_1)\cos(t_2) + \sin(t_1)\sin(t_2)$, which indeed coincides with $\cos(\theta) = \cos(t_2 - t_1)$.

If P has mixed signature, then in orthonormal coordinates the inner product is written $\langle (x, y), (x, y)\rangle = x^2 - y^2$. The unit "circle" $x^2 - y^2 = 1$ can be parametrized by $\gamma(t) = (\cosh t, \sinh t)$. This is an arclength parametrization: $\|\gamma'(t)\| = \sqrt{\left|\sinh^2 t - \cosh^2 t\right|} = 1$. Let $u = \gamma(t_1)$ and $v = \gamma(t_2)$. Without loss of generality, assume that $t_1 < t_2$. By definition, the angle between u and v is $\theta = \int_{t_1}^{t_2} \|\gamma'(t)\| \, \mathrm{d}t = t_2 - t_1$. On the other hand, we have $\langle u, v\rangle = \langle \gamma(t_1), \gamma(t_2)\rangle$, that is $\langle u, v\rangle = \cosh(t_1)\cosh(t_2) - \sinh(t_1)\sinh(t_2)$, which indeed coincides with $\cosh(\theta) = \cosh(t_2 - t_1)$. ■

Remark 3.50. Our discussion of angles is deeper than it first appears: we essentially defined (one-dimensional) spherical, hyperbolic, and Euclidean geometries! I invite readers to revisit this section after reading Chapter 5 and Chapter 8 to ponder this claim.

3.6 Space and time orientation

Like any finite-dimensional vector (or affine) space, a pseudo-Euclidean space can be given an orientation. In addition, we will see that it can have a *time orientation* and a *space orientation*.

3.6.1 Orientation of a vector space

Let us briefly recall how the orientation of a finite-dimensional vector space V is defined. Say that a linear automorphism of $f\colon V \to V$ ***preserves*** [resp. ***reverses***] ***orientation*** when it has positive [resp. negative] determinant. Next, say that two bases $(e_1, \dots, e_n)$ and $(e'_1, \dots, e'_n)$ are compatible if the linear automorphism of defined by $f(e_k) = e'_k$ is orientation-preserving. (Equivalently, the transition matrix has positive determinant.) This is an equivalence relation on the set of all bases and there are two equivalence classes.

An ***orientation*** of V consists in picking one class of bases as positively oriented; the others are then negatively oriented. Evidently, there are two choices of orientation. A choice can be simply made by declaring that some given basis is positively oriented. For instance, $\mathbb{R}^n$ has a canonical basis, hence a canonical orientation.

Remark 3.51. Let $\mathrm{GL}^+(V) < \mathrm{GL}(V)$ denote the normal subgroup of linear automorphisms of V with positive determinant. The previous paragraph can be summarized pedantically by saying that an orientation is the choice of a group isomorphism $\mathrm{GL}(V)/\mathrm{GL}^+(V) \xrightarrow{\sim} \{\pm 1\}$.

Remark 3.52. For advanced readers: It is cleanest to define determinants and orientation using tensor algebra. If V is a vector line, an orientation of V is the choice of a direction. If V is n-dimensional, one can reduce to the one-dimensional case by taking the exterior power $\Lambda^n V$.

Remark 3.53. An orientation of V does not induce an orientation of subspaces of V. That being said, if $V = W_1 \oplus W_2$, then an orientation of both W_1 and W_2 determine one for V.

3.6.2 Orientation of space and time

Let V be a pseudo-Euclidean vector space of signature (p, q). An ***orientation of space*** [resp. ***orientation of time***] is a consistent choice of orientation for all maximal positive [resp. negative] subspaces. To see that such a choice is possible, we need to do some work[3].

First we show that there is a canonical way to identify any two maximal subspaces:

Lemma 3.54. *Let W and W' be two maximal positive subspaces. Then the orthogonal projection $p_{W,W'}$ of W onto W' is well-defined, and it is a linear isomorphism.*

Proof. The orthogonal projection $p_{W'} : V \longrightarrow W'$ is well-defined by Corollary 3.12. Since $\dim W = \dim W'$ (by Theorem 3.19), the restriction $p_{W,W'} := (p_{W'})_{\restriction W}$ is bijective if and only if it has trivial kernel. If $p_{W,W'}(v) = 0$, then $v \in W'^{\perp}$, which is a negative subspace by Proposition 3.20. Since W is positive, we must have $v = 0$. ■

The next lemma constructs an interpolation between any two maximal subspaces:

Lemma 3.55. *Let W and W' be two maximal positive subspaces. For any $w \in W$ and $t \in [0, 1]$, put $f(w) := p^{-1}_{W',W}(w) - w$ and $p_t(w) := w + tf(w)$.*

(i) f is a linear map $W \longrightarrow W^{\perp}$.

(ii) Let W_t denote the graph of tf, i.e. the image of p_t. Then $W_0 = W$, $W_1 = W'$, and W_t is a maximal positive subspace for all $t \in [0, 1]$.

Proof. (i) The map f is well-defined and linear; let us show that $f(W) \subseteq W^{\perp}$. Let $w' = f(w) \in f(W)$. Then $p_W(w') = w - p_W(w) = 0$ since $w \in W$. This proves that $w' \in \ker p_W = W^{\perp}$.

(ii) Let q denote the quadratic form $q(v) = \langle v, v\rangle$. Since $w \perp tf(w)$, we have $q(w + tf(w)) = q(w) + t^2 q(f(w)) > 0$ (unless $w = 0$). This shows that W_t is positive definite. Finally, W_t is maximal because it has the same dimension as W, being the graph of tf. ■

The two preceding lemmas show that an orientation of any maximal subspace can be consistently transported to all the others. Let us prove this properly:

Proposition 3.56. *Say that two oriented maximal positive subspaces W and W'* ***have the same orientation*** *if the orthogonal projection of W onto W' is orientation-preserving. This is an equivalence relation on all such oriented subspaces and there are two equivalence classes.*

Proof. The relation is clearly reflexive. Let us show that it is symmetric. This amounts to showing that for any maximal subspaces W and W', the map $h : W \longrightarrow W$ defined by $h = p^{-1}_{W',W} \circ p_{W,W'}$ is orientation-preserving. Denote $h_t : W \longrightarrow W$ the same map as h, only replacing W' by W_t as in Lemma 3.55. Then $\det(h_t)$ depends continuously on t, and it is nonvanishing, therefore it is always positive since $\det(h_0) = \det(\mathrm{id}_W) = 1$. This proves that $h = h_1$ is orientation-preserving. Transitivity is proven similarly: show that $h = p^{-1}_{W_1,W_3} \circ p_{W_2,W_3} \circ p_{W_1,W_2}$ is orientation-preserving by interpolating between W_2 and W_3.

[3] I thank Jérémy Toulisse for helping me figure this out.

It is trivial that there are two equivalence classes: given an oriented maximal subspace W_0, any other W is equivalent to either W_0 or W_0 with the opposite orientation, since $p_{W_0,W}$ is either orientation-preserving or reversing. ■

Remark 3.57. The preceding discussion applies to maximal negative subspaces: just switch "positive" and "negative" everywhere. (Or, to be slick, take the opposite pseudo-inner product.)

Definition 3.58. An ***orientation of space*** [resp. ***of time***] is the choice of one of the two equivalence classes of oriented maximal positive [resp. negative] subspaces.

We emphasize that an orientation of space [resp. of time] is determined by an orientation on any one positive [resp. negative] maximal subspace.

Example 3.59. There is a canonical orientation of space and of time on $\mathbb{R}^{p,q}$: declare that $(e_1, \ldots, e_p)$ [resp. $(e_{p+1}, \ldots, e_{p+q})$] is a positively oriented.

Remark 3.60. There are always $2 \times 2 = 4$ orientations of spacetime: two for space, *and*, independently, two for time. Each configuration induces an orientation of V: choose a maximal positive subspace W, then $W^\perp$ is maximal negative and $W = W \oplus W^\perp$, conclude.

3.7 Isometries

3.7.1 Linear isometries

Linear maps transport symmetric bilinear forms backwards, by ***pullback***: if $f\colon V \to V'$ is a linear map and b' is a symmetric bilinear form on V, then one can define $b = f^*b'$ on V by $b(u,v) := f(b'(u), b'(v))$. (See § 3.1.1 and Exercise 3.3 for more discussion.)

By definition, a linear map $f\colon (V,b) \to (V',b')$ between two pseudo-Euclidean spaces is called ***isometric*** if $f^*b' = b$. Concretely, this means that f ***preserve the inner product***:

$$\forall\, u,v \in V \quad \langle f(u), f(v)\rangle = \langle u,v\rangle\,.$$

Remark 3.61. We abusively use the same notation for the pseudo-inner product in V and V'; we will do the same for the length, distance, etc.

Definition 3.62. A ***linear isometry*** $f\colon V \to V'$ is an isometric linear isomorphism.

Remark 3.63. If $V' = V$, some—I am on to the French—also say ***orthogonal automorphism***.

Example 3.64. Let $V = \mathbb{R}^{1,1}$. The linear map $r_t\colon V \to V$ whose matrix representation is

$$R_t = \begin{bmatrix} \cosh t & \sinh t \\ \sinh t & \cosh t \end{bmatrix}$$

is a linear isometry which we call a ***pseudo-rotation*** or a ***Lorentz boost*** (see § 4.2).

The next proposition is trivial:

Proposition 3.65. *A linear isometry preserves causal characters (of vectors and subspaces), lengths (of vectors and curves), angles (between vectors in a plane), distances (between vectors).*

Proof. All these notions only depend on the pseudo-inner product. ■

Lemma 3.66. *Let $f : V \to V'$ be an arbitrary map between pseudo-Euclidean vector spaces that preserves the inner product. If f is surjective or linear, then it is in fact linear and injective.*

Proof. Let us leave this proof as a semi-elementary exercise, it is a good one. For the case where f is assumed linear, start by observing that if $W \subseteq V$ is a positive [resp. negative] subspace, then $f(W)$ is too and has the same dimension. ■

Remark 3.67. The proof still works if the linear span of $f(V)$ is nondegenerate in V'. However, in contrast to the Euclidean setting, the result may be false without this assumption: consider $f : \mathbb{R} \to \mathbb{R}^{2,1}$ defined by $f(x) = (x, h(x), h(x))$ where h is any function $\mathbb{R} \to \mathbb{R}$.

Proposition 3.68. *Let $f : V \to V'$ be a bijective map between pseudo-Euclidean vector spaces. f is a linear isometry if and only if it preserves the length and the causal character of any vector.*

Proof. It is trivial that if f preserves the inner product, then it preserves lengths and causal characters. The converse is also easy: if $\|f(v)\| = \|v\|$, then $\langle f(v), f(v)\rangle = \pm\langle v, v\rangle$, and the sign is correct because f preserves causal characters. Conclude by polarization (see § 3.1.1). Since f is assumed bijective, it is a linear isomorphism by Lemma 3.66. ■

Remark 3.69. It is not true that any length preserving linear isomorphism is a linear isometry: consider $f : \mathbb{R}^{1,1} \to \mathbb{R}^{1,1}$ defined by $f(x, y) = (y, x)$.

3.7.2 Affine and pseudo-Riemannian isometries

Definition 3.70. Let E and E' be pseudo-Euclidean affine spaces. An affine map $f : E \to E'$ is called an ***affine isometry*** when the underlying linear map $\vec{f} : \vec{E} \to \vec{E}'$ is a linear isometry.

Remark 3.71. As in the Euclidean setting, affine isometries are essentially linear isometries composed by translations. More precisely, given any point $O \in E$, an affine isometry $f : E \to E'$ is uniquely determined by $f(O)$ and $\vec{f}$ by $f(M) = f(O) + \vec{f}\big(\overrightarrow{OM}\big)$. When $E = E'$ is a vector space, a linear isometry is the same thing as an affine isometry that fixes the origin.

There is also a differential version of isometries:

Definition 3.72. A map $f : E \to E'$ is called a ***pseudo-Riemannian isometry*** if f is bijective, of class $\mathcal{C}^1$, and for every $M \in E$ the differential $\mathrm{d}f_{|M} : \vec{E} \to \vec{E}'$ is a linear isometry.

The reader should understand that, *a priori*, the linear isometry $\mathrm{d}f_{|M}$ changes with M. However the next theorem says that any pseudo-Riemannian must be affine, hence $\mathrm{d}f_{|M} = \vec{f}$ is in fact independent of M. I would like to emphasize that this "rigidity" is remarkable:

Remark 3.73. We will see in Chapter 9 that a $\mathcal{C}^1$ map $f: E \to E'$ between Euclidean spaces is *conformal* if, for every $M \in E$, the differential $\mathrm{d}f_{|M}$ is a linear similarity. By analogy, one could expect that any conformal map is in fact an affine similarity. This is far from true in dimension 2: holomorphic functions provide many counter-examples.

Theorem 3.74. *Let $f: E \to E'$ be a bijective map of class $\mathcal{C}^1$ between pseudo-Euclidean spaces. The following are equivalent:*

(i) f is an affine isometry.
(ii) f is a pseudo-Riemannian isometry.
(iii) f is energy preserving: $\mathcal{E}(f \circ \gamma) = \mathcal{E}(\gamma)$ for any $\mathcal{C}^1$ curve $\gamma: I \to E$.
(iv) f preserves causal characters and is length preserving: $\ell(f \circ \gamma) = \ell(\gamma)$.
(v) f preserves causal characters and maps any geodesic to a geodesic with the same speed.
(vi) f preserves causal characters and is "distance" preserving: $d(f(A), f(B)) = d(A, B)$.

Remark 3.75. We say that "f preserves causal character" when $f(B) - f(A)$ as the same causal character as $B - A$ for all $A, B \in E$, which makes sense for any map $f: E \to E'$.

Proof. Thankfully we have already done the hard part of the work in § 3.4.2. We prove the following implications:

$$\begin{tikzcd}
& & (i) \arrow[dll, Rightarrow] & & \\
(ii) \arrow[rr, Rightarrow] \arrow[dr, Rightarrow] & & (iv) \arrow[rr, Rightarrow] & & (vi) \arrow[ull, Rightarrow] \\
& (iii) \arrow[rr, Rightarrow] & & (v) \arrow[ur, Rightarrow] &
\end{tikzcd}$$

(i) $\Rightarrow$ (ii): If f is an affine isometry, then $\mathrm{d}f_{|M} = \vec{f}$ is a linear isometry and it coincides with the differential $\mathrm{d}f_{|M}$ at any point. This proves that f is a pseudo-Riemannian isometry.

(ii) $\Rightarrow$ (iii): By definition, $\mathcal{E}(f \circ \gamma) = \int_I \langle (f \circ \gamma)'(t), (f \circ \gamma)'(t) \rangle \,\mathrm{d}t$. By the chain rule, $(f \circ \gamma)'(t) = \mathrm{d}f_{|\gamma(t)}(\gamma'(t))$, and since f is a pseudo-Riemannian isometry we have

$$\left\langle \mathrm{d}f_{|\gamma(t)}(\gamma'(t)), \mathrm{d}f_{|\gamma(t)}(\gamma'(t)) \right\rangle = \langle \gamma'(t), \gamma'(t) \rangle .$$

It follows that $\mathcal{E}(f \circ \gamma) = \int_I \langle \gamma'(t), \gamma'(t) \rangle \,\mathrm{d}t = \mathcal{E}(\gamma)$.

(ii) $\Rightarrow$ (iv): The proof that $\ell(f \circ \gamma) = \ell(\gamma)$ is the same as for the energy. The fact that $\vec{f}$ preserves causal character is not easily derived from (ii), but it is harmless to derive it from (i) (replace (ii) by (i)+(ii) in the diagram above).

(iii) $\Rightarrow$ (v): Assume f is energy preserving. In particular, a smooth regular curve γ is a critical point of the energy if and only if $f \circ \gamma$ is a critical point of the energy. By Theorem 3.42, this proves that γ is a geodesic if and only if $f \circ \gamma$ is a geodesic. These two geodesics have the same energy if and only if they have the same speed.

(v) $\Rightarrow$ (vi): Let $\gamma\colon [0,1] \longrightarrow E$ be the unique geodesic with $\gamma(0) = A$ and $\gamma(1) = B$: it is given by $\gamma(t) = A + t(B - A)$ where $u = B - A$. Similarly, let Let $\tilde{\gamma}\colon [0,1] \longrightarrow E'$ be the unique geodesic from $f(A)$ to $f(B)$, given by $\tilde{\gamma}(t) = A + t(f(B) - f(A))$. By assumption, we must have $\tilde{\gamma} = f \circ \gamma$, and the speed of $\tilde{\gamma}$ must be equal to the speed of γ, in other words $\|f(B) - f(A)\| = \|B - A\|$, i.e. $d(f(A), f(B)) = d(A, B)$.

(iv) $\Rightarrow$ (vi): This follows from Corollary 3.45.

(vi) $\Rightarrow$ (i): f is distance-preserving if and only if $\vec{f}$ preserves the lengths of vectors. Thus (i) $\Leftrightarrow$ (vi) is the content of Proposition 3.68. ∎

3.7.3 Isometry groups

When $V = V'$, linear isometries form a group under composition, called the ***linear isometry group*** or ***(pseudo-)orthogonal group*** of V, denoted $\mathrm{O}(V)$ (or $\mathrm{O}(q)$, where q is the quadratic form on V associated to the inner product). Let us record this:

Definition 3.76. Let V be a pseudo-Euclidean vector space. The ***orthogonal group*** of V is the group of linear automorphisms of V that preserve the inner product:

$$\mathrm{O}(V) := \{f \in \mathrm{GL}(V) \mid \forall u, v \in V\ \langle f(u), f(v)\rangle = \langle u, v\rangle\}\ .$$

In terms of matrices: choose a basis $(e_1, \dots, e_n)$. Denote $B = \left[\langle e_i, e_j\rangle\right]_{1\leqslant i,j\leqslant n}$ the matrix representation of the inner product and M the matrix of some linear map $f\colon V \longrightarrow V$.

Proposition 3.77. *f is a linear isometry if and only if $M^\mathsf{T} B\, M = B$.*

Proof. Let U and V be the column vectors representing vectors u and v. Then $\langle u, v\rangle = U^\mathsf{T} B\, V$ and $\langle f(u), f(v)\rangle = (MU)^\mathsf{T} B\,(MV) = U^\mathsf{T}\left(M^\mathsf{T} B\, M\right) V$. It follows that f preserves the inner product if and only if $M^\mathsf{T} B\, M = B$. By Lemma 3.66, since f is linear, it is then automatically injective, hence a linear automorphism. ∎

In particular, if we choose an orthonormal basis so that the matrix of the inner product is $I_{p,q}$ as in Theorem 3.19, then the isometry group of V is identified to the group:

$$\mathrm{O}(p, q) := \left\{M \in \mathrm{M}(n, \mathbb{R}) \mid M^\mathsf{T} I_{p,q}\, M = I_{p,q}\right\}\ .$$

called the ***indefinite orthogonal group*** of signature (p, q).

Proposition 3.78. *The orthogonal group of any pseudo-Euclidean vector space of signature (p, q) is isomorphic to $\mathrm{O}(p, q)$.*

Remark 3.79. In $V = \mathbb{R}^{p,q}$, we have a canonical orthonormal basis, therefore $\mathrm{O}(\mathbb{R}^{p,q})$ is canonically identified to $\mathrm{O}(p, q)$. Thus Proposition 3.78 can also be derived from Theorem 3.29.

Affine isometry group.

3.7.4 Preserving space and time orientation

3.7.5 The Cartan–Dieudonné theorem

3.8 The orthogonal group

3.8.1 Algebra

Cartan–Dieudonné theorem

Normal form

Self-adjoint endomorphisms, Polar decomposition

3.8.2 Topology

4 connected components

Maximal compact

3.9 Exercises

Exercise 3.1. Bilinear forms and duality

Let V be a vector space and denote V^* its dual space. Recall that one can associate to any bilinear form b two linear maps $b_{\mathrm{L}}, b_{\mathrm{R}} \colon V \longrightarrow V^*$ defined by $b_{\mathrm{L}}(u) = b(u, \cdot)$ and $b_{\mathrm{R}}(v) = b(\cdot, v)$.

(1) Show that b is symmetric if and only if $b_{\mathrm{L}} = b_{\mathrm{R}}$.

(2) Assume V is finite-dimensional and equipped with a basis. Denote by B the matrix representation of b, and denote B_{L} and B_{R} the matrix representations of b_{L} and b_{R} using the basis of V and the dual basis of V^*. Show that $B = B_{\mathrm{L}} = {B_{\mathrm{R}}}^{\mathsf{T}}$.

(3) Denote $j \colon V \longrightarrow V^{**}$ the natural map of V into its ***bidual*** (or ***double dual***) defined by $j(x)(\varphi) = \varphi(x)$. Check that j is an injective linear map, and is an isomorphism if (*Optional:* and only if) V is finite-dimensional.

(4) We recall that the ***dual*** (or ***transpose***) of a linear map $f \colon V \longrightarrow W$ is the map $f^* \colon W^* \longrightarrow V^*$ defined by $f^*(\psi) = \psi \circ f$. Show that $b_{\mathrm{R}} = b_{\mathrm{L}}^* \circ j$. Recover (2).

(5) Recall why any perfect pairing is nondegenerate, and the converse is true if V is finite-dimensional. What if V is infinite-dimensional?

Exercise 3.2. Cone vs radical and definiteness

Let V be a vector space and let b a symmetric bilinear form.

(1) Show that b is definite (has trivial isotropic cone) if and only if the restriction of b to any subspace is nondegenerate.

(2) Assume $\mathbb{K} = \mathbb{R}$. Show that the isotropic cone of b is equal to its radical if and only if b is positive [or negative] semidefinite.

(3) Characterize definite symmetric bilinear forms when $\mathbb{K} = \mathbb{R}$ and $\mathbb{K} = \mathbb{C}$.

Exercise 3.3. Pullback and equivalence of symmetric bilinear forms

If $f \colon V \longrightarrow W$ is a linear map and b is a bilinear form on W, the ***pullback*** of b by f is the bilinear form on V defined by $f^*b(v_1, v_2) = b(f(v_1), f(v_2))$.

(1) Show that in terms of quadratic forms, the pullback is $f^*q = q \circ f$.

(2) Describe the pullback in terms of matrix representations.

(3) Let $W = V$. Show that pullback defines a right action of $\mathrm{GL}(V)$ on the set of all symmetric bilinear forms on V.

(4) Two symmetric bilinear forms are called ***equivalent*** when they lie in the same $\mathrm{GL}(V)$-orbit. Show that this amounts to their matrix representations being congruent.

(5) Characterize the equivalence classes when $\mathbb{K} = \mathbb{C}$ and $\mathbb{K} = \mathbb{R}$.

Exercise 3.4. Length minimizing curves

Let E be a pseudo-Euclidean space and let $A, B \in E$. Is it true that a $\mathcal{C}^1$ curve γ from A to B is a geodesic if and only if it is length minimizing?

Exercise 3.5. Pseudo-Euclidean similarities

Let V be a pseudo-Euclidean vector space of mixed signature and let $f \in GL(V)$. Prove that the following are equivalent:

(i) There exists $\lambda > 0$ and $g \in \mathrm{O}(E)$ such that $f = \lambda g$.
(ii) f preserves causal characters.
(iii) f preserves angles.

Is it also equivalent to: There exists $\lambda > 0$ such that $\|f(v)\| = \lambda\|v\|$ for all $v \in V$?

It is natural to call a map satisfying (i)–(iii) a ***pseudo-Euclidean (linear) similarity***.

CHAPTER 4

Minkowski space and the Lorentz group

Disclaimer: This chapter is a draft.

4.1 Minkowski spaces

4.1.1 Definition

Minkowski spaces are just a special case of pseudo-Euclidean spaces:

Definition 4.1. A ***Minkowski space*** is a pseudo-Euclidean space of negative index $q = 1$.

In other words, a Minkowski space of dimension $n+1$ is a pseudo-Euclidean space of signature $(n, 1)$. (We assume $n > 0$, although $n = 0$ can be tolerated.)

Remark 4.2. A pseudo-inner product of negative index $q = 1$ is also called ***Lorentzian***[1].

Remark 4.3. A *Minkowski space* either refers to a Minkowski vector space or to an affine space modelled on one, just like a (pseudo-)Euclidean space.

Refer to ?? for the def of spacelike, timelike, lightlike and to Figure 3.1 for pic of light cone.

Remark 4.4. A subspace of a Minkowski space is also a Minkowski space if and only if it is timelike. We naturally call this a ***Minkowski subspace***.

As a specialization of the discussion of pseudo-Euclidean spaces in § 3.2, we obtain:

Theorem 4.5. *Any Minkowski vector space of dimension $n + 1$ admits a coordinate system relative in which the pseudo-inner product is written $\langle x, y\rangle = x_1y_1 + \cdots + x_ny_n - x_{n+1}y_{n+1}$. It is thus isomorphic to $\mathbb{R}^{n,1}$, called the **canonical Minkowski space** of dimension $n + 1$.*

4.1.2 Spheres

The discussion of pseudo-Euclidean spheres in § 3.3 also applies to Minkowski spaces:

Theorem 4.6. *Let $S = S_m$ be the sphere of square radius $m \neq 0$ in a Minkowski space V.*

(i) S is a proper quadric and a smooth hypersurface.

(ii) S restricts to a lower-dimensional sphere on any Minkowski subspace.

(iii) S is preserved by any linear isometry.

(iv) The linear tangent space to S at a point $v \in S$ is the hyperplane $\mathrm{T}_v S = v^\perp$.

In addition, if $m > 0$:

*(v) Unless $\dim V = 2$, S is connected and called a **one-sheeted hyperboloid**.*

[1]This terminology is especially used in differential geometry. A ***pseudo-Riemannian*** (some authors such as [ONe83] say ***semi-Riemannian***) ***metric*** in a manifold is a (smoothly varying) pseudo-inner product in the tangent space. Such a metric is called ***Lorentzian*** when it has negative index $q = 1$.

(vi) Any tangent space to S has negative index 1.

Otherwise, if $m < 0$:

*(v) S has two components and called a **two-sheeted hyperboloid**.*

(vi) Any tangent space to S is positive definite.

Let us also record the specialization of Proposition 3.20 from which (vi) derives:

Proposition 4.7. *Let V be a Minkowski vector space. If $v \in V$ is timelike, then $v^\perp$ is a spacelike hyperplane and $V = \mathbb{R}v \oplus v^\perp$.*

4.1.3 Time orientation

One special feature of Minkowski spaces is that timelike vectors fall into two classes:

Proposition 4.8. *Let V be a Minkowski vector space. The binary relation on timelike vectors: $u \sim v \overset{def}{\Leftrightarrow} \langle u, v\rangle < 0$ is an equivalence relation and there are two equivalence classes.*

Proof. It is a good exercise to try to write your own proof instead of reading further.

The relation is clearly reflexive and symmetric; let us prove that it is transitive. Let u, v, w be three timelike vectors such that $\langle u, v\rangle < 0$ and $\langle v, w\rangle < 0$. Let $v_t = v + tw$; this is a timelike vector for all $t \geqslant 0$, and $\langle u, v_t\rangle = 0$ for $t = -\frac{\langle u,v\rangle}{\langle u,w\rangle} =: t_0$. By Proposition 4.7, v_{t_0} is spacelike, therefore t_0 must be negative i.e. $\langle u, w\rangle < 0$.

We now show that there are two equivalence classes. Fix a timelike vector u. Clearly, u and $-u$ are not in the same equivalence class. The equivalence class of any timelike vector v is determined by the sign of $\langle u, v\rangle$: we have $v \sim u$ if $\langle u, v\rangle < 0$ and $v \sim -u$ if $\langle u, v\rangle > 0$. The case $\langle u, v\rangle = 0$ is ruled out by Proposition 4.7. ∎

Geometrically, the two equivalence classes are the two connected components of the interior of the light cone shown in Figure 3.1. (Indeed, for any timelike vector v, the two components lie on either side of the hyperplane $v^\perp$.)

A ***time orientation*** of a Minkowski space is the choice of one of the two classes of timelike vectors, whose elements are called ***future-pointing***, while the others are ***past-pointing***.

Remark 4.9. The light cone with the origin removed also has two connected components, which are the boundary (minus the origin) of the components of timelike vectors. Lightlike vectors may also be called *future-pointing* or *past-pointing* accordingly.

For the canonical Minkowski space $\mathbb{R}^{n,1}$, we choose the time orientation according to which the future is upwards, more precisely:

Definition 4.10. A causal (i.e. timelike or lightlike) vector $(x_1, \dots, x_{n+1}) \in \mathbb{R}^{n,1}$ is *future-pointing* if $x_{n+1} > 0$, and *past-pointing* if $x_{n+1} < 0$.

If V is oriented (as a vector space), a time orientation induces a ***space orientation***, that is a consistent orientation of all spacelike hyperplanes $H \subseteq V$: a basis of H is declared positively oriented if it can be completed into a positively oriented basis of V by appending a future-pointing vector.

Remark 4.11. For a general pseudo-Euclidean space, it is not possible to consistently define future- and past-pointing vectors; nevertheless it is possible generalize the notion of time and space orientation: see Exercise 4.2.

4.1.4 Timelike angles

Proposition 4.12. *Let u and v be timelike vectors in a Minkowski vector space. Then*

$$\langle u, v\rangle^2 \geqslant \langle u, u\rangle \langle v, v\rangle \tag{4.1}$$

Proof. If u and v are collinear, we easily see that there is equality in (4.1). Otherwise, consider the function $p(t) = \langle u + tv, u + tv\rangle$. This function is negative at $t = 0$, and it cannot have constant sign, otherwise b would be negative definite on the plane spanned by u and v, which is excluded because b has negative index 1. On the other hand, notice that $p(t)$ is a polynomial of degree 2 in t: $p(t) = t^2\langle v, v\rangle + 2t\langle u, v\rangle + \langle u, u\rangle$, and since it does not have constant sign, it must have nonnegative discriminant: $\langle u, v\rangle^2 - \langle u, u\rangle \langle v, v\rangle \geqslant 0$, as desired. ■

4.1.5 The Minkowski plane

Hyperbolic numbers. $\mathrm{SL}_2(\mathbb{C}_\tau)$.

4.2 Lorentz group

When $V = \mathbb{R}^{p,q}$, the orthogonal group is denoted $\mathrm{O}(p, q)$. As a matrix group:

$$\mathrm{O}(b) = \left\{M \in \mathrm{GL}(n, \mathbb{R}) \,:\, M^{\mathsf{T}} I_{p,q}\, M = I_{p,q}\right\}$$

where $I_{p,q}$ is the matrix representation of the pseudo-inner product.

When $q = 1$, i.e. $V = \mathbb{R}^{n,1}$ is a Minkowski space, the group $\mathrm{O}(n, 1)$ is called the ***Lorentz group***.

Isometries of $\mathbb{R}^{p,q}$ can be independently be either space-orientation preserving/reversing and time-orientation preserving/reversing. Unfortunately, I was not able to find a good in general. But in the case of Minkowski space, one can give the following definition:

Definition 4.13. An isometry $f \in \mathrm{O}(n, 1)$ is called ***time-orientation preserving*** [resp. ***time-orientation reversing***] if, for some (equivalently all) timelike vector v, $\langle f(u), v\rangle$ has the same sign (resp. opposite sign) as $\langle u, v\rangle$ for all $u \in V$ (equivalently, for some u not orthogonal to v).

One can proceed to declare that $f \in \mathrm{O}(n, 1)$ is space orientation-preserving if and only if: f is globally orientation-preserving and time orientation-preserving, or globally orientation-reversing and time orientation-reversing. Of course, $f \in \mathrm{O}(n, 1)$ is declared space orientation-reversing in the opposite scenario.

We denote $\mathrm{O}^+(n, 1)$ the subgroup of isometries that are time orientation-preserving (physicists call it the ***orthochronous Lorentz group***). The subgroup $\mathrm{SO}^+(n, 1)$ consists of isometries that are both time and space orientation-preserving.

We will see further properties of $\mathrm{O}(n, 1)$ in Chapter 5, and propose a classification of isometries in Chapter 12.

4.3 Exercises

Exercise 4.1. Orthogonal subspace to a timelike vector

Prove Proposition 4.7 (copied below) directly, without using the results of § 3.1.

Proposition . *Let V be a Minkowski vector space. If $v \in V$ is timelike, then $v^{\perp}$ is a spacelike hyperplane and $V = \mathbb{R}v \oplus v^{\perp}$.*

Exercise 4.2. Time and space orientation

The goal of this exercise is to define the orientation of time and space in a pseudo-Euclidean space, generalizing the case of a Minkowski space (see § 4.1.3).

Let V be a pseudo-Euclidean vector space of mixed signature (p, q).

(1) Show that the binary relation on timelike vectors $u \sim v \overset{def}{\Leftrightarrow} \langle u, v \rangle < 0$ is an equivalence relation if and only if V is a Minkowski space. (So, no good!)

Exercise 4.3. Time orientation-preserving criterion

Let M be a matrix in $\mathrm{O}(n, 1)$. Show that f is time orientation-preserving if and only if the bottom-right coefficient of M is positive.

CHAPTER 5

The hyperboloid model

Disclaimer: This chapter is a draft.

In this chapter, we introduce our first model of hyperbolic space, namely the hyperboloid model. The hyperboloid is a pseudo-Euclidean sphere: it is the unit "sphere" of negative square radius in Minkowski space.

This model is much analogous to the sphere in Euclidean geometry: the hyperboloid, a pseudosphere in Minkowski space, plays the role of the sphere in Euclidean space.

For many purposes, the hyperboloid is the best model of hyperbolic space: we shall see in particular that it is fairly easy and elegant to derive all the relevant geometric properties: the Riemannian metric, group of isometries, geodesics, distance function, and sectional curvature.

Historically, the idea of an imaginary sphere goes back to Lambert in 1766, and in 1826 Taurinus performs trigonometry calculations on a "sphere of imaginary radius". The connection with hyperbolic geometry and the other models was established by Poincaré in the 1880s, and the relation to Minkowski space followed the development of special relativity in the early 20th century. We refer to [Rey93, §14] for a more detailed historical account.

5.1 Description of the hyperboloid

Other pseudo-Euclidean spheres are also interesting pseudo-Riemannian manifolds. Exercise: the hyperboloid is the only Riemannian pseudo-Euclidean sphere. The pseudo-Lorentzian spheres are dS and AdS.

5.1.1 Hyperboloid of dimension 2

Let $M = \mathbb{R}^{n,1}$ be Minkowski space. For the moment, let us take $n = 2$.

Consider the set

$$\mathcal{H} := \{v \in M \colon \langle v, v \rangle = -1\} .$$

By definition, this is a pseudosphere: it is a level set of the quadratic form in the pseudo-Euclidean space M. In other words, abusing notations, this is the “sphere” $\{\|v\|^2 = R^2\}$ in M, with $R = \sqrt{-1}$.

Let us use coordinates $v = (x, y, z)$ on M, so that the quadratic form of Minkowski space is $\langle v, v \rangle = x^2 + y^2 - z^2$. In these coordinates, $\mathcal{H}$ is the quadric defined by the equation

$$x^2 + y^2 - z^2 + 1 = 0 .$$

Such a surface is called a ***hyperboloid of two sheets***.

It is easy to check that $\mathcal{H}$ is invariant by rotations around the z-axis and by reflections through vertical planes through the origin. (Note that this is a particular case of Theorem 5.7).

The intersection of $\mathcal{H}$ with horizontal planes $\{z = z_0\}$ is empty for $|z_0| < 1$, and is the circle $x^2 + y^2 = z_0^2 - 1$ for $|z_0| > 1$. In particular, it is clear that $\mathcal{H}$ has two connected components $\mathcal{H}^+$ and $\mathcal{H}^-$, called upper and lower sheets. On the other hand, its intersection with a vertical plane is a hyperbola. Indeed, by rotational symmetry, it is enough to consider the plane $y = 0$; it intersects the hyperboloid is the hyperbola $z^2 - x^2 = 1$.

Note that the upper arc of this hyperbola can be parametrized using the ***hyperbolic*** trig functions: $(x = \sinh t, z = \cosh t)$ (this is the explanation for the name of these functions). We shall see in § 5.4 that this parametrized curve is a geodesic. The hyperbola is asymptotic to its axes with equation $z^2 - x^2 = 0$, i.e. $z = \pm x$. The hyperboloid $\mathcal{H}$ itself is asymptotic to the cone $x^2 + y^2 - z^2 = 0$ (which can be obtained by rotating the hyperbola’s axes), in other words to the light cone $\langle v, v \rangle = 0$. See Figure 5.1 for an illustration.

In this chapter, we are interested in the upper sheet $\mathcal{H}^+$ of the hyperboloid.

5.1.2 Hyperboloid of dimension n

The previous story naturally generalizes to an arbitrary dimension $n \geqslant 1$. The (unit) hyperboloid of two sheets $\mathcal{H} \subseteq \mathbb{R}^{n,1}$ is still defined by

$$\mathcal{H} := \{v \in M \colon \langle v, v \rangle = -1\} .$$

In coordinates $v = (x_1, \dots, x_{n+1})$ on Minkowski space, $\mathcal{H}$ is the quadric defined by the equation

$$x_1^2 + \cdots + x_n^2 - {x_{n+1}}^2 + 1 = 0 .$$

This quadric is invariant by rotations around the x_{n+1}-axis and by reflections through vertical planes through the origin. The intersection of $\mathcal{H}$ with horizontal hyperplanes $\{x_{n+1} = z_0\}$ is

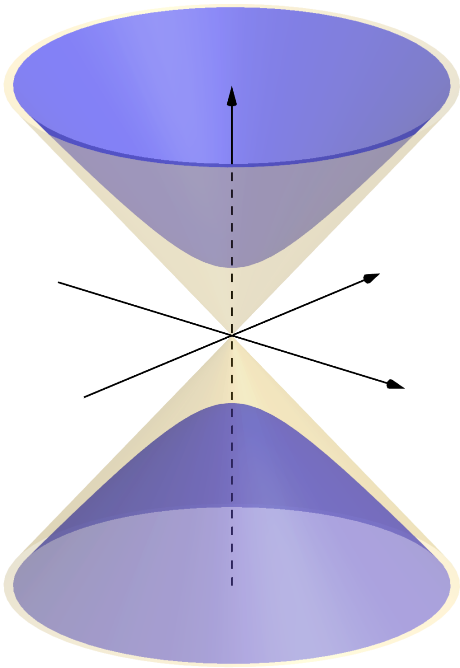


Figure 5.1: The hyperboloid $\mathcal{H}$ in $\mathbb{R}^{n,1}$.

empty for $|z_0| < 1$, and is the sphere $x^2 + \cdots + x_n^2 = z_0^2 - 1$ for $|z_0| > 1$. Again, $\mathcal{H}$ has two connected components (sheets) $\mathcal{H}^+$ and $\mathcal{H}^-$, distinguished by the sign of x_{n+1}.

It is interesting to note that intersecting $\mathcal{H}$ with subspaces of M intersecting it yields lower dimensional hyperboloids:

Proposition 5.1. *Let W be a subspace of M intersecting $\mathcal{H}$. Then W is a Minkowski space, and $W \cap \mathcal{H}$ is the unit hyperboloid in W.*

Proof. Elementary: left as exercise. ∎

Again, the hyperboloid $\mathcal{H}$ is asymptotic to the light cone, which is the isotropic cone in Minkowski space. In coordinates:

$$x_1^2 + \cdots + x_n^2 - {x_{n+1}}^2 = 0\,.$$

In the rest of this chapter, we use the notation $\mathcal{H}^+$ or $\mathbb{H}^n$ indistinctly to refer to the upper sheet of the hyperboloid equipped with the Riemannian metric defined below.

5.2 Riemannian metric

First we identify the tangent space:

Proposition 5.2. *The hyperboloid $\mathcal{H}$ is a smooth embedded surface in $\mathbb{R}^{n,1}$. Its linear tangent space $\mathrm{T}_p\mathcal{H}$ at a point $p \in \mathcal{H} \subseteq \mathbb{R}^{n,1}$ is the plane $p^\perp$.*

Remark 5.3. A couple of clarifications:

- Note that when we write $p^\perp$, we think of p as a vector in $\mathbb{R}^{n,1}$.
- We use the phrase ***linear tangent space*** to make it clear that it is a vector space. The ***affine tangent space*** at p is the affine plane through p in $\mathbb{R}^{n,1}$ with underlying vector space $\mathrm{T}_p\mathcal{H}$. See Figure 5.2.

Proof. The hyperboloid is defined by the equation $q(p) = -1$, where $q(p) = \langle p, p\rangle$. The function q is $\mathcal{C}^\infty$ (it is a degree 2 polynomial), with derivative given by $\mathrm{d}q_p(h) = \langle p, h\rangle$. For any p, the derivative $\mathrm{d}q_p$ is not the zero linear form, since $\mathrm{d}q_p(p) = -1$. The map q is therefore a submersion, and it is a classical fact of differential geometry that any level set such as $q^{-1}(-1)$ is a smooth hypersurface. Moreover, the tangent space at p is the kernel of $\mathrm{d}q_p$, which is precisely $p^\perp$. ∎

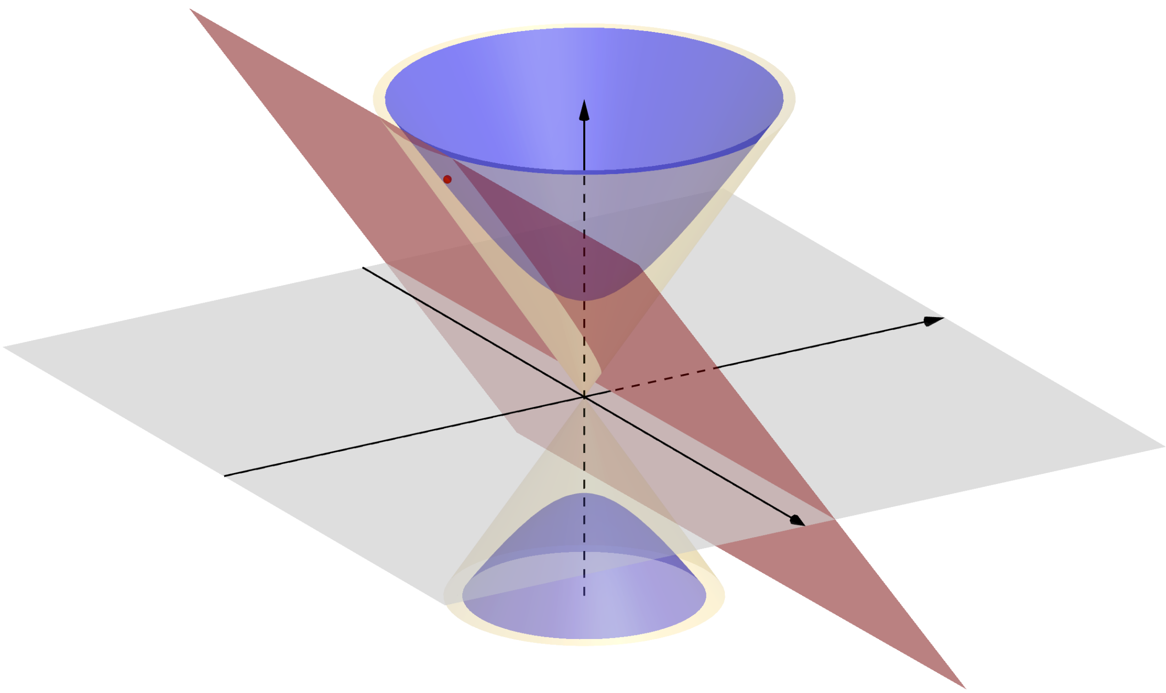


Figure 5.2: The affine tangent space to the hyperboloid at a point.

We can apply Proposition 4.7 to see that $\mathrm{T}_v = p^\perp$ is spacelike. In other words, the restriction on the inner product of $\mathbb{R}^{n,1}$ is positive definite. We thus get:

Proposition 5.4. *The restriction of the inner product of $\mathbb{R}^{n,1}$ to $\mathcal{H}$ is a Riemannian metric.*

We now have a precise definition of the hyperboloid model:

Definition 5.5. The hyperboloid model of the hyperbolic plane is the upper sheet $\mathcal{H}^+$ equipped with the Riemannian metric induced from the Minkowski inner product.

Let us looks at the case $n = 2$ with coordinates (x, y, z) on $\mathbb{R}^{2,1}$: $\mathcal{H}^+$ is defined implicitly by the equation $x^2 + y^2 - z^2 = -1$ with $z > 0$, and the Riemannian metric is the restriction to $\mathcal{H}^+$ of the Minkowski metric

$$\mathrm{d}s^2 = \mathrm{d}x^2 + \mathrm{d}y^2 - \mathrm{d}z^2 .$$

5.3 Isometries

We have seen in § 4.2 that the group of linear isometries of Minkowski space is $\mathrm{O}(n, 1)$. It is clear that these preserve the quadratic form $q(v) = \langle v, v\rangle$, therefore it preserves its level sets. In particular, $\mathcal{H}$ is invariant under the action of $\mathrm{O}(n, 1)$. Since the action of $\mathrm{O}(n, 1)$ on $\mathbb{R}^{n,1}$ is linear and preserves the inner product, its induced action on $\mathcal{H}$ is by Riemannian isometries. Of course, the action of an element $f \in \mathrm{O}(n, 1)$ is orientation-preserving if and only if $f \in \mathrm{SO}(n, 1)$. It is not hard to see that an element of $\mathrm{O}(n, 1)$ preserves $\mathcal{H}^+$ and $\mathcal{H}^-$ if it is time-orientation preserving, and exchanges them otherwise.

Theorem 5.6. *The groups $\mathrm{O}^+(n, 1)$ and $\mathrm{SO}^+(n, 1)$ act isometrically on $\mathcal{H}^+$. Moreover:*

(i) The action of $\mathrm{O}^+(n, 1)$ and $\mathrm{SO}^+(n, 1)$ on $\mathcal{H}^+$ is transitive.

(ii) For any $p \in \mathcal{H}^+$, the stabilizer $K_p \subseteq \mathrm{O}^+(n, 1)$ [resp. $K_p \subseteq \mathrm{SO}^+(n, 1)$] acts transitively on the set of [positive] orthonormal bases of $\mathrm{T}_p\,\mathcal{H}^+$. In particular, the action of K_p in $\mathrm{T}_p\,\mathcal{H}^+$ is transitive.

By definition, Theorem 5.6 shows that $\mathcal{H}^+$ is ***homogeneous*** and ***isotropic***. In particular, it satisfies Euclid's fourth postulate (E4).

Loosely speaking, Theorem 5.6 says $\mathcal{H}^+$ has a very big group of isometries. More precisely, it *contains* a very big group of isometries, namely $\mathrm{O}^+(n, 1)$; but in fact the full group of isometries cannot be bigger:

Theorem 5.7. *The group of isometries of $\mathcal{H}^+$ is $\mathrm{O}^+(n, 1)$, and the subgroup of orientation-preserving isometries is $\mathrm{SO}^+(n, 1)$.*

The proof of Theorem 5.6 and Theorem 5.7, as well as an expansion of the discussion preceding it, are treated in Exercise 5.1.

5.4 Geodesics

Recall that the hyperboloid is the analog of a sphere in Minkowski space (§ 5.1). In order to find the geodesics of the hyperboloid, we are going to follow the same strategy that we used to describe the geodesics on the sphere (Exercise 2.5).

Let $p \in \mathcal{H}^+ \subseteq \mathbb{R}^{n,1}$ and $v \in \mathrm{T}_p\,\mathcal{H}^+$, i.e. $v \perp p$. Let us determine the geodesic γ_v. Denote by P the 2-plane in $\mathbb{R}^{n,1}$ containing p and v. Note that P has signature $(1,1)$ so it is nondegenerate, therefore the reflection r through P is a well-defined element of $\mathrm{O}(n,1)$, moreover it fixes p so it must be in $\mathrm{O}^+(n,1)$. Call f the induced isometry of $\mathcal{H}^+$. The set of fixed points of f is the one-dimensional hyperboloid $\mathcal{H}^+ \cap P$, which is an arc of hyperbola (see Figure 5.3). Since $v \in P$, the geodesic γ_v must be contained in $\mathcal{H}^+ \cap P$ by Proposition 2.10. Given that $\mathcal{H}^+ \cap P$ is 1-dimensional, γ_v is just the constant speed parametrization of it. This parametrization has a nice closed expression:

Theorem 5.8. *The geodesic γ_v in $\mathcal{H}$ with initial velocity $v \in \mathrm{T}_p\,\mathcal{H}^+$ is given by:*

$$\gamma_v(t) = \cosh(\|v\|t)p + \sinh(\|v\|t)\frac{v}{\|v\|}\,. \tag{5.1}$$

Proof. By the previous discussion, it is enough to check that $\gamma(0) = p$, $\gamma'(0) = v$, γ has constant speed, and γ is contained in $\mathcal{H}^+ \cap P$. All these verifications are immediate. ∎

Remark 5.9. To be perfectly rigorous, the previous argument only shows that γ_v is *contained* in $\mathcal{H}^+ \cap P$, not equal to it. Therefore we only proved that the expression (5.1), call it $\tilde{\gamma}(t)$, coincides with $\gamma_v(t)$ for t in some interval containing 0. However, by repeating the argument at another point $p_1 = \tilde{\gamma}(t_1)$ with $v_1 = \tilde{\gamma}'(t_1)$, we see that the curve $\tilde{\gamma}$ must also be the geodesic with these initial conditions. This proves that $\tilde{\gamma}$ is a geodesic for all t, hence it is a maximal geodesic.

Corollary 5.10. *The hyperboloid model $\mathcal{H}^+$ is complete.*

Proof. Theorem 5.8 shows that geodesics are defined for all time, i.e. $\mathcal{H}^+$ is geodesically complete. By the Hopf–Rinow theorem, this is equivalent to any of the well-known characterizations of complete Riemannian manifolds. ∎

Corollary 5.11. *Any two distinct points p and q in $\mathcal{H}^+$ are joined by a unique geodesic segment γ (up to parametrization), moreover γ is length-minimizing: $d(p,q) = L(\gamma)$.*

Proof. The discussion above shows that any geodesic through p and q must be contained in a 2-dimensional subspace $P \subseteq \mathbb{R}^{n,1}$. There is only one choice: it is the space spanned by p and q. This yields both existence and uniqueness of the geodesic, up to parametrization.

The fact that γ is length-minimizing is an immediate consequence of the standard fact in Riemannian geometry that there exists a length-minimizing geodesic between any two points in a complete Riemannian manifold: see [Lee18, Cor. 6.21]. ∎

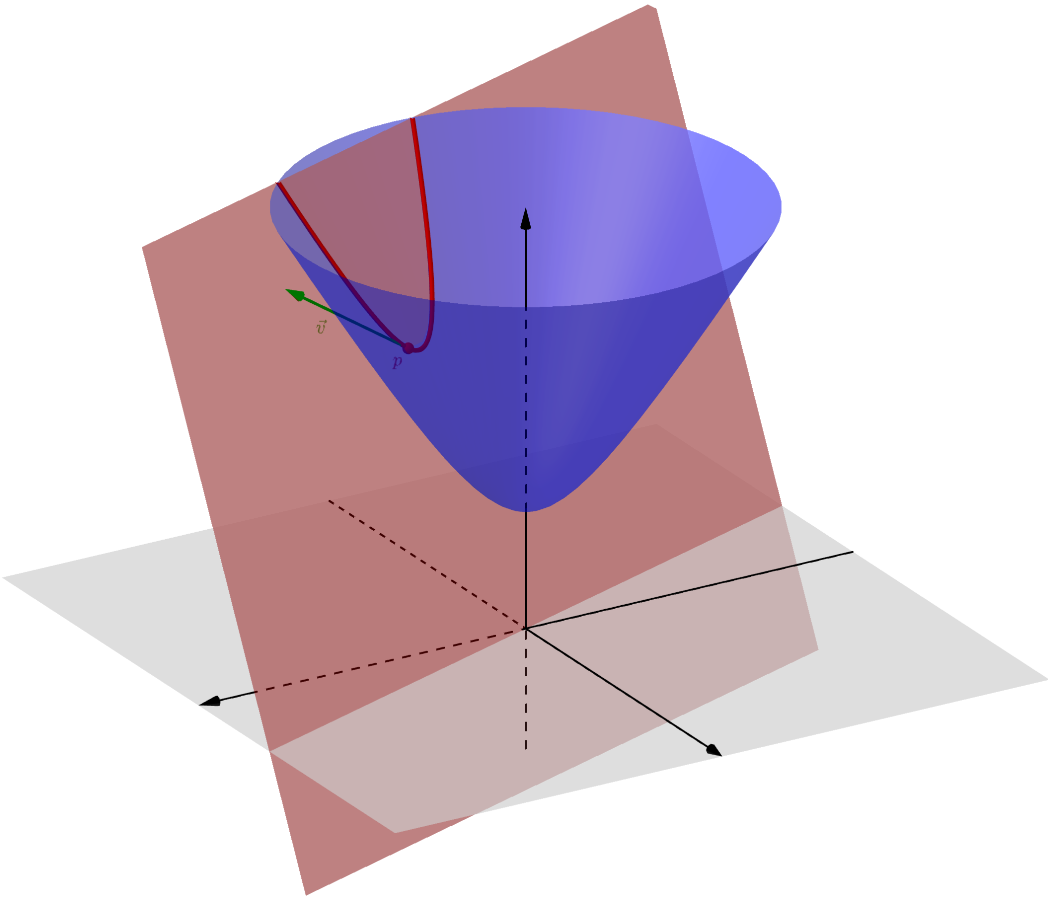


Figure 5.3: Geodesic on the hyperboloid.

Notice that Corollary 5.11 shows that the hyperboloid model satisfies Euclid's first postulate (E1), in its strictest interpretation, while Corollary 5.10 shows that it satisfies the second postulate (E2).

5.5 Distance

Theorem 5.12. *The distance between any two points p and q in $\mathcal{H}^+$ is given by*

$$d(p,q) = \measuredangle(p,q) = \operatorname{arcosh}(-\langle p,q\rangle)$$

where $\measuredangle(p,q)$ is the hyperbolic angle in $\mathbb{R}^{n,1}$.

Proof. By Corollary 5.11, it is enough to show that $d(p,q) = \measuredangle(p,q)$ when $p = \gamma(t_0)$ and $q = \gamma(t_1)$ where γ is any geodesic. After reparametrizing, we can assume that $t_0 = 0$, $t_1 > 0$, and γ has unit speed. On the one hand, $d(p,q)$ is the length of γ between t_0 and

t_1 since γ is the unique geodesic, that is $d(p,q) = t_1$. On the other hand, by Theorem 5.8, $\gamma(t) = \cosh(t)p + \sinh(t)v$ for some unit vector v, so we have $q = \cosh(t_1)p + \sinh(t_1)v$. Recall that the hyperbolic angle $\measuredangle(p,q)$ is given by

$$\langle p,q\rangle^2 = \langle p,p\rangle \langle q,q\rangle \cosh^2 \measuredangle(p,q) .$$

Here $\langle p,p\rangle = \langle q,q\rangle = -1$ and $\langle p,q\rangle = -\cosh t_1$, so we find $d(p,q) = t_1 = \operatorname{arcosh}(-\langle q,q\rangle)$, and $\cosh^2 t_1 = \cosh^2 \measuredangle(p,q)$ yields $\measuredangle(p,q) = t_1$. ■

5.6 Curvature

The goal of this section is to prove:

Theorem 5.13. *$\mathcal{H}^+ \subseteq \mathbb{R}^{n,1}$ has constant curvature sectional curvature -1.*

This result holds in any dimension $n \geqslant 2$. Note that for $n = 1$, the hyperboloid $\mathcal{H}^+$ is still well-defined (it is an arc of hyperbola in $\mathbb{R}^{1,1}$), but the notion of sectional curvature is irrelevant for one-dimensional manifolds. First let us argue that it is enough to prove Theorem 5.13 in the case $n = 2$.

Consider a 2-plane $P \subseteq \mathrm{T}_p \mathcal{H}^+$. By definition, the sectional curvature $K_p(P)$ of $\mathcal{H}^+$ at p in the direction P is the Gaussian curvature at p of the surface $S_P \subseteq \mathcal{H}^+$ is the union of all geodesics whose initial tangent vector is in P (in the language of Riemannian geometry, $S_P = \exp_p(P)$). We know from Theorem 5.8 that the geodesic with initial tangent vector v is $\mathcal{H}^+ \cap P_v$, where we have denoted P_v the 2-plane spanned by p and v. It follows that

$$\begin{aligned} S_P &= \bigcup_{v \in P} \mathcal{H}^+ \cap P_v \\ &= \mathcal{H}^+ \cap W \end{aligned}$$

where $W \subseteq \mathbb{R}^{n,1}$ is the 3-dimensional subspace spanned by p and P. Now, by Proposition 5.1, $S_P = \mathcal{H}^+ \cap W$ is the unit hyperboloid in the Minkowski space W, which has signature $(2,1)$. Thus it is enough to show that cH^+ has sectional (i.e. Gaussian) curvature in the $n = 2$ case:

Theorem 5.14. *$\mathcal{H}^+ \subseteq \mathbb{R}^{2,1}$ has constant Gaussian curvature -1.*

In order to prove Theorem 5.14, we would like to use the fact that for surfaces in $\mathbb{R}^3$, the Gaussian curvature is equal to the product of the principal curvatures (i.e. the determinant of the second fundamental form in an orthonormal basis). For surfaces in Minkowski space $\mathbb{R}^{2,1}$, this result is still true, but with the opposite sign:

Lemma 5.15. *The Gaussian curvature of a surface $S \subseteq \mathbb{R}^{2,1}$ is equal to* minus *the product of the principal curvatures, i.e.* minus *the determinant of the second fundamental form in an orthonormal basis.*

The proof of this lemma requires some knowledge of Riemannian geometry, readers who have not taken a course in Riemannian geometry may skip what follows. Lemma 5.15 is a special case of the modified ***Gauss equation***:

Theorem 5.16. *Let S be a spacelike hypersurface in a Lorentzian manifold M. The sectional curvature $\bar{K}$ of M and the sectional curvature K of S are related by:*

$$\bar{K} = K + \det B\,. \tag{5.2}$$

Remark 5.17. More precisely, (5.2) means that for any orthonormal pair $\{X, Y\} \subseteq \mathrm{T}\, S$:

$$\bar{K}(X, Y) = K(X, Y) + B(X, X)B(Y, Y) - B(X, Y)^2\,.$$

We recall that in the Riemannian case, the Gauss equation is instead $\bar{K} = K - \det B$.

Proof. Choose a local unit normal N to S. Note that since S is a spacelike hypersurface, N must be timelike: $\langle N, N\rangle = -1$. As in the Riemannian case, the second fundamental form B may be defined by the formula

$$\bar{\nabla}_X Y = \nabla_X Y + B(X, Y)N \tag{5.3}$$

where $\bar{\nabla}$ [resp. ∇] denotes the Levi-Civita connection of M [resp. S]. It is an elementary exercise which we leave to the reader to check that while this gives the formula

$$B(X, Y) = +\langle \bar{\nabla}_X N, Y\rangle$$

(instead of $B(X, Y) = -\langle \nabla_X N, Y\rangle$), and that (5.3) and (5.6) yield the modified Gauss equation (5.2) (use the definition of sectional curvature with the Riemann curvature tensor). ∎

We can now prove Theorem 5.14:

Proof of Theorem 5.14 with extrinsic curvatures. By Lemma 5.15, we would like to show that the determinant of the second fundamental form of $\mathcal{H}^+ \subseteq \mathbb{R}^{2,1}$ is equal to 1 at any point p. Clearly, it is enough to show that the extrinsic curvature $\rho_p(v) = \langle \gamma_v''(0), N\rangle$ (which coincides with $B(v, v)$) is equal to 1 (or -1, depending on the choice of unit normal) for every unit vector $v \in \mathrm{T}_p\, M$. This is immediate to check with the explicit expression of γ_v given in (5.1). ∎

In the exercises, we will give two other nice proofs of the fact that $\mathcal{H}^+$ has constant sectional curvature -1:

- A classical proof using the Riemannian geometry notion of *Jacobi fields* is proposed in Exercise 5.3.
- A proof using distance between geodesics is proposed in Exercise 5.2.

5.7 Hyperbolic space of radius R

Instead of considering the "unit" hyperboloid $\mathcal{H} \subseteq \mathbb{R}^{n,1}$, we could instead have defined the hyperboloid $\mathcal{H}$ "of radius $R > 0$" by:

$$\mathcal{H}_R \coloneqq \{v \in M \colon \langle v, v \rangle = -R^2\} .$$

Figure 5.4 shows hyperboloids of different radii.

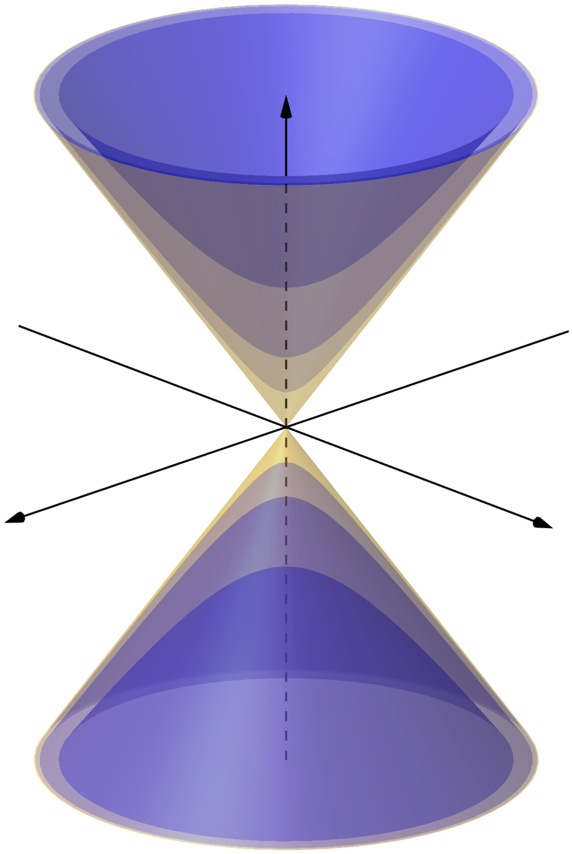


Figure 5.4: Hyperboloids of radii $R = \frac{1}{2}$, $R = 1$, and $R = 2$.

Everything we have seen about the unit hyperboloid $\mathcal{H} = \mathcal{H}_1$ works the same for $\mathcal{H}_R$, with some minor differences:

Riemannian metric. We still equip $\mathcal{H}_R^+$ with the metric induced from Minkowski space $\mathbb{R}^{n,1}$, which is positive definite.

Isometries. It is still the case that the group of isometries of $\mathcal{H}_R^+$ is the orthochronous Lorentz group $\mathrm{O}^+(n, 1)$ and its group of orientation-preserving isometries is the special orthochronous Lorentz group $\mathrm{SO}^+(n, 1)$.

Geodesics. Geodesics in $\mathcal{H}_R^+$ are still intersections on $\mathcal{H}_R^+$ with 2-planes in $\mathbb{R}^{n,1}$, but now the parametrization of the geodesic has to be written:

$$\gamma_v(t) = \cosh\left(\frac{\|v\|}{R}t\right)p + R\sinh\left(\frac{\|v\|}{R}t\right)\frac{v}{\|v\|}\,.$$

Distance. For the distance on $\mathcal{H}_R^+$, we now have

$$d(p,q) = R\measuredangle(p,q)$$

where $\measuredangle(p,q)$ is the hyperbolic angle in $\mathbb{R}^{n,1}$. This amounts to

$$d(p,q) = R\,\mathrm{arcosh}\left(\frac{-\langle p,q\rangle}{R^2}\right)\,.$$

Curvature. Following the same proof as before, the modified expression of geodesics leads to finding that $\mathcal{H}_R^+$ has constant sectional curvature $k = -\frac{1}{R^2}$. Let us recap the most important information:

Theorem 5.18. *The upper sheet $\mathcal{H}_R^+$ of the hyperboloid of radius R is a complete and uniquely geodesic Riemannian manifold of constant sectional curvature $k = -\frac{1}{R^2}$.*

We recall that in Riemannian geometry, one shows that such a model for the *space form of curvature* $k < 0$ is essentially unique: see the discussion of § 2.4 and in particular Theorem 2.21.

5.8 Exercises

Exercise 5.1.

Isometries of the hyperboloid

The goal of this exercise is to determine the group of isometries of hyperbolic space in the hyperboloid model, in particular to write a careful proof of Theorem 5.7.

Let $\mathcal{H} = \{v \in M : \langle v, v \rangle = -1\}$ denote the hyperboloid of two sheets in Minkowski space $M = \mathbb{R}^{n,1}$ and let $\mathcal{H}^+$ indicate the upper sheet (with $x_{n+1} > 0$).

(1) Let us prove that $\mathrm{O}^+(n, 1)$ acts by isometries on $\mathcal{H}^+$.

   (a) Show that the linear action of $\mathrm{O}(n, 1)$ on M preserves $\mathcal{H}$.
   (b) Show that $\mathrm{O}(n, 1)$ acts on $\mathcal{H}$ by Riemannian isometries.
   (c) Show that $f \in \mathrm{O}(n, 1)$ preserves $\mathcal{H}^+$ if and only if $f \in \mathrm{O}^+(n, 1)$. Conclude that $\mathrm{O}^+(n, 1) \subseteq \mathrm{Isom}(\mathcal{H}^+)$.
   (d) *Optional.* Show that $f \in \mathrm{O}^+(n, 1)$ is orientation-preserving on $\mathcal{H}^+$ if and only if $f \in \mathrm{SO}^+(n, 1)$. Conclude that $\mathrm{SO}^+(n, 1) \subseteq \mathrm{Isom}^+(\mathcal{H}^+)$.

(2) Let us prove that, conversely, any isometry of $\mathcal{H}^+$ is induced by an element of $\mathrm{O}^+(n, 1)$.

   (a) Show that the action of $\mathrm{O}^+(n, 1)$ on $\mathcal{H}^+$ is transitive.
   (b) Derive from the previous question that it is enough to show that any isometry of $\mathcal{H}^+$ fixing some point is induced by some element of $\mathrm{O}(n, 1)$ fixing that point.
   (c) Identify the subgroup K of $\mathrm{O}(n, 1)$ fixing $v_0 = (0, \ldots, 0, 1)$. Show that the induced action of K in $\mathrm{T}_{v_0} \mathcal{H}^+$ is transitive on the set of orthonormal bases of $\mathrm{T}_{v_0} \mathcal{H}^+$.
   (d) Let f be an isometry of $\mathcal{H}^+$ fixing v_0. Show that f is completely determined by its derivative at v_0.
   (e) Conclude that $\mathrm{Isom}(\mathcal{H}^+) = \mathrm{O}^+(n, 1)$ and $\mathrm{Isom}^+(\mathcal{H}^+) = \mathrm{SO}^+(n, 1)$.

Exercise 5.2.

Geodesic deviation on the hyperboloid

We denote as usual $\mathcal{H}^+ \subseteq \mathbb{R}^{n,1}$ the upper sheet of the hyperboloid in Minkowski space. Let $p \in \mathcal{H}^+$ and let $v, w \in \mathrm{T}_p \mathcal{H}^+$ be an orthonormal pair of tangent vectors. It is a general fact of Riemannian geometry that the distance between the geodesics $\gamma_v(t)$ and $\gamma_w(t)$ satisfies

$$d\left(\gamma_v(t), \gamma_w(t)\right)^2 = 2t^2 - \frac{1}{3} K\, t^4 + O(t^5)$$

as $t \longrightarrow 0$, where K denotes the sectional curvature of the plane spanned by v and w. (This is discussed in § 2.3.3.)

(1) Show that $d\left(\gamma_v(t), \gamma_w(t)\right) = \operatorname{arcosh}\left(\cosh^2 t\right)$.

(2) Find the Taylor expansion of $\operatorname{arcosh}(\cosh^2 x)$ to order 3 as $x \longrightarrow 0$.

(3) Conclude that $K = -1$.

(4) Show likewise that $\mathcal{H}_R^+$ has constant sectional curvature $-\frac{1}{R^2}$.

Exercise 5.3.

Jacobi fields on the hyperboloid

We denote as usual $\mathcal{H}^+ \subseteq \mathbb{R}^{n,1}$ the upper sheet of the hyperboloid in Minkowski space.

(1) Let $v, w \in \mathrm{T}_p\,\mathcal{H}^+$ be an orthonormal pair. Let us define $\gamma\colon \mathbb{R} \times \mathbb{R} \longrightarrow \mathcal{H}^+$ by

$$\gamma(s,t) = \cosh(t)p + \sinh(t)\left[\cos(s)v + \sin(s)w\right] .$$

Show that:

(i) $\gamma(s,\cdot)$ is a unit geodesic for all $s \in \mathbb{R}$,

(ii) $\gamma(0,\cdot) = \gamma_v$.

Such a family γ is called a ***variation of geodesics***.

(2) Let $J(t) = \frac{\partial}{\partial s}_{|s=0}\gamma(s,t)$. Check that $J(0) = 0$ and $J'(0) = w$. This is called a ***normal Jacobi field***.

(3) We admit the fact: if $J(t)$ is a normal Jacobi field along a unit geodesic and $J''(t) + k(t)J(t) = 0$, then the sectional curvature of the plane spanned by $\gamma'(t)$ and $J(t)$ is equal to $k(t)$ for all $t > 0$[1]. Show that the plane spanned by v and w has curvature -1.

(4) Conclude that $\mathcal{H}^+$ has constant sectional curvature -1.

(5) Show similarly that the hyperboloid of radius R has constant sectional curvature $-\frac{1}{R^2}$.

Exercise 5.4.

Horocycles on the hyperboloid

Let P be an affine plane in Minkowski space $\mathbb{R}^{2,1}$ whose underlying vector space $\vec{P}$ is the orthogonal of an isotropic vector n. The curve $\mathcal{H}^+ \cap P$ is called a ***horocycle***[2].

[1]Students who know Riemannian geometry should recall why this is true. It follows from the ***Jacobi equation*** $J''(t) + R(J(t),\gamma'(t))\gamma'(t) = 0$.

[2]Horocycles will be more naturally defined and studied in Chapter 11.

(1) Show that $P = \{p \in \mathbb{R}^{2,1} \colon \langle p, n \rangle = c\}$ where c is a constant.

(2) *Optional.* Show that any two horocycles are ***congruent***, i.e. differ by an isometry of $\mathcal{H}^+$.

(3) Show that any horocycle is a parabola.

(4) (*) Show that all the geodesics in $\mathcal{H}^+$ perpendicular to a given horocycle are asymptotic.

Exercise 5.5.

Comparing hyperboloids

We denote $(\mathcal{H}^+_R, g_R)$ the upper sheet of the hyperboloid of radius R in $\mathbb{R}^{n,1}$ equipped with its Riemannian metric,

(1) Find a natural map $f \colon \mathcal{H}^+_R \to \mathcal{H}^+_1$.

(2) Compare g_R and $f^* g_1$. Recover the results of § 5.7.

Exercise 5.6.

Euclid's fifth postulate for the hyperboloid

Does Euclid's fifth postulate hold for the hyperboloid model? Compute the angle of parallelism as a function of the distance a (see Figure 1.3).

Exercise 5.7.

Curvature of the hyperboloid in $\mathbb{R}^3$ (*)

(1) Compute the Gaussian curvature of the hyperboloid in $\mathbb{R}^3$. Extend the result to $\mathbb{R}^n$.

(2) Observe that the hyperboloid has positive Gaussian curvature everywhere. Recover this result without any computations after doing Exercise 7.17.

CHAPTER 6

Relativity theory

Disclaimer: This chapter is a draft.

6.1 Relativistic addition of velocities

Part III

Projective geometry and the Klein model

Projective geometry is all geometry.

– Arthur Cayley[1]

Projective geometry has opened up for us with the greatest facility new territories in our science, and has rightly been called the royal road to our particular field of knowledge.

– Felix Klein[2]

[1]Quoted from [New00].
[2]Quoted from [Bel86].

CHAPTER 7

Projective geometry

This chapter is an introduction to projective spaces and projective geometry. I initially planned to make it shorter but ended up expanding it in order to give a fair overview of the subject. As a result, it contains more information than what is strictly needed in subsequent chapters; in particular the last section (§ 7.6) can safely be skipped, as well as § 7.1.4, § 7.3.4, § 7.3.5.

The main prerequisite for this chapter is a good understanding of the basics of abstract linear algebra. Knowledge of general topology is necessary to understand a few points (especially § 7.1.2), but they are not critical.

* * *

Intuitively, a projective space is an affine space, such as the Euclidean plane, that has been augmented with a hyperplane at infinity, the “horizon”. Points at infinity represent directions, i.e. equivalence classes of parallel lines (in perspective drawing, they are called *vanishing points*). Although this point of view is heuristically useful, it does not provide a very satisfactory mathematical definition of a projective space for several reasons, firstly because it breaks the symmetry: points at infinity should not be special.

The synthetic approach consists in defining projective spaces axiomatically, *à la* Euclid, starting with the primitive notions of points, lines, and incidence. For instance, a possible definition of the projective plane uses the three axioms: Any two points lie on a unique line; Any two lines meet at a unique point; There exists four points, no three of which are collinear. (We shall see a slightly different axiomatization in § 7.1.4.)

In this book, we favor instead the modern (now standard) approach according to which a projective space is, by definition, the set of vector lines in a vector space. This somewhat formal viewpoint is very effective because it enables one to study projective geometry with all the might of linear algebra. As an example, any projective transformation is just a linear transformation of the associated vector space. Another example: projective conics and quadrics are defined by symmetric bilinear forms and easily classified with linear algebra.

The reader may legitimately wonder what a chapter on projective geometry is doing in a book on hyperbolic geometry. While these are two different subjects, they are in fact deeply related. We shall see in the next chapter that projective geometry offers the most elegant model of hyperbolic space, called the Cayley–Klein model (and its affine version, the Beltrami–Klein model). More generally, the realization that projective geometry offers a unifying frame for both Euclidean and non-Euclidean geometries led Felix Klein to offer his famous *Erlangen program* in 1872 [Kle93].

Beside the Klein models of hyperbolic space, there is another concrete reason that hyperbolic geometry appeals to projective geometry. It is almost by accident that, in dimensions 2 and 3, the Poincaré models introduced in Chapter 10 are best described in terms of (one-dimensional) *complex* projective geometry. This coincidence has to do with the conformal nature of the Poincaré models and the fact that conformal transformations are the same as complex automorphisms in dimension 2. In particular, the group of orientation-preserving isometries of $\mathbb{H}^2$ [resp. $\mathbb{H}^3$] will be identified to $\mathrm{PSL}(2,\mathbb{R})$ [resp. $\mathrm{PGL}(2,\mathbb{C})$]. (Not to worry, all of this will be explained in Part IV!) This association with complex numbers is very useful and one of the reasons that the Poincaré models are often favored for $\mathbb{H}^2$ and $\mathbb{H}^3$. It is therefore essential to understand projective linear transformations, cross-ratios, and other fundamental concepts of real and complex projective geometry.

For more background on projective geometry, its history, the classical approach, the connection to hyperbolic geometry, and more; some great references are the books of Bruce Meserve [Mes83], Harold Coxeter [Cox89; Cox94], Albrecht Beutelspacher and Ute Rosenbaum [BR98], John Stillwell [Sti10, Chap. 8], Jürgen Richter-Gebert [Ric11], and Christopher Baltus [Bal20]. For a more in-depth treatment of the modern approach to projective geometry, I recommend the wonderful book of Pierre Samuel [Sam86; Sam88], the relevant chapters of Marcel Berger's Geometry I and II [Ber77; Ber09], or the more student-oriented books of Jean-Claude Sidler [Sid00] (in French) and Michèle Audin [Aud03; Aud06].

7.1 Projective spaces

7.1.1 Definition

Let $\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$ be the field of real or complex numbers.

Remark 7.1. There is nothing wrong with projective spaces over any field $\mathbb{K}$, but for our purposes we shall only be interested in real and complex projective geometry.

Let V be a vector space over $\mathbb{K}$. We recall that a *vector line* in V is a one-dimensional subspace of V (i.e. a line through the origin).

Definition 7.2. The ***projective space of*** V or ***projectivization of*** V, denoted $\mathbf{P}(V)$, is the set of vector lines in V.

If V is finite-dimensional, the ***dimension*** of $\mathbf{P}(V)$ is defined as $\dim V - 1$.

Example 7.3. Let us look at the lowest dimensions:

- If $\dim V = 0$, then $\mathbf{P}(V)$ is empty (and has dimension -1 or undefined).

- If $\dim V = 1$, then $\mathbf{P}(V)$ contains one single point (and has dimension 0).
- If $\dim V = 2$, then $\mathbf{P}(V)$ is called a ***projective line*** over $\mathbb{K}$ (and has dimension 1).
- If $\dim V = 3$, then $\mathbf{P}(V)$ is called a ***projective plane*** over $\mathbb{K}$ (and has dimension 2).

Example 7.4. The projective space $\mathbf{P}(\mathbb{K}^{n+1})$ is the standard projective space of dimension n and is denoted $\mathbb{K}\mathbf{P}^n$ (other common notations are $\mathbf{P}^n_{\mathbb{K}}$, $\mathbf{P}^n\,\mathbb{K}$, etc.). For instance, the standard real projective plane is $\mathbb{R}\mathbf{P}^2 := \mathbf{P}(\mathbb{R}^3)$.

For $x \in V - \{0\}$, we denote $[x] (= \mathbb{K}x)$ the vector line containing x. Thus $[x]$ is an element of $\mathbf{P}(V)$; in fact, $\mathbf{P}(V) = \{[x],\ x \in V - \{0\}\}$. The vector line $[x]$ is also the equivalence class of $x \in V - \{0\}$ for the equivalence relation defined by $x \sim y$ if and only if x and y are collinear (i.e., if and only if $[x] = [y]$). Therefore the next definition is equivalent to the previous one:

Definition 7.5. The ***projective space of*** V is the quotient set $\mathbf{P}(V) := {}^{\textstyle (V - \{0\})}/_{\textstyle \sim}$.

Remark 7.6. To be precise, the line $[x] \subseteq V$ contains 0 whereas the equivalence class of x does not, therefore the sets $\mathbf{P}(V)$ in Definition 7.2 and Definition 7.5 are technically not the same. But there is an obvious identification between the two: add 0 to each equivalence class.

7.1.2 Topology

Assume that $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$ and V is finite-dimensional. There is a canonical topology on V: the topology induced by any norm, since any two norms are equivalent. Definition 7.5 allows us to equip $\mathbf{P}(V)$ with the quotient topology (see § A.3.2). In order to understand this topology, it helps to identify $\mathbf{P}(V)$ as a quotient of a smaller set:

Proposition 7.7. *Equip V with any norm and let $S \subseteq V$ denote the unit sphere. The inclusion $S \hookrightarrow V$ induces a homeomorphism $S/\mathbb{U} \xrightarrow{\sim} \mathbf{P}(V)$ where $\mathbb{U} = \mathrm{O}(1) = \{\pm 1\}$ [resp. $\mathbb{U} = \mathrm{U}(1) = \{|z| = 1\}$] is the subgroup of $\mathbb{K}^\times$ of scalars of unit norm.*

Proof. This is an elementary exercise of general topology; I encourage you to not keep reading and try to write your own proof for practice.

Let $F \colon S \to \mathbf{P}(V)$ denote the composition of the inclusion $S \hookrightarrow V - \{0\}$ by the quotient map $V - \{0\} \to \mathbf{P}(V)$. The action of $\mathbb{U}$ on V by scalar multiplication preserves S, moreover F is constant on each orbit, therefore it induces a quotient map $f \colon S/\mathbb{U} \to \mathbf{P}(V)$. The line $[x] \in \mathbf{P}(V)$ intersects S exactly along the $\mathbb{U}$-orbit of x, which proves that f is injective. It is also clearly surjective, since any line $[x]$ intersects the sphere (at $\frac{x}{\|x\|}$).

Since $f \colon S/\mathbb{U} \to \mathbf{P}(V)$ is bijective and $S/\mathbb{U}$ is compact (see below), it is enough to show that f is continuous. This is a general fact: see § A.3.2. The compactness of $S/\mathbb{U}$ derives from that of S by continuity of the projection $S \to S/\mathbb{U}$. Finally, the continuity of f is just a matter of unraveling definitions: let $U \subseteq \mathbf{P}(V)$ be an open set; meaning that $\hat{U} := \bigcup_{[x] \in U} [x]$ is an open subset of $V - \{0\}$. We want to show that the preimage of U by f is open in $S/\mathbb{U}$, i.e. its preimage in S is open. But the subset of S we are talking about is just $\hat{U} \cap S$, which is open in S by definition of the subspace topology. ∎

Corollary 7.8. *If V is a finite-dimensional vector space over $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$, then $\mathbf{P}(V)$ is a compact Hausdorff topological space.*

Proof. By Proposition 7.7, it is enough to show that $S/\mathbb{U}$ is compact and Hausdorff. The compactness of $S/\mathbb{U}$ is immediately derived from that of S, as we already explained in the proof of Proposition 7.7. The separation ("Hausdorffness") of $S/\mathbb{U}$ is also derived from that of S, but that is less immediate; the reader may either try to prove it manually or accept it without proof. The general argument is: a continuous action of a compact Hausdorff topological group is always proper ([Bou71, TG III.28, Prop. 2]), and the quotient of a topological space by a proper group action is always Hausdorff ([Bou71, TG III.29, Prop. 3]). ∎

Example 7.9. The real projective plane $\mathbb{R}\mathbf{P}^2$ is homeomorphic to $S^2/\{\pm 1\}$. This is a compact non-orientable surface, which is not easy to visualize. Like other non-orientable surfaces, it cannot be embedded in $\mathbb{R}^3$, but it can be immersed. An immersion of $\mathbb{R}\mathbf{P}^2$ in $\mathbb{R}^3$ can be realized by sewing a Möbius strip to the edge of a disk. Depending on the gluing, the resulting surface is homeomorphic to either a *(sphere with a) cross-cap*, *Boy's surface*, or *Roman's surface*. (I encourage curious readers to look these up for pictures and further description).

Example 7.10. The complex projective line $\mathbb{C}\mathbf{P}^1$ is homeomorphic to $S^3/\mathrm{U}(1)$. It turns out that this is homeomorphic to the 2-sphere S^2. This is known as the ***Hopf fibration***, which is the subject of Exercise 7.5.

7.1.3 Projective subspaces and projective duality

Let $\mathcal{P} = \mathbf{P}(V)$ be a finite-dimensional projective space.

Definition 7.11. A ***projective subspace*** of $\mathcal{P}$ is a subset of the form $\mathbf{P}(W)$ where W is a subspace of V.

Example 7.12. When $\dim W = 1$, $\mathbf{P}(W)$ contains just one point. When $\dim W = 2$, $\mathbf{P}(W)$ is called a ***(projective) line*** in $\mathcal{P}$; when $\dim W = 3$, $\mathbf{P}(W)$ is a ***(projective) plane***, etc. When W is a hyperplane (has codimension 1), $\mathbf{P}(W)$ is called a ***(projective) hyperplane*** in $\mathcal{P}$.

Recall that V^* denotes the dual space of V, i.e. the space of linear forms $V \to \mathbb{K}$.

Definition 7.13. The projective space $\mathbf{P}(V^*)$ is called ***dual projective space*** of $\mathcal{P} := \mathbf{P}(V)$ and denoted $\mathcal{P}^*$.

Points of $\mathcal{P}^*$ correspond to hyperplanes of $\mathcal{P}$: one can associate to any $[\varphi] \in \mathcal{P}^*$ the projective hyperplane $\mathbf{P}(\ker \varphi) \subseteq \mathcal{P}$. Conversely, if $\mathbf{P}(H) \subseteq \mathcal{P}$ is a hyperplane, then the set of linear forms that vanish on H is 1-dimensional, therefore it defines a point in $\mathcal{P}^*$.

More generally, for a subspace $W \subseteq V$, denote $W^\circ \subseteq V^*$ the ***annihilator*** (or ***polar***) of W: by definition, it consists of all the linear forms $\varphi \colon V \to \mathbb{K}$ that vanish on W. (The notation $W^\perp$ is also sometimes used.) Taking the annihilator is a decreasing map with respect to inclusion: if $W_1 \subset W_2$ then $W_2^\circ \subset W_1^\circ$. Moreover, it is an basic exercise of linear algebra that $\dim W^\circ = \operatorname{codim} W$. Taking annihilators of subspaces induces a map between projective subspaces of V and projective subspaces of V^*, namely

$$\mathcal{Q} = \mathbf{P}(W) \mapsto \mathcal{Q}^\circ := \mathbf{P}(W^\circ)\,.$$

This map is called ***projective duality***.

Proposition 7.14. *Let $\mathcal{P}$ be a finite-dimensional projective space. Projective duality is a bijective correspondence between projective subspaces of $\mathcal{P}$ and projective subspaces of $\mathcal{P}^*$, and it is decreasing with respect to inclusion. Moreover, projective duality is involutive in the sense that $\mathcal{Q}^{\circ\circ} = \mathcal{Q}$ under the identification $P^{**} \approx P$.*

The proof of Proposition 7.14 is elementary and left to the reader (Exercise 7.1).

Example 7.15. Let $\mathcal{P}$ be a projective plane. Projective duality defines a bijective correspondence between points [resp. lines] of $\mathcal{P}$ and lines [resp. points] of $\mathcal{P}^*$.

Projective duality is a beautiful and powerful tool; readers should further their understanding of it by working on Exercise 7.1 and Exercise 7.16.

7.1.4 Axioms of projective geometry

It is possible to develop projective geometry axiomatically, in the spirit of Euclid, starting with the primitive notions of points, lines, and incidence. This approach is sometimes called *synthetic*, especially to emphasize that it avoids the use of coordinates. Despite its historical importance, the synthetic approach fell out of fashion for obvious reasons: the power of analytic and algebraic techniques, and the unification of (almost) all modern mathematics in the language of set theory and algebraic structures.

Out of interest, let us give an example of axiomatization of projective geometry. Many choices are possible for the axioms, not always giving exactly the same theory. The following axioms, suggested by Veblen and Young[1] [VY08], are remarkable for their simplicity: A ***projective space*** is a set $\mathcal{P}$ (the set of points), together with a set $\mathcal{L}$ of subsets of $\mathcal{P}$ (the set of lines), satisfying the axioms:

(P1) There exists a unique line through any two points.

(P2) If A, B, C, D are distinct points and the lines AB and CD meet, then so do AC and BD.

(P3) Any line has at least three points.

Starting with these axioms, one can define, for instance, a *projective subspace* by requiring that it contains each line through any two of its points; a *collineation* as a map that preserves alignment, etc.

[1] The mathematician and philosopher Alfred Whitehead had suggested a similar system of axioms in 1906 [Whi60]. The great David Hilbert had proposed a more general axiomatization of geometry in the famous *Grundlagen der Geometrie* [Hil99] in 1899.

Remark 7.16. The axiom (P2) is sometimes called *Veblen's axiom*. It is a clever way of saying that any two coplanar lines must intersect.

Veblen and Young ([VY08], see also [VY65a; VY65b]) proved that this axiomatic definition of projective geometry is equivalent to Definition 7.2 for dimensions > 2:

Theorem 7.17. *If $\mathcal{P}$ is a projective space in the sense above, and has dimension at least* 3 *(i.e. there exists two non-intersecting lines), then it is isomorphic to some projective space* $\mathbf{P}(V)$ *over a division ring* $\mathbb{K}$.

Remark 7.18. The more precise theorem is that a projective space $\mathcal{P}$ is isomorphic to some $\mathbf{P}(V)$ if and only if Desargues's theorem holds (see Exercise 7.9), which is always the case if $\dim \mathcal{P} \geqslant 3$; furthermore $\mathbb{K}$ is commutative if and only if Pappus's theorem holds (see Exercise 7.8). This beautiful result was proven by David Hilbert in the *Grundlagen* [Hil99] (with a different axiomatization). Hilbert also exposed many examples of "non-Desarguesian planes". For the proof, I recommend [BR98, §2.2.2, §3.4.2]

For readers interested in learning projective geometry via the synthetic approach, I recommend [Cox89; Cox94] and [BR98] (or the German edition [BR04]).

7.1.5 Affine charts and hyperplane at infinity

Let $\mathcal{P} = \mathbf{P}(V)$ be a projective space. Choose an affine hyperplane $H \subseteq V$ not containing the origin, and denote by $\vec{H}$ vector hyperplane in V parallel to H (i.e. the vector space underlying H). Every vector line in V intersects H (once) except the lines contained in $\vec{H}$: see Figure 7.1.

By the previous observation, taking the intersection of every line not in $\vec{H}$ with H defines a bijection $\varphi_H \colon \mathcal{P} - \mathcal{H} \to H$ where $\mathcal{H} = \mathbf{P}\left(\vec{H}\right)$. The map φ_H thus identifies $\mathcal{P}$ minus a projective hyperplane to an affine space, and is called an ***affine chart*** (or ***affine patch***). We leave it to the reader to check that it is a homeomorphism. When an affine chart φ_H as above is chosen, the projective hyperplane $\mathcal{H}$ is called ***hyperplane at infinity*** (the term ***horizon*** would also make sense).

Conversely, starting with any affine space H, one can reconstruct the projective space $\mathcal{P}$: first introduce the underlying vector space $\vec{H}$, then the "hyperplane at infinity" $\mathcal{H} = \mathbf{P}\left(\vec{H}\right)$. Geometrically, the points at infinity (the elements of $\mathcal{H}$) can be described as "directions", i.e. equivalence classes of parallel lines in H. This corresponds to the intuition that parallel lines meet at a single point of the horizon. It remains to identify $H \cup \mathcal{H}$ as a projective space, which is achieved by tracing the previous paragraph backwards. More precisely, embed H as an affine hyperplane in a vector space V: for instance, put $V = \vec{H} \times \mathbb{K}$ and fix an identification of H with $\vec{H} \times \{1\}$. Then $\mathbf{P}(V) \approx H \cup \mathcal{H}$.

The extension of an affine space H to a projective space $\mathcal{P} = H \cup \mathcal{H}$ described above (or, more precisely, the embedding $\varphi_H^{-1} \colon H \to \mathcal{P}$) is called the ***projective completion*** of H. The projective completion of H is in particular a nice compactification: $\mathcal{P}$ is Hausdorff and is a smooth manifold (see § 7.2.2).

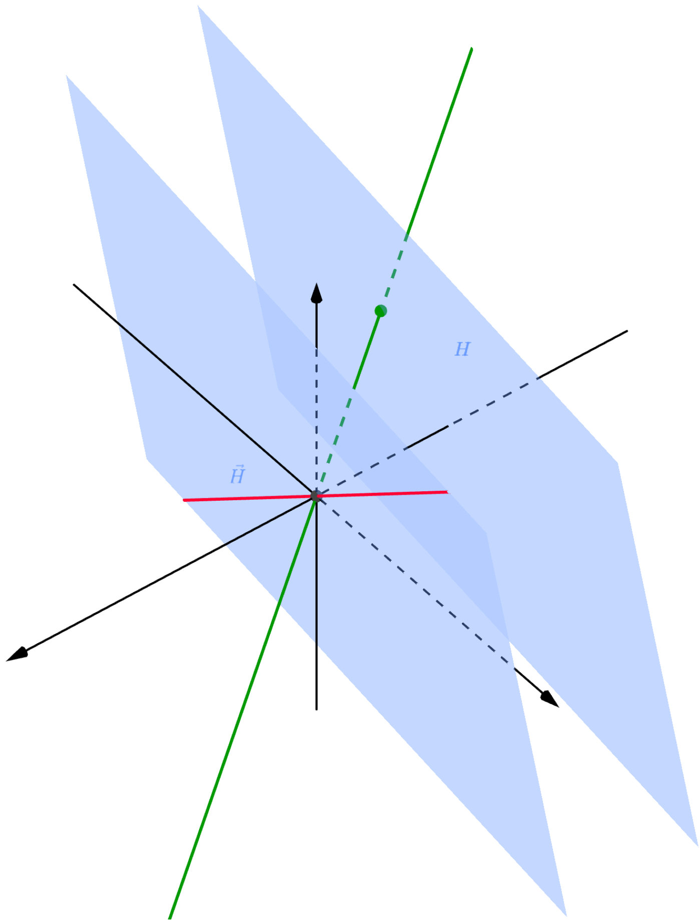


Figure 7.1: Every vector line, unless contained in $\vec{H}$, intersects the affine hyperplane H.

Example 7.19. Let E be an affine space of dimension 1. Show that the projective completion of E is topologically the same as its one-point compactification. Argue that it is homeomorphic to S^1 for $\mathbb{K} = \mathbb{R}$ and S^2 for $\mathbb{K} = \mathbb{C}$.

Example 7.20. Let E be a real affine space of dimension 2. The projective completion of E, obtained by adding the “horizon”, is topologically a closed disk with diametrically opposed points identified. This is a well-known topological description of the projective plane.

Remark 7.21. In an affine space E, a ***central projection*** from a point O to an affine plane H not containing O sends each point M to the intersection of the line OM with H. The projection is not defined for points of the plane through O parallel to H. With our setup, when $E = V$ is a vector space and O is the origin, the central projection is the map $V - \vec{H} \longrightarrow H$ obtained by composition of the quotient map $V - \vec{H} \longrightarrow \mathcal{P} - \mathcal{H}$ with the affine chart $\varphi_H \colon \mathcal{P} - \mathcal{H} \longrightarrow H$. We shall see in § 7.3.4 that central projections are more elegantly defined as projective maps.

We shall further discuss affine charts in § 7.2.2 and show that they define an atlas on $\mathcal{P}$ in the sense of differential geometry, giving it the structure of a manifold.

7.2 Coordinates

There are three ways that one can talk about coordinates on a projective space $\mathcal{P} = \mathbf{P}(V)$: homogeneous coordinates, affine coordinate charts, and projective frames. These are not independent concepts; on the contrary, they are easily related, and can all be derived from a choice of basis for the vector space V.

For the remainder of this section, let V be a vector space over $\mathbb{K}$ of dimension $n+1$ and let $\mathcal{P} = \mathbf{P}(V)$ be its projectivization. We typically denote $(e_1, \ldots, e_{n+1})$ a basis of V and $x = (x_1, \ldots, x_{n+1})$ the associated system of coordinates, meaning that $x = \sum_{k=1}^{n+1} x_i e_i$.

Remark 7.22. From the viewpoint of "coordinate charts", a system of coordinates on V is a linear isomorphism $\varphi \colon V \longrightarrow \mathbb{K}^{n+1}$. Of course, this amounts to the choice of a basis.

7.2.1 Homogeneous coordinates

Recall that we denote $[x] \in \mathcal{P}$ the vector line through x. When using coordinates, we abbreviate $[x] = [(x_1, \ldots, x_{n+1})]$ to the notation $[x] = [x_1 \colon \ldots \colon x_{n+1}]$. This notation is called ***homogeneous coordinates***.

Homogeneous coordinates were introduced by the German mathematician August Ferdinand Möbius in 1827 [Möb27]. (Möbius actually introduced *barycentric coordinates*, but these are essentially the same thing: understand this by working on Exercise 7.4.)

Homogeneous coordinates are not coordinates in the usual sense. Indeed, they are not unique: $[x_1 \colon \ldots \colon x_{n+1}] = [y_1 \colon \ldots \colon y_{n+1}]$ whenever there exists $\lambda \in \mathbb{K}^\times$ such that $y_k = \lambda x_k$ for all $k \in \{1, \ldots, n+1\}$. Also, mind that $[0 \colon \ldots \colon 0]$ is not allowed, because $[0]$ is not well-defined.

Example 7.23. The standard basis (e_1, e_2, e_3) of $\mathbb{R}^3$ gives linear coordinates (x, y, z) on $\mathbb{R}^3$ and homogeneous coordinates $[x \colon y \colon z]$ on $\mathbb{R}\mathbf{P}^2$. The equation of a generic vector plane $P \subseteq \mathbb{R}^3$ is $ax + by + cz = 0$ with a, b, c not all zero. This is also the equation of a generic projective line $\ell = \mathbf{P}(P)$ in $\mathbb{R}\mathbf{P}^2$ in homogeneous coordinates. Prove as an exercise that $[a \colon b \colon c]$ represent the point in $(\mathbb{R}\mathbf{P}^2)^*$ dual to ℓ.

Some elementary properties of homogeneous coordinates are discussed in Exercise 7.3.

7.2.2 Affine coordinate charts

Let $\mathcal{P} = \mathbf{P}(V)$ be a projective space of dimension n. We have seen in § 7.1.5 that any choice of an affine hyperplane $H \subseteq V$ not containing the origin yields an identification $\varphi_H \colon \mathcal{P} - \mathcal{H} \xrightarrow{\sim} H$ (where $\mathcal{H} = \mathbf{P}\left(\vec{H}\right)$) which we called an *affine chart*.

Suppose that we have chosen a basis of V such that in the associated coordinate system $(X_1, \ldots, X_{n+1})$, the equation of H is $X_{n+1} = 1$. Such a choice of basis is always possible: just take any basis $(e_1, \ldots, e_n)$ of H and complete it by taking any $e_{n+1} \in H$. (We are now using uppercase letters X_k for the coordinates on V, to distinguish it from the affine coordinates on

H that we are about to introduce.) The equation of $\vec{H}$ in these coordinates is $X_{n+1} = 0$, which is also the equation of $\mathcal{H}$ in homogeneous coordinates.

There is an obvious choice of coordinates $(x_1, \ldots, x_n)$ on H, which we call ***affine coordinates***: the point with coordinates $(x_1, \ldots, x_n)$ in H is the point with coordinates $(x_1, \ldots, x_n, 1)$ in V. It is easy to compute the expression of φ_H in these coordinates:

Proposition 7.24. *The affine chart $\varphi_H \colon \mathcal{P}-\mathcal{H} \xrightarrow{\sim} H$ is given by $x_k = \frac{X_k}{X_{n+1}}$ for all $k \in \{1, \ldots, n\}$.*

Proof. Let $p = [X_1 : \ldots : X_{n+1}] \in \mathcal{P} - \mathcal{H}$, i.e. $X_{n+1} \neq 0$. To determine $\varphi_H(p)$, we are looking for $\lambda \in \mathbb{K}^\times$ such that $\lambda(X_1, \ldots, X_{n+1}) \in H$. Since the equation of H is $X_{n+1} = 1$, the unique solution is $\lambda = \frac{1}{X_{n+1}}$. The conclusion follows. ∎

Let us now show that such affine charts φ_H give us a coordinate atlas on $\mathcal{P}$ in the sense of differential geometry (manifolds). Instead of considering all possible hyperplanes H, we choose $n+1$ such that the open sets $\mathcal{P}-\mathcal{H}$ cover $\mathcal{P}$. Fix a coordinate system $(X_1, \ldots, X_n)$ on V, and take for H_i the hyperplane with equation $X_i = 1$. The open sets $U_i \coloneqq \mathcal{P} - \mathcal{H}_k = \{X_i \neq 0\}$ cover $\mathcal{P}$, since a point $[X_1 : \ldots : X_{n+1}]$ being in the complement of all would mean that $X_i = 0$ for all k, which is not allowed.

There is a natural choice of coordinates on H_i, which we denote $(x_1, \ldots, \hat{x}_i, \ldots x_{n+1})$, where $\hat{x}_i$ means that the i-th entry of $(x_1, \ldots, x_{n+1})$ is omitted (for instance, $(x, \hat{y}, z) = (x, z)$). As in Proposition 7.24, the map $\varphi_i \colon U_i \longrightarrow \mathbb{K}^n$ is written $x_k = \frac{X_k}{X_i}$ for all $k \neq i$.

At this point we have: a Hausdorff topological space $\mathcal{P}$ covered by open sets U_i, and homeomorphisms $\varphi_i \colon U_i \longrightarrow \mathbb{K}^n$. By definition, this shows that $\mathcal{P}$ is a ***topological manifold***. Each φ_i is called a ***coordinate chart***, and the collection of all charts is a ***coordinate atlas***.

Remark 7.25. Topological manifolds are often required to be ***secound-countable***, i.e. have a countable basis of open sets. For $\mathcal{P}$ this follows easily from the fact that it has a finite atlas. More generally, second-countability is equivalent to paracompactness for Hausdorff and locally Euclidean space, and the paracompactness of $\mathcal{P}$ is trivial because it is compact. For more details, refer to [Lee11] and [Noi].

We can further show that the atlas $\{\varphi_i\}$ defines a ***differential structure*** on $\mathcal{P}$ by proving that any two charts φ_i and φ_j are compatible. By definition, this means that the ***transition function*** (change of coordinates) $\varphi_j \circ \varphi_i^{-1}$ is smooth. Our previous computation shows that it is written $x'_k = \frac{x_k}{x_j}$ for all $k \neq j$, with the exception $x'_i = \frac{1}{x_j}$. This map is smooth, even analytic: it is a rational fraction. In conclusion, we proved that $\mathcal{P}$ is a ***smooth manifold***, and even an ***analytic manifold*** (real analytic when $\mathbb{K} = \mathbb{R}$, and even complex analytic when $\mathbb{K} = \mathbb{C}$).

Remark 7.26. As an exercise, the reader may show more generally that any two charts φ_H are compatible. This amounts to showing the analyticity of any *perspectivity* (see § 7.3.4).

For the basics and more on manifolds, I recommend [Lee13] or [Laf15]. Both cover the example of a projective space, unsurprisingly: it is a great example of manifold.

7.2.3 Projective frames

Projective frames allow the definition of homogeneous coordinates on a projective space $\mathcal{P}$ without referring to the vector space V such that $\mathcal{P} = \mathbf{P}(V)$.

Definition 7.27. Let $\mathcal{P}$ be a projective space of dimension n. A ***projective frame*** is a $(n+2)$-tuple of points $(p_1, \dots, p_{n+2})$ such that no projective hyperplane contains $n+1$ of them.

Example 7.28. Here are some fundamental examples of projective frames:

- A frame for a projective line is a triple of distinct points.
- A frame for a projective plane is a quadruple of points such that no three are collinear.
- The ***standard frame*** of $\mathcal{P} = \mathbb{K}\mathbf{P}^n$ is $([e_1], \dots, [e_{n+2}])$, where $(e_1, \dots, e_{n+1})$ is the standard basis of $\mathbb{K}^{n+1}$ and $e_{n+2} = \sum_{k=1}^{n+1} e_k$.

Lemma 7.29. *Let $\mathcal{P} = \mathbf{P}(V)$ be a projective space of dimension n. If $(p_1, \dots, p_{n+2})$ is a projective frame, there exists $e_1, \dots, e_{n+2} \in V$, unique unique up to multiplication by the same $\lambda \in \mathbb{K}^\times$, such that $p_k = [e_k]$ and $\sum_{k=1}^{n+2} e_k = 0$. Moreover, $(e_1, \dots, e_{n+1})$ is a basis of V.*

Proof. This is an excellent elementary exercise of linear algebra, try to write the proof yourself without reading further! You can also show that the converse is true.

Existence: let $v_1, \dots, v_{n+2} \in V - \{0\}$ such that $p_k = [v_k]$. Since $\dim V = n+1$, the vectors $v_1, \dots, v_{n+2}$ cannot be linearly independent: there exists $\lambda_1, \dots, \lambda_{n+2} \in \mathbb{K}$ such that $\sum_{k=1}^{n+2} \lambda_k v_k = 0$. Observe that none of the λ_k can be zero: for if $\lambda_j = 0$, then the vectors v_k for $k \neq j$ would lie on a vector hyperplane, so that the points p_k for $k \neq j$ would lie on a projective hyperplane. For the same reason, $(v_1, \dots, v_{n+1})$ must be linearly independent, hence a basis of V. Set $e_k = \lambda_k v_k$. Since all λ_k's are $\neq 0$, $(e_1, \dots, e_{n+1})$ is still a basis of V, $p_k = [e_k]$, and the relation $\sum_{k=1}^{n+2} \lambda_k v_k = 0$ is rewritten $\sum_{k=1}^{n+2} e_k = 0$.

Uniqueness: assume $(e_1, \dots, e_{n+2})$ and $(e'_1, \dots, e'_{n+2})$ are two solutions. Since $[e'_k] = [e_k]$, there exists $\lambda_k \in \mathbb{K}^\times$ such that $e'_k = \lambda_k e_k$. In particular, we have $e'_{n+2} = \lambda e_{n+2} = \lambda \sum_{k=1}^{n+1} e_k$ where $\lambda := \lambda_{n+2}$; on the other hand $e'_{n+2} = \sum_{k=1}^{n+1} e'_k = \sum_{k=1}^{n+1} \lambda_k e_k$. Equating the two expressions yields $\sum_{k=1}^{n+1} (\lambda - \lambda_k) e_k = 0$. Since $(e_1, \dots, e_{n+1})$ is a basis of V (otherwise $p_1, \dots, p_n$ are contained in a hyperplane), it follows that $\lambda_k = \lambda$ for all $k \in \{1, \dots, n+1\}$, hence $e'_k = \lambda e_k$. ∎

Since the basis $(e_1, \dots, e_{n+1})$ of Lemma 7.29 is unique up to scalar multiplication, the homogeneous coordinates $[x_1 : \dots : x_{n+1}]$ of a point $p \in \mathcal{P}$ in this basis are well-defined. By definition, these are the ***homogeneous coordinates of** p **in the frame*** $(p_1, \dots, p_{n+2})$.

Example 7.30. In $\mathcal{P} = \mathbb{K}\mathbf{P}^n$, one can take $p_k = [e_k]$ where $(e_1, \dots, e_{n+1})$ is the canonical basis of $\mathbb{K}^{n+1}$ and $e_{n+2} = -\sum_{k=1}^{n+1} e_k$. In that case, the homogeneous coordinates of $[v]$ in the frame $(p_1, \dots, p_{n+2})$ are just $[v_1 : \dots : v_{n+1}]$.

If an affine chart $H \subseteq V$ is chosen, one can choose a collection of affinely independent points $p_1, \dots, p_{n+1} \in H$, i.e. a basis of V consisting of elements of H. Homogeneous coordinates in $\mathcal{P}$ relative to that basis are the same thing as *barycentric coordinates* in H: see

Exercise 7.4 for details. Such homogeneous coordinates correspond to the projective frame $(p_1, \ldots, p_{n+2})$ where $p_{n+2} = \left[\sum_{k=1}^{n+1} p_k\right]$, i.e. p_{n+2} is the ***isobarycenter*** of $\{p_1, \ldots, p_{n+1}\}$ in H.

An important property of projective frames is that a projective transformation is uniquely determined by the image of any projective frame: see Theorem 7.39.

7.3 Projective transformations

Projective transformations can be defined in several ways: as transformations that lift to linear transformations of the associated vector space, or as products of elementary transformations called *perspectivities*, or as (maybe a subclass of) maps that preserve alignment (known as *collineations*). We favor the "projective linear" point of view, but shall also briefly discuss the others for completeness.

7.3.1 Projective linear maps

Let $\mathcal{P} = \mathbf{P}(V)$ and $\mathcal{P}' = \mathbf{P}(V')$ be two projective spaces (we are especially interested in $\mathcal{P}' = \mathcal{P}$). Any linear map $F\colon V \longrightarrow V'$ sends a vector line in V to a vector line in V', unless it is contained in $\ker F$. Therefore F induces a map $f = [F]\colon \mathcal{P} - \mathcal{Q} \longrightarrow \mathcal{P}'$, where $\mathcal{Q} = \mathbf{P}(\ker F)$. Note $f = [F]$ can be described as a quotient map, defined by $[F]([x]) = [F(x)]$. We call f the ***projectivization*** of F, and F a ***lift*** of f.

Definition 7.31. A map $f\colon \mathcal{P} \longrightarrow \mathcal{P}'$, possibly only defined in the complement of a projective subspace, is called ***projective linear*** (or simply ***projective***) if it has a linear lift $F\colon V \longrightarrow V'$.

Example 7.32. Consider the *central projection* $f\colon \mathbb{R}^3 - (\mathbb{R}^2 \times \{0\}) \longrightarrow \mathbb{R}^2 \times \{1\}$ defined by $f(x, y, z) = (x/z, y/z, 1)$ (see Remark 7.21). This map extends to the projective completions $\mathbb{R}\mathbf{P}^3 \longrightarrow \mathbb{R}\mathbf{P}^2$ by $f([x\colon y\colon z\colon t]) = [x\colon y\colon z]$. f is projective linear since it lifts to $F\colon \mathbb{R}^4 \longrightarrow \mathbb{R}^3$, $(x, y, z, t) \longmapsto (x, y, z)$ (observe how F looks nicer than the initial map f!). The kernel of F is $\mathbb{R}e_4$ where $e_4 = (0, 0, 0, 1)$, hence f is defined outside of $O = [e_4]$, which is the origin of our original $\mathbb{R}^3$. More generally, any central projection $E - \vec{H} \longrightarrow H$ extends everywhere outside of the origin when taking the projective completions.

If $f\colon \mathcal{P} \longrightarrow \mathcal{P}'$ is a projective linear map, then its linear ***lift*** $F\colon V \longrightarrow V'$ (i.e. $f = [F]$) is unique up to a scalar factor $\lambda \in \mathbb{K}^\times$ by the next proposition:

Proposition 7.33. *If $F, G\colon V \longrightarrow V'$ are linear maps such that $[F] = [G]$, then there exists $\lambda \in \mathbb{K}^\times$ such that $F = \lambda G$.*

Proof. By definition, if $[F] = [G]$, then F and G have same kernel $W \subseteq V$, and for every $x \in V - W$, $[F(x)] = [G(x)]$. In other words, there exists $\lambda_x \in \mathbb{K}^\times$ such that $G(x) = \lambda_x F(x)$. Using the linearity of F and G, it is readily checked that λ_x must be independent of x. ■

Remark 7.34. Given a projective linear map $f: \mathcal{P} \to \mathcal{P}'$, "the" linear lift F of f is also called the ***homogenization*** of f.

A projective linear map $f: \mathcal{P} \to \mathcal{P}'$ is defined everywhere if and only if its linear lift $F: V \to V'$ has trivial kernel, i.e. is injective, in which case f is also injective. When $\mathcal{P}$ and $\mathcal{P}'$ (hence V and V') have the same dimension, F is injective if and only if it is bijective, in which case f is also bijective. In particular, when $\mathcal{P} = \mathcal{P}'$, we have:

Proposition 7.35. *A projective linear map $f: \mathcal{P} \to \mathcal{P}$ is well-defined everywhere if and only if its linear lift $F: V \to V$ is an element of $\mathrm{GL}(V)$, in which case f is bijective.*

Henceforth we assume that a projective linear map $f: \mathcal{P} \to \mathcal{P}$ is well-defined everywhere, hence bijective; such a map is called ***projective (linear) transformations*** (or ***automorphisms***) of $\mathcal{P}$. The terms ***homography*** and ***projectivity*** (see § 7.3.4) are also synonyms, as is ***collineation*** under suitable assumptions (see § 7.3.5).

An elementary property of projective transformations is that they preserve subspaces:

Proposition 7.36. *Let $f: \mathcal{P} \to \mathcal{P}$ be a projective transformation. The image $f(\mathcal{Q})$ of any projective subspace $\mathcal{Q} \subseteq \mathcal{P}$ is a projective subspace of the same dimension.*

Proof. This is almost trivial: write $\mathcal{Q} = \mathbf{P}(W)$ and $f = [F]$. Since $F \in \mathrm{GL}(V)$, $F(W)$ is a subspace of V of the same dimension as W. The conclusion quickly follows. ∎

7.3.2 Projective transformations in coordinates

A projective transformation in homogeneous coordinates is simply represented by the matrix of its linear lift. To make this precise, suppose that we have homogeneous coordinates $[x_1, : \dots : x_{n+1}]$ on $\mathcal{P}$. This means that we have chosen a basis $(e_1, \dots, e_{n+1})$ of V up to a scalar factor $\lambda \in \mathbb{K}^\times$, or equivalently a projective frame $(p_1, \dots, p_{n+2})$ of $\mathcal{P}$: see § 7.2. Let $p \in \mathcal{P}$ and let $X = [x_k] \in \mathbb{R}^{n+1}$ [resp. $Y = [y_k] \in \mathbb{R}^{n+1}$] denote the column vector of the homogeneous coordinates of p [resp. $f(p)$]. Note that each of M, X, and Y are all only defined up to a scalar $\lambda \in \mathbb{K}^\times$.

Proposition 7.37. *Let $f: \mathcal{P} \to \mathcal{P}$ be a projective transformation. Let $M \in \mathrm{GL}(n+1, \mathbb{K})$ denote the matrix of a linear lift $F: V \to V$ in the basis $(e_1, \dots, e_{n+1})$. Then M is uniquely defined up to $\lambda \in \mathbb{K}^\times$, and f is given in homogeneous coordinates by $Y = MX$, in the sense that $f(p) = [y_1 : \dots : y_{n+1}]$ if $p = [x_1 : \dots : x_{n+1}]$.*

Proof. This is trivial: F is given in the basis $(e_1, \dots, e_{n+1})$ by $Y = MX$, therefore so is f after passing to the quotient. ∎

Remark 7.38. In homogeneous coordinates, the components of f, equivalently F, are homogeneous polynomials of degree 1 in the variables $(x_1, \dots, x_n)$ (in contrast to rational fractions in an affine chart, see Example 7.32). This explains the term *homogenization*.

A projective transformation is uniquely determined by the image of a projective frame:

Theorem 7.39. *The image of a projective frame by a projective transformation is a projective frame. Given two projective frames (p_k) and (q_k) for $\mathcal{P}$, there exists a unique projective transformation $f\colon \mathcal{P} \to \mathcal{P}$ such that $f(p_k) = q_k$ for all k.*

Proof. Let $(p_1, \dots, p_{n+2})$ be a projective frame and let $q_1, \dots, q_{n+2} \in \mathcal{P}$. We want to examine the existence and uniqueness of a projective transformation such that $q_k = f(p_k)$. Let $e_1, \dots, e_{n+2} \in V$ such that $p_k = [e_k]$ as in Lemma 7.29. If a hyperplane of $\mathcal{P}$ contained $n+1$ of the q_k, then a hyperplane of V would contain the corresponding $F(e_k)$. Since any $n+1$ of the e_k form a basis of V (otherwise a hyperplane would contain the corresponding p_k), the range of F would be contained in a hyperplane of V, which contradicts $F \in \mathrm{GL}(V)$. This proves that $(q_1, \dots, q_{n+2})$ must be a projective hyperplane for f to exist.

The existence and uniqueness of f derives from the existence and uniqueness of a linear map when prescribing the image of a basis. Indeed, assuming (q_k) is also a projective frame, let $e'_k \in V$ such that $q_k = [e'_k]$ as in Lemma 7.29. A linear lift F of f must satisfy $f(e_k) = \lambda_k f_k$ for some $\lambda_k \in \mathbb{K}^\times$ for all $k \in \{1, \dots n+2\}$. Since $e_{n+2} = -\sum_{k=1}^{n+1} e_k$ and F is linear, we have $\lambda_{n+2} e'_{n+2} = -\sum_{k=1}^{n+1} \lambda_k e'_k$. On the other hand, $e'_{n+2} = -\sum_{k=1}^{n+1} e'_k$, therefore we obtain $\sum_{k=1}^{n+1} (\lambda_{n+2} - \lambda_k) e'_k = 0$. In conclusion we must have $\lambda_k = \lambda_{n+2} =: \lambda$ for all k. This proves that F is uniquely determined up to $\lambda \in \mathbb{K}^\times$, hence f is unique. We also obtain existence: setting $F(e_k) = e'_k$ for all $k \in \{1, \dots n+1\}$ uniquely defines an element of $\mathrm{GL}(V)$, which moreover satisfies $F(e_{n+2}) = e'_{n+2}$, so that the quotient map f verifies $f(p_k) = q_k$ for all k. ∎

Remark 7.40. Theorem 7.39, or the slight generalization that *a projective map between projective spaces of the same dimension is uniquely determined by the image of a projective frame*, is sometimes referred to as the ***first fundamental theorem of projective geometry***.

7.3.3 The projective linear group

Projective transformations of $\mathcal{P} = \mathbf{P}(V)$ form a group $\mathrm{Aut}(\mathcal{P})$ under composition, which can be identified as the quotient of $\mathrm{GL}(V)$ by the subgroup of homotheties $\mathbb{K}^\times \mathrm{id}_V$; called *projective linear group* $\mathrm{PGL}(V)$. We explain this in what follows.

The map $\mathrm{GL}(V) \to \mathrm{Aut}(\mathcal{P})$ which assigns to any $F \in \mathrm{GL}(V)$ the quotient map $[F] \in \mathrm{Aut}(\mathcal{P})$ is a group homomorphism, in other words it defines a group action of $\mathrm{GL}(V)$ on $\mathcal{P}$. The image of this homomorphism is $\mathrm{Aut}(\mathcal{P})$ by definition, and by the next proposition its kernel is composed of the ***homotheties*** of V, i.e. the nonzero scalar multiples of the identity:

Proposition 7.41. *The kernel of $\mathrm{GL}(V) \to \mathrm{Aut}\,\mathbf{P}(V)$ is the group $\mathbb{K}^\times \mathrm{id}_V$ of homotheties. More generally, for any subgroup $G < \mathrm{GL}(V)$, the kernel of $G \to \mathrm{Aut}\,\mathbf{P}(V)$ is $G \cap \mathbb{K}^\times \mathrm{id}_V$.*

Proof. Immediate consequence of Proposition 7.33. ∎

For any subgroup $G \leqslant \mathrm{GL}(V)$, it is clear $G \cap \mathbb{K}^\times \mathrm{id}_V$ is a normal subgroup of G, and we shall call the quotient group $\mathbf{P}(G) := G \big/ (G \cap \mathbb{K}^\times \mathrm{id}_V)$ the ***projective group of*** G. Note that $\mathbf{P}(G)$ is the quotient of G by the equivalence relation $F \sim G$ if and only if $F = \lambda G$ for some $\lambda \in \mathbb{K}^\times$, similarly to the definition of $\mathbf{P}(V)$. By Proposition 7.41, $\mathbf{P}(G)$ is the largest quotient of G that acts faithfully on $\mathcal{P}$. In particular, the projective group of $\mathrm{GL}(V)$, denoted $\mathrm{PGL}(V)$, is called the ***projective linear group***. Our discussion leads to:

Proposition 7.42. *The group* $\mathrm{Aut}(\mathcal{P})$ *of projective transformations of* $\mathcal{P} = \mathbf{P}(V)$ *is canonically isomorphic to* $\mathrm{PGL}(V)$.

Remark 7.43. It is classical fact that for $G = \mathrm{GL}(V)$, the subgroup $G \cap \mathbb{K}^\times \mathrm{id}_V$ coincides with the center $\mathrm{Z}(G)$. In general, $G \cap \mathbb{K}^\times \mathrm{id}_V$ is only a proper subgroup of $\mathrm{Z}(G)$: consider for instance an abelian group $G \leqslant \mathrm{GL}(V)$ (e.g. diagonal matrices). There is nothing wrong with the quotient group $G/\mathrm{Z}(G)$, but it should not be confused with the projective group of G. One could call it instead the ***inner group*** of G, since the action of G on itself by inner automorphisms yields an isomorphism between $G/\mathrm{Z}(G)$ and the group of inner automorphisms $\mathrm{Inn}\, G$. I am grateful to Andy Sanders for helping me figure this out.

7.3.4 Central projections, perspectivities, central projective maps

Central projections, perspectivities, and central projective transformations are fundamental examples of projective maps, and are instrumental in many classical theorems of projective geometry. That being said, we shall not need them in other chapters of the book.

Central projections

Let $\mathcal{P}$ be a projective space. Let $\mathcal{H} \subseteq \mathcal{P}$ be a hyperplane and let $C \in \mathcal{P}$ be a point not on $\mathcal{H}$.

Lemma 7.44. *Any line* ℓ *through* C *intersects* $\mathcal{H}$ *exactly once.*

Proof. Write $\mathcal{P} = \mathbf{P}(V)$, $\mathcal{H} = \mathbf{P}(H)$, $\ell = \mathbf{P}(P)$, and $C = [c]$. Grassmann's formula says that $\dim(H \cup P) = \dim H + \dim P - \dim(H \cap P)$. Writing $\dim H = \dim V - 1$, and $\dim P = 2$, we obtain $\dim(H \cup P) = \dim V + 1 - \dim(H \cap P)$. *A priori* $\dim(H \cap P) \in \{0, 1, 2\}$, but 0 is excluded because it would imply $\dim(H \cup P) > \dim V$. Also, $\dim(H \cap P) = 2$ is excluded because $P \subsetneq H$, since c is in P but not in H. We conclude that $\dim(H \cap P) = 1$, which proves that $\mathcal{H} \cap \ell = \mathbf{P}(H \cap P)$ is a point. ∎

Definition 7.45. The ***central projection*** $p_{\mathcal{H},C} \colon \mathcal{P} - \{C\} \to \mathcal{H}$ is the map which assigns to any point $M \neq C$ the intersection of the line CM with $\mathcal{H}$.

Remark 7.46. The affine notion of central projection seen in Remark 7.21 is extended by Definition 7.45: the latter restricts to the former in any affine chart; conversely, the former extends to the latter in the projective completion.

The next proposition guarantees that central projections are projective linear maps:

Proposition 7.47. *Let $\mathcal{P} = \mathbf{P}(V)$ be a projective space, $\mathcal{H} = \mathbf{P}(H)$ be a hyperplane, and $C = [c]$ a point not on $\mathcal{H}$. The central projection $p_{\mathcal{H},C} \colon \mathcal{P} - \{C\} \to \mathcal{H}$ is the projectivization of the linear projection $p_{H,[c]} \colon V \to H$ on H parallel to $[c]$.*

Proof. Write $\mathcal{P} = \mathbf{P}(V)$, $H = \mathbf{P}(H)$, $C = [c]$. Since $\dim V = \dim H + \dim[c]$ and $H \cap [c] = \{0\}$, we have $V = H \oplus [c]$. This shows that the projection $p_{H,[c]} \colon V \to H$ is well-defined: any vector $v \in V$ is uniquely written as $v = v_H + \lambda c$, with $p_{H,[c]}(v) =:= v_H \in H$.

Let $M = [v] \in \mathcal{P}$ be any point. The line $\ell = CM$ is the projectivization of the plane $P = [v] + [c]$. Writing $v = v_H + \lambda c$, we have $P = [v_H] + [c]$. It follows that the intersection of P with H is $[v_H]$, in other words $p_{\mathcal{H},C}(M) = \left[p_{H,[c]}(v)\right]$. ∎

Perspectivities

Let $\mathcal{P} = \mathbf{P}(V)$ be a projective space and let $\mathcal{H}_1, \mathcal{H}_2 \subseteq \mathcal{P}$ be two hyperplanes.

Definition 7.48. A ***perspectivity*** $f \colon \mathcal{H}_1 \to \mathcal{H}_2$ is the restriction to $\mathcal{H}_1$ of a central projection $\mathcal{P} \to \mathcal{H}_2$.

Remark 7.49. It is implicitly assumed in the definition above that the center of the projection does not lie on $\mathcal{H}_1$, otherwise f would not be well-defined on all $\mathcal{H}_1$. The center cannot lie on $\mathcal{H}_2$ either by definition of a central projection.

Proposition 7.50. *Perspectivities are projective linear isomorphisms.*

Proof. The restriction of a projective linear map to a projective subspace is projective linear. Since $\mathcal{H}_1$ and $\mathcal{H}_2$ have same dimension, it remains to show that f is well-defined everywhere on $\mathcal{H}_1$, which is the case by assumption. ∎

Figure 7.2 can be interpreted as an illustration of a perspectivity (from the line ℓ to ℓ').

Central projective transformations

Let $\mathcal{P} = \mathbf{P}(V)$ be a projective space and $f \colon \mathcal{P} \to \mathcal{P}$ be a projective transformation.

Theorem 7.51. *Assume f is not the identity. The following are equivalent:*

*(i) f admits a **center**, i.e. a point $C \in \mathcal{P}$ such that any line through C is preserved by f.*

*(ii) f admits an **axis**, i.e. a hyperplane $\mathcal{H} \subseteq \mathcal{P}$ fixed pointwise by f.*

Moreover, when a center and axis exist, they are both unique (if $\dim \mathcal{P} > 1$ for the center).

Proof. Assume that f has a center $C = [c]$. Let F denote a linear lift of f. The fact that f preserves any line through C means that for any $v \in V$, the plane (or line) spanned by v and c is preserved by F. In particular there exists $\lambda_v, \mu_v \in \mathbb{K}$ such that $F(v) = \lambda_v v + \mu_v c$. For v not on $[c]$, λ_v and μ_v are uniquely determined; and by considering $F(v + w)$, it is not too hard to

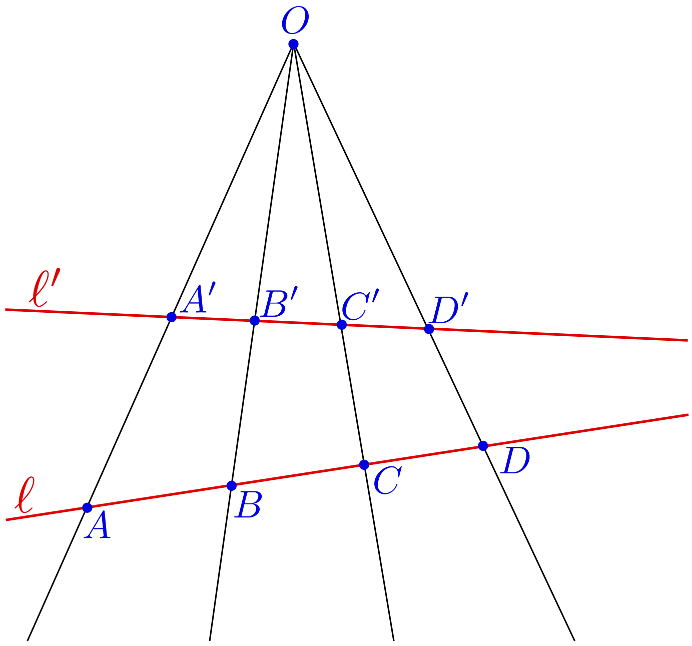


Figure 7.2: Two sets of points in perspectivity.

argue that $\lambda_v =: \lambda$ must be constant and $\mu_v =: \varphi(v)$ is a linear function of v. Note that φ is not zero, otherwise f would be the identity, therefore its kernel is a hyperplane $H \subseteq V$, and is the λ-eigenspace of F. Its projectivization $\mathcal{H}$ is fixed by f.

Conversely, assume f admits an axis $\mathcal{H} = \mathbf{P}(H)$. Denote F a linear lift of f, then H is the λ-eigenspace of F for some $\lambda \in \mathbb{K}^\times$. If F has an eigenvector $c \notin H$, then $V = H \oplus \mathbb{K}c$ and one quickly checks that c is a center. Otherwise, λ is the only eigenvalue of F, and any $v \notin H$ satisfies $F(v) = \lambda v + c$ for some $c \in H$. With the decomposition $V = H \oplus \mathbb{K}c$, one quickly checks that $[c]$ is a center of f.

Uniqueness of the axis is easily proved with linear algebra, bearing in mind that two distinct vector hyperplanes span the whole vector space. Assume $\dim \mathcal{P} \geqslant 2$ and f admits two distinct centers C_1, C_2. Let $M \in \mathcal{P}$ be any point not on the line C_1C_2, so that the lines C_1M and C_2M are well-defined and distinct. Since both are preserved by f, we must have $f(M) = M$. It quickly follows that f is the identity, contrary to the assumption. ∎

Remark 7.52. If $\dim \mathcal{P} \geqslant 2$, the center C must be fixed by f, since any two distinct lines through C are preserved and f preserves incidence.

Definition 7.53. A projective transformation $f : \mathcal{P} \to \mathcal{P}$ having a center (equivalently, an axis) is called a ***central*** projective transformation.

Remark 7.54. A central projective transformation is not to be confused with a central projection (Definition 7.45) nor a perspectivity (Definition 7.48). Traditionally, they are called ***central collineations***, but we emphasize that they are projective linear maps rather than the weaker notion of collineation (see § 7.3.5). The two notions coincide in this case: any collineation with a center is projective linear [BR98, Thm. 3.6.1].

Remark 7.55. The center of a central projective transformation f may or may not lie on the axis: f is called an ***elation*** if it does and a ***homology*** otherwise.

If $\mathcal{P}$ is (embedded as) a hyperplane in an another projective space $\mathcal{P}'$, then central projective transformations of $\mathcal{P}$ admit the further classical characterization:

Theorem 7.56. *Let $\mathcal{P}$ be a hyperplane in a projective space $\mathcal{P}'$. A map $f \neq \mathrm{id}\colon \mathcal{P} \to \mathcal{P}$ is a central projective transformation if and only if $f = h \circ g$ is the composition of two perspectivities $g\colon \mathcal{P} \to \mathcal{Q}$ and $h\colon \mathcal{Q} \to \mathcal{P}$.*

Proof. The composition $f = h \circ g$ of two perspectivities $\mathcal{P} \to \mathcal{Q} \to \mathcal{P}$ fixes $\mathcal{P} \cap \mathcal{Q}$, which is a hyperplane of $\mathcal{P}$. This shows that f is central by Theorem 7.51.

Conversely, let $f\colon \mathcal{P} \to \mathcal{P}$ be a central projective map with center C and axis $\mathcal{H}$. Let $\mathcal{Q} \neq \mathcal{P}$ be another hyperplane containing $\mathcal{H}$ and take any perspectivity $g\colon \mathcal{P} \to \mathcal{Q}$. Let $C' = g(C)$. Let D be any point on the line CC' distinct from C and C', and let $h\colon \mathcal{Q} \to \mathcal{P}$ be the perspectivity with center D. Then $h \circ g$ is a central projective transformation of $\mathcal{P}$ with center C and axis $\mathcal{H}$. Central projective transformations with a given center and axis form a one-parameter family, see Exercise 7.7. Consequently, while $h \circ g$ and f are not necessarily equal on the nose, D can be adjusted on the line CC' so that $h \circ g$ and f coincide. A few more explanations are needed to make this argument rigorous but we spare the details. ∎

The next result is sometimes known as the ***third fundamental theorem of projective geometry***:

Theorem 7.57. *Let $\mathcal{P} = \mathbf{P}(V)$ be a projective space. Any projective transformation $f\colon \mathcal{P} \to \mathcal{P}$ is a finite product of central projective transformations.*

Proof. Let $F \in GL(V)$ be a linear lift of f. Choose a basis of V so that F is represented by a matrix $M \in \mathrm{GL}(n+1, \mathbb{K})$. It is a standard theorem of linear algebra that any invertible matrix is a finite product of elementary matrices. By definition, an ***elementary matrix*** is either:

- A ***transvection matrix*** (also called ***shear matrix***), i.e. a matrix obtained from the identity matrix by replacing a 0 entry by some $\lambda \in \mathbb{K}^\times$.
- A ***dilation matrix***, i.e. a matrix obtained from the identity matrix by replacing a 1 entry by some $\lambda \neq 1 \in \mathbb{K}^\times$.
- A ***transposition matrix***, i.e. a matrix obtained from the identity matrix by switching two rows.

(Even if you are not aware of this theorem, you apply it constructively every time you solve a linear system!)

It is easy to see that all elementary matrices have a $+1$-eigenspace of codimension 1, in other words they fix a hyperplane. It follows that the corresponding projective transformations have an axis, therefore are central by Theorem 7.51. ∎

In view of Theorem 7.56, if $\mathcal{P}$ is embedded as a hyperplane in another projective space, we further obtain:

Corollary 7.58. *Any projective transformation of $\mathcal{P}$ is a finite product of perspectivities.*

Remark 7.59. Traditionally, a ***projectivity*** is defined as a finite product of perspectivities. The previous result says that *projectivity* is a perfect synonym of *projective transformation*.

7.3.5 Collineations

By definition, a collineation is a bijective map that preserves alignment of points:

Definition 7.60. Let $\mathcal{P}, \mathcal{P}'$ be projective spaces. A map $f\colon \mathcal{P} \to \mathcal{P}'$ is called a ***collineation*** if f is bijective and $f(\ell)$ is a line whenever ℓ is a line.

Remark 7.61. Collineations are well-suited to the synthetic (or axiomatic) approach to projective geometry: the definition above immediately makes sense, whereas defining projective linear maps is difficult without referring to a vector space.

Let us focus on $\mathcal{P}' = \mathcal{P}$ and denote $\mathcal{P} = \mathbf{P}(V)$. Clearly, any projective transformation of $\mathcal{P}$ is a collineation: this is an immediate consequence of Proposition 7.36. The converse is not true, as show the following counter-examples:

- If $\dim \mathcal{P} = 1$, any bijective map $f\colon \mathcal{P} \to \mathcal{P}$ is a collineation.
- Let V be a $\mathbb{C}$-vector space and assume $F\colon V \to V$ is antilinear: $F(v+w) = F(v) + F(w)$ and $F(\lambda v) = \bar{\lambda} F(v)$ for all $\lambda \in \mathbb{C}$, $v, w \in V$. (Equivalently, F is the composition of a linear map with complex conjugation.) Then $f = [F]$ preserves alignment.

Remark 7.62. The second example may be generalized for a projective space $\mathcal{P} = \mathbf{P}(V)$ over any field: Assume $F\colon V \to V$ satisfies $F(v+w) = F(v) + F(w)$ and $F(\lambda v) = \sigma(\lambda) F(v)$ for all $\lambda \in \mathbb{C}$ and $v, w \in V$, where $\sigma\colon \mathbb{K} \to \mathbb{K}$ is a nontrivial field automorphism. Such a map F is called ***semilinear***, and its projectivization $f = [F]\colon \mathcal{P} \to \mathcal{P}$ is called ***projective semilinear***.

The next theorem, known as the *(second) fundamental theorem of real projective geometry*, guarantees that there are no other counter-examples:

Theorem 7.63. *Let $\mathcal{P}$ be a projective space of dimension > 1. A bijective map $f\colon \mathcal{P} \to \mathcal{P}$ is a collineation if and only if it is projective semilinear.*

As can be expected, the proof of Theorem 7.63 is not trivial. The curious readers can read it in the excellent [BR98], although it is far from necessary as far as this book is concerned.

Corollary 7.64. *Let $\mathcal{P}$ be a real projective space of dimension > 1. In this setting, the terms "collineation" and "projective transformation" are perfect synonyms.*

Proof. Recall or prove as an exercise that $\mathbb{K} = \mathbb{R}$ has no nontrivial field automorphisms. (Hint: prove that a field automorphism σ is the identity on $\mathbb{Q}$ and is increasing.) Conclude. ∎

Remark 7.65. Theorem 7.63 implies that the group of collineations of $\mathcal{P} = \mathbf{P}(V)$ is isomorphic to the ***projective semilinear group*** $\mathrm{P\Gamma L}(V)$, i.e. the projective group of the group $\Gamma\mathrm{L}(V)$ of semilinear automorphisms of V. Those of you who like algebra will promptly agree that $\Gamma\mathrm{L}(V)$ is a split extension of $\mathrm{GL}(V)$ by the group of field automorphisms of $\mathbb{K}$ (i.e. the Galois group $\mathrm{Gal}(\mathbb{K}/k)$ where k is the prime subfield of $\mathbb{K}$), although this extension is not canonical (for $\mathbb{K} = \mathbb{C}$, it amounts to the choice of a real structure on V).

7.4 The projective line

Let us take a closer look at the 1-dimensional case: let $\mathcal{P}$ be a projective line, in other words $\mathcal{P} = \mathbf{P}(V)$ where V is a 2-dimensional vector space over a field $\mathbb{K}$.

7.4.1 Coordinates

Let us review § 7.2 in the case of the projective line $\mathcal{P} = \mathbf{P}(V)$.

Choosing a basis of V amounts to choosing an isomorphism $V \approx \mathbb{K}^2$, which induces an identification $\mathcal{P} \approx \mathbb{K}\mathbf{P}^1$. A point of $\mathcal{P}$ is represented by homogeneous coordinates $[X : Y]$, where X and Y are elements of $\mathbb{K}$ that are not simultaneously 0.

Choose the "hyperplane at infinity" $Y = 0$: it contains a single point $[1 : 0]$, which we denote ∞. Following § 7.2.2, we get an affine chart $\varphi = z \colon \mathbb{K}\mathbf{P}^1 - \{\infty\} \to \mathbb{K}$ defined by $[X : Y] \mapsto \frac{X}{Y}$. We call this the ***standard affine chart*** (or ***standard affine coordinate***) on $\mathbb{K}\mathbf{P}^1$. This allows us to identify the projective line $\mathbb{K}\mathbf{P}^1$ with the ***extended line*** $\hat{\mathbb{K}} := \mathbb{K} \cup \{\infty\}$. In homogeneous coordinates, this identification is given by $[X : Y] \mapsto z = \frac{X}{Y}$, with the convention that that $\frac{X}{0} = \infty$ for $X \neq 0$. Let us record this:

Proposition 7.66. *The standard affine chart $[X : Y] \mapsto z = \frac{X}{Y}$, extended by $\frac{1}{0} = \infty$, induces an identification $\mathbb{K}\mathbf{P}^1 \approx \hat{\mathbb{K}}$ between the projective line $\mathbb{K}\mathbf{P}^1$ and the extended line $\hat{\mathbb{K}} = \mathbb{K} \cup \{\infty\}$.*

When $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$, the extended line $\hat{\mathbb{K}}$ can be given the topology of the one-point compactification of $\mathbb{K}$, and it is an elementary exercise of topology to show that the identification $\mathbb{K}\mathbf{P}^1 \approx \hat{\mathbb{K}}$ is a homeomorphism. For $\mathbb{K} = \mathbb{R}$, the extended line $\hat{\mathbb{R}}$ is a topological circle. For $\mathbb{K} = \mathbb{C}$, the extended line $\hat{\mathbb{C}}$ is a topological 2-sphere. By the discussion of § 7.2.2, $\mathbb{C}\mathbf{P}^1 \approx \hat{\mathbb{C}}$ is a complex-analytic manifold, known as the ***Riemann sphere***. The identification $\mathbb{C}\mathbf{P}^1 \approx \hat{\mathbb{C}} \approx S^2$ and its relation to the *Hopf fibration* is further discussed in Exercise 7.5.

By definition, a projective frame of a projective line consists of 3 distinct points. For $\mathcal{P} = \mathbb{K}\mathbf{P}^1$, the standard projective frame is the triple of points $[1 \colon 0]$, $[0 \colon 1]$, $[1 \colon 1]$. In the standard affine coordinate z, this is the triple of points $\infty, 0, 1$. Let us put this on the record:

Proposition 7.67. *The standard projective frame of $\mathcal{P} = \mathbb{K}\mathbf{P}^1$, in the standard affine chart $z = \frac{X}{Y}$, is the triple of points $(\infty, 0, 1)$.*

7.4.2 Projective transformations

Assume $\mathcal{P} = \mathbb{K}\mathbf{P}^1$. Following § 7.3, a projective transformation $f \colon \mathcal{P} \to \mathcal{P}$ coincides with an element of $\mathrm{PGL}(2, \mathbb{K})$. In other words, it is given by a matrix $M = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ in $\mathrm{GL}(2, \mathbb{K})$, unique up to scalar multiplication.

In homogeneous coordinates, f is given by $f([X \colon Y]) = [aX + bY \colon cX + dY]$. In the standard affine chart z described above, this is rewritten:

$$f(z) = \frac{az + b}{cz + d}\,. \tag{7.1}$$

Note that we could have defined a map f from the extended line $\hat{\mathbb{K}}$ to itself by the expression (7.1) above, without any knowledge of projective transformations. Such maps are called ***linear fractional***. We have thus established that:

Proposition 7.68. *Under the identification $\mathbb{K}\mathbf{P}^1 \approx \hat{\mathbb{K}}^1$ provided by the standard affine chart, projective transformations of $\mathbb{K}\mathbf{P}^1$ correspond to fractional linear transformations of $\hat{\mathbb{K}}$.*

Remark 7.69. For any $M \in \mathrm{GL}(2, \mathbb{K})$, denote f_M the fractional linear transformation as above. It is an elementary exercise, which does not require knowledge of projective spaces, to check that $M \mapsto f_M$ is a group homomorphism. However, the "deep" reason is Proposition 7.68.

Since a projective frame of $\mathbb{K}\mathbf{P}^1$ is a triple of 3 distinct points, Theorem 7.39 says that for any triples of distinct points (p_1, p_2, p_3) and (q_1, q_2, q_3) in $\mathbb{K}\mathbf{P}^1$, there is a unique projective transformation such that $f(p_j) = q_j$. There is a special terminology for this property:

Theorem 7.70. *The action of $\mathrm{PGL}(2, \mathbb{K})$ on $\mathbb{K}\mathbf{P}^1$ is **simply** 3-**transitive**.*

Remark 7.71. It is worth noting once again that while Theorem 7.70 can be checked by direct computation, we proved it more elegantly with projective geometry.

7.4.3 Cross-ratios

Let $\mathcal{P}$ be a projective line and let a, b, c, d be four distinct points on $\mathcal{P}$. By Theorem 7.39, there exists a unique projective linear map $f \colon \mathcal{P} \to \mathbb{K}\mathbf{P}^1 \approx \hat{\mathbb{K}}$ which sends the triple (a, b, c) to the standard projective frame $(\infty, 0, 1)$.

Definition 7.72. The ***cross-ratio*** of four distinct points a, b, c, d on a projective line $\mathcal{P}$ is the "number" $[a, b, c, d] \in \hat{\mathbb{K}}$ equal to the image of d by the unique projective linear map $\mathcal{P} \longrightarrow \mathbb{K}\mathbf{P}^1 \approx \hat{\mathbb{K}}$ which maps (a, b, c) to $(\infty, 0, 1)$.

Remark 7.73. When $\mathcal{P} = \mathbb{K}\mathbf{P}^1$, an equivalent definition is that $[a, b, c, d] \in \mathbb{K}\mathbf{P}^1$ is the point whose homogeneous coordinates relative to the standard projective frame $(\infty, 0, 1)$ are equal to the homogeneous coordinates of d relative to the projective frame (a, b, c).

Proposition 7.74. *Under the identification $\mathbb{K}\mathbf{P}^1 \approx \hat{\mathbb{K}}$ given by the standard affine chart, the cross-ratio of four distinct points is given by:*

$$[z_1, z_2, z_3, z_4] = \frac{(z_3 - z_1)(z_4 - z_2)}{(z_3 - z_2)(z_4 - z_1)} .$$

Proof. It is easy to prove the formula by guessing the right fractional linear transformation: see Exercise 7.10. However, let us be stubborn and write a "projective" proof based on Remark 7.73.

Denote a, b, c, d the points of $\mathbb{K}\mathbf{P}^1$ corresponding to z_1, z_2, z_3, z_4, so that if we have homogeneous coordinates $a = [a_1 : a_2]$, etc, then $z_1 = \frac{a_1}{a_2}$, etc.

In order to work in the frame (a, b, c), we remember that up to a multiplicative scalar, there exists a unique choice of λ_1, λ_2, λ_3 such that $\lambda_1(a_1, a_2) + \lambda_2(b_1, b_2) = \lambda_3(c_1, c_2)$. We can solve this for λ_1 and λ_2, this is a 2×2 linear system which admits the unique solution

$$\lambda_1 = \frac{c_1 b_2 - c_2 b_1}{a_1 b_2 - a_2 b_1} \lambda_3 \qquad \lambda_2 = \frac{a_1 c_2 - a_2 c_1}{a_1 b_2 - a_2 b_1} \lambda_3 . \tag{7.2}$$

Call $e_1 = \lambda_1(a_1, a_2)$ and $e_2 = \lambda_2(b_1, b_2)$. By definition, d has homogeneous coordinates $[k_1 : k_2]$ relative to the frame (a, b, c) means that

$$\lambda(d_1, d_2) = k_1 e_1 + k_2 e_2 \tag{7.3}$$

for some scalar λ, which can be chosen equal to 1 by scaling (k_1, k_2). Again, (7.3) can be solved for k_1 and k_2:

$$k_1 = \frac{d_1 b_2 - d_2 b_1}{a_1 b_2 - a_2 b_1} \frac{\lambda_3}{\lambda_1} \qquad k_2 = \frac{a_1 d_2 - a_2 d_1}{a_1 b_2 - b_1 a_2} \frac{\lambda_3}{\lambda_2} . \tag{7.4}$$

Remark 7.73 says that $[z_1, z_2, z_3, z_4] = \frac{k_1}{k_2}$. With (7.4) and (7.2) we find

$$[z_1, z_2, z_3, z_4] = \frac{d_1 b_2 - d_2 b_1}{a_1 d_2 - a_2 d_1} \frac{\lambda_2}{\lambda_1} = \frac{(d_1 b_2 - d_2 b_1)(a_1 c_2 - a_2 c_1)}{(a_1 d_2 - a_2 d_1)(c_1 b_2 - c_2 b_1)} .$$

Dividing the numerator and denominator by $a_2 b_2 c_2 d_2$ yields the result. ■

A fundamental property of cross-ratios is their invariance under projective maps:

Theorem 7.75. *Let $\mathcal{P}$ be a projective line. For any four distinct points $a, b, c, d \in \mathcal{P}$ and for any projective map $f : \mathcal{P} \to \mathcal{P}'$,*

$$[f(a), f(b), f(c), f(d)] = [a, b, c, d] .$$

Proof. We can safely assume that $\mathcal{P}'$ is a projective line and f is bijective: just put $\mathcal{P}' = f(\mathcal{P})$. Let $f_0 : \mathcal{P} \to \mathbb{K}\mathbf{P}^1$ be the unique projective map that sends (a, b, c) to $(\infty, 0, 1)$. By definition of the cross-ratio, $[a, b, c, d] = f_0(d)$. Define $f_1 : \mathcal{P}' \to \mathbb{K}\mathbf{P}^1$ by $f_1 = f_0 \circ f^{-1}$ and observe that f_1 sends $(f(a), f(b), f(c))$ to $(\infty, 0, 1)$. By definition of the cross-ratio, $[f(a), f(b), f(c), f(d)] = f_1(f(d))$. Since $f_1(f(d)) = f_0(d) = [a, b, c, d]$, we are done. ■

Remark 7.76. Appreciate the elegance of the proof of Theorem 7.75 compared to a proof by direct computation.

Remark 7.77. Theorem 7.75 implies that the formula for the cross-ratio (Proposition 7.74) holds on any projective line equipped with any affine chart.

Note that in any projective space, the cross-ratio of four distinct collinear points is well-defined. (Also, the cross-ratio of any four concurrent hyperplanes, by projective duality.) The next theorem is an immediate consequence of Theorem 7.75:

Theorem 7.78. *Projective linear maps preserve the cross-ratios of* 4*-tuples of collinear points.*

Remark 7.79. The converse of Theorem 7.78 is true: see Exercise 7.12.

Example 7.80. In Figure 7.2, the cross-ratios $[A, B, C, D]$ and $[A', B', C', D']$ must be equal, since the two 4-tuples differ by a perspectivity, which is a projective linear by Proposition 7.50.

Example 7.81. Exercise 7.11 features an application of Theorem 7.78 to metrology borrowed from Wikipedia [Wik21a]: using cross-ratios to measure real-world dimensions from a photo.

7.5 Quadrics

In this section, we define projective and affine quadrics and discuss some basic properties. This introduction to quadrics is far from exhaustive! For a more thorough treatment, I recommend Samuel's book [Sam86; Sam88] or Berger's [Ber77; Ber09], or the more approachable book of Audin [Aud03; Aud06]. It is useful to first review quadratic forms in § 3.1.

7.5.1 Homogeneous functions

Let V be a finite-dimensional vector space over a field $\mathbb{K}$.

A function $f : V \to \mathbb{K}$ is called ***polynomial*** if, for some (equivalently any) linear coordinate system $(x_1, \ldots, x_m)$ (where $m = \dim V$), f coincides with a polynomial of m variables with coefficients in $\mathbb{K}$. As an example, $f(x, y, z) = x^2yz^3 - 2xz^4$ is a polynomial function of total degree 6 on $V = \mathbb{K}^3$.

A polynomial of several variables is called ***homogeneous of degree*** d if it is a sum of monomials, each of degree d. For instance, $P(X,Y) = X^2Y^3 - XY^4 + 2Y^5$ is homogeneous of degree 5. This notion can be generalized to arbitrary functions on a vector space:

Definition 7.82. A function $f\colon V \to \mathbb{K}$ is called ***homogeneous of degree*** d if $f(\lambda v) = \lambda^d f(v)$ for all $\lambda \in \mathbb{K}$ and $v \in V$.

When f is a polynomial function on V and $\mathbb{K}$ is infinite, it is an elementary exercise of algebra to check that the two notions of homogeneity coincide.

Recall that a *linear form* is a linear map $f\colon V \to \mathbb{K}$, and that we call ***quadratic form*** a function that can be written $q(v) = B(v,v)$ where $B\colon V \times V \to \mathbb{K}$ is a symmetric bilinear form. B is uniquely determined by q by the *polarization formula* $B(u,v) = \frac{1}{2}\left(q(u+v) - q(u) - q(v)\right)$ (unless $\mathbb{K}$ has characteristic 2). Review § 3.1 for more details on quadratic forms.

Proposition 7.83. *Let V be a finite-dimensional vector space over $\mathbb{K}$ (with* $\operatorname{char}\mathbb{K} \neq 2$*).*

$$\begin{aligned} f \text{ polynomial + homogeneous of degree } 1 &\Leftrightarrow f \text{ is a linear form} \\ f \text{ polynomial + homogeneous of degree } 2 &\Leftrightarrow f \text{ is a quadratic form} \end{aligned}$$

Proof. This is an elementary exercise of linear algebra (do it!). ∎

Remark 7.84. Proposition 7.83 admits a natural generalization for any degree: consider a symmetric multilinear form $F\colon V \times \cdots \times V \to \mathbb{K}$ and let $f(v) = F(v, \dots, v)$. Then f is polynomial and homogeneous of degree d. Conversely, any polynomial and homogeneous function is of this form. As for polarization: $F(v_1, \dots, v_d)$ can be recovered as the coefficient of $t_1 \dots t_d$ in the polynomial $p(t_1, \dots, t_d) \coloneqq f(t_1v_1 + \dots t_dv_d)$.

Remark 7.85. Not all homogeneous functions are polynomial: for instance, consider the functions $f(x,y,z) = (x^3+y^3+z^3)^{\frac{1}{3}}$ on $\mathbb{R}^3$ and $g(z,w) = \frac{z^pw^qz\bar{w}}{|z|^2+|w|^2}$ on $\mathbb{C}^2$.

Homogeneous functions are relevant to projective geometry in the following way. A homogeneous function $f\colon V \to \mathbb{K}$ is not constant on vector lines unless it has degree 1, so it does not induce a function on $\mathbf{P}(V)$. That being said, the zero level set $\mathrm{Z}(f) \coloneqq \{v \in V \colon f(v) = 0\}$ is invariant by scalar multiplication: it is clear that $f(\lambda v) = \lambda^d f(v) = 0$ if and only if $f(v) = 0$. In other words, $\mathrm{Z}(f)$ is a union of vector lines; such a set is called a ***(linear) cone***. The projectivization of any such cone is a well-defined subset of $\mathbf{P}(V)$. In summary:

Definition 7.86. If $f\colon V \to \mathbb{K}$ is a homogeneous function, the set $\mathbf{P}(\{f = 0\})$ is well-defined subset of $\mathbf{P}(V)$, called ***projectivized cone*** of f.

Remark 7.87. In the language of algebraic geometry, the projectivized cone of any homogeneous polynomial is a projective variety. More generally, a ***projective variety*** is the projectivization of the common zero set of a finite family of homogeneous polynomials.

Example 7.88. On $V = \mathbb{R}^3$, the function $q(x, y, z) = x^2 + y^2 - z^2$ is a homogeneous polynomial of degree 2, i.e. a quadratic form. Its projectivized cone $\mathcal{C}$ is the subset of $\mathbb{R}\mathbf{P}^2$ with equation $x^2 + y^2 - z^2 = 0$ in homogeneous coordinates. In the standard affine chart $z = 1$, the equation of $\mathcal{C}$ becomes $x^2 + y^2 = 1$: it is the unit circle in this affine plane.

Any polynomial can be ***homogenized*** by adding an extra dimension:

Lemma 7.89. *Assume $\mathbb{K}$ is infinite. For any $P \in \mathbb{K}[X_1, \dots, X_n]$ of degree d, there exists a unique homogeneous $\hat{P} \in \mathbb{K}[X_1, \dots, X_{n+1}]$ of degree d such that $P(X_1, \dots, X_n) = \hat{P}(X_1, \dots, X_n, 1)$:*

$$\hat{P}(X_1, \dots, X_n, X_{n+1}) = X_{n+1}^d P\left(\frac{X_1}{X_{n+1}}, \dots, \frac{X_n}{X_{n+1}}\right). \tag{7.5}$$

Proof. The expression $X_{n+1}^d P\left(\frac{X_1}{X_{n+1}}, \dots, \frac{X_n}{X_{n+1}}\right)$ is *a priori* a rational fraction, but it is actually a polynomial because P has total degree d, so that the denominator is cancelled by X_{n+1}^d. Thus $\hat{P}$ is a well-defined polynomial, and it is quickly checked that it has the required qualities.

Uniqueness of $\hat{P}$ is clear "for $X_{n+1} \neq 0$": the identity (7.5), seen as an identity between functions on $\mathbb{K}^{n+1}$, must hold for $x_{n+1} \neq 0$ by homogeneity of $\hat{P}$. In order to conclude "by continuity", we need a slightly subtle algebraic argument: if two polynomials coincide on the complement of $\{x_{n+1} = 0\}$, then they must be equal. This is true if $\mathbb{K}$ is infinite; it is a generalization (that can be proven by induction) of the fact that a polynomial of one variable is determined by its values on any infinite set. In the language of algebraic geometry: any nonempty Zariski open set of $\mathbb{K}^{n+1}$, such as $\{x_{n+1} \neq 0\}$, is dense in the Zariski topology. ∎

Example 7.90. The recipe to homogenize any polynomial (function) is clear in practice:

$$\begin{aligned} x - y^3 &\longrightarrow xz^2 - y^3 \\ ax^2 + bx + c &\longrightarrow ax^2 + bxy + cy^2 \\ x^3y^7 - z^4 + 1 &\longrightarrow x^3y^7 - z^4t^6 + t^{10} \end{aligned}$$

Lemma 7.89 can be generalized as follows:

Theorem 7.91. *Let H be an affine space and $\mathcal{P} = \mathbf{P}(V)$ its projective completion. Any polynomial function $f\colon H \to \mathbb{K}$ admits a unique homogeneous extension $\hat{f}\colon V \to \mathbb{K}$ of the same degree. (Here the base field $\mathbb{K}$ is assumed infinite.)*

By definition, the function $\hat{f}$ in Theorem 7.91 is the ***homogenization*** of f.

Proof. The proof is a straightforward extension of Lemma 7.89 by choosing an appropriate coordinate system. Let $\vec{H}$ denote the vector space underlying H. The projective completion of H can be constructed as $\mathcal{P} = \mathbf{P}(V)$ where $V = \vec{H} \times \mathbb{K}$, and H is identified as $\vec{H} \times \{1\}$: see § 7.1.5 for details. As usual, choose a coordinate system $(x_1, \dots, x_{n+1})$ on V such that H has equation $x_{n+1} = 1$: this is achieved by choosing any basis $(e_1, \dots, e_n)$ of $\vec{H}$, and completing it with $e_{n+1} = (0, \dots, 0, 1)$. Now, a polynomial function f of degree d on H is given by a polynomial $P \in \mathbb{K}[X_1, \dots, X_n]$ in our coordinate system: for any $x \in H$ with coordinates $(x_1, \dots, x_n, 1)$ in V, we have $f(x) = P(x_1, \dots, x_n)$. Denote by $\hat{P}$ the homogenization of P as in (7.5). Then $\hat{f}(x) := \hat{P}(x_1, \dots, x_{n+1})$ is a homogeneous extension of f. Uniqueness quickly follows from the uniqueness of Lemma 7.89. ■

Let H be an affine space and denote $\mathcal{P}$ its projective completion. It follows from our discussion that if $S \subseteq H$ is the zero set of some polynomial function $f\colon H \to \mathbb{K}$, then S admits a natural ***projective completion*** $\hat{S} \subseteq \mathcal{P}$, defined as the projectivization of the zero set of the homogenization $\hat{f}$ of f.

Example 7.92. The set $y^2 = x^3 + ax + b$ is a cubic in $\mathbb{K}^2$ called ***elliptic curve***. Its projective completion is the set in $\mathbb{K}\mathbf{P}^2$ defined by $y^2 z = x^3 + axz^2 + bz^3$ in homogeneous coordinates.

7.5.2 Projective quadrics

Let $\mathcal{P} = \mathbf{P}(V)$ be a projective space. We have seen that given a homogeneous function $f\colon V \to \mathbb{K}$, its projectivized cone $\mathcal{C} := \mathbf{P}(\{f = 0\}) \subseteq \mathcal{P}$ is well-defined. As a special case, when $f = q$ is a homogeneous polynomial of degree 2 i.e. a quadratic form (Proposition 7.83), $\mathcal{C}$ is called a *quadric*. If $\dim \mathcal{P} = 2$, the term *conic* is preferred. Let us record these definitions:

Definition 7.93. Let $\mathcal{P} = \mathbf{P}(V)$ be a projective space over a field $\mathbb{K}$ (with $\operatorname{char} \mathbb{K} \neq 2$). Let $q\colon V \to \mathbb{K}$ be a quadratic form, and let $\mathcal{C} := \mathbf{P}(\{f = 0\})$ denote its projectivized cone.

- $\mathcal{C}$ is called a ***(projective) conic*** if $\dim \mathcal{P} = 2$.
- $\mathcal{C}$ is called a ***(projective) quadric*** if $\dim \mathcal{P} \geqslant 3$ ($\dim \mathcal{P} = 1, 2$ can also be tolerated).

We call the quadric $\mathcal{C}$ ***nondegenerate*** if the symmetric bilinear form B associated to q is nondegenerate, and ***proper*** if it is nondegenerate and nonempty.

The quadratic form q associated to a quadric $\mathcal{C}$ is not unique: any scalar multiple λq (with $\lambda \in \mathbb{K}^\times$) defines the same quadric. Unfortunately, this is not the only reason for non-uniqueness: for instance, the quadratic forms $q_1(x, y, z) = x^2 + y^2$ and $q_2(x, y, z) = x^2 + 2y^2$ define the same conic $\mathcal{C} = \{[0 : 0 : 1]\} \subseteq \mathbb{R}\mathbf{P}^2$. For this reason, "serious" books define a quadric by its equation rather than as a subset of $\mathcal{P}$—this is a standard approach in algebraic geometry. Thankfully, the situation is not so terrible over $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$:

Theorem 7.94. *Let $q_1, q_2\colon V \to \mathbb{K}$ be two quadratic forms having the same zero set, i.e. defining the same quadric $\mathcal{C} \subseteq \mathbf{P}(V)$.*

- *If $\mathbb{K} = \mathbb{C}$, then $q_2 = \lambda q_1$ for some $\lambda \in \mathbb{K}^\times$.*
- *If $\mathbb{K} = \mathbb{R}$ and $\mathcal{C}$ is proper, the same conclusion holds.*

(More generally, the conclusion holds if $\mathbb{K}$ is algebraically closed or if $\mathcal{C}$ has a simple point.)

Proof. For $\mathbb{K} = \mathbb{C}$ or any algebraically closed field, the theorem can be seen as a simple application of Hilbert's *Nullstellensatz*, a fundamental theorem of algebraic geometry, although this is a case of using a sledge hammer to kill a fly[2]. A more pedestrian proof consists in first proving the theorem for a projective line, and second the general case by restricting to any line. I leave out the details, which can be found in Berger [Ber09, Thm. 14.1.6.2].

An elementary algebraic proof of the theorem when $\mathcal{C}$ has a simple point can be found in Samuel's book (Thm. 46 in [Sam86] or [Sam88]). I suggest the following alternative proof when $\mathbb{K} = \mathbb{R}$ and q has mixed signature[3], which is a weaker assumption than $\mathcal{C}$ proper (why?). Let $x \in V$. If x is isotropic for q_1 or q_2, there is nothing to show. Otherwise, there exists $y \in V$ such that q has signature $(1, 1)$ in restriction to the 2-plane spanned by x and y. Thus it is (nearly) enough to prove the theorem on any such 2-plane, which is a straightforward exercise of algebra. ∎

Our knowledge of symmetric bilinear forms will prove useful to study projective quadrics. For instance, when $\mathbb{K} = \mathbb{R}$, Sylvester's law of inertia (Theorem 3.19) yields:

Theorem 7.95. *Let $\mathcal{C} \subseteq \mathcal{P} = \mathbf{P}(V)$ be a quadric in a real projective space. There exists a unique unordered pair of nonnegative integers p, q with $p + q \leqslant \dim V$ such that, in suitable homogeneous coordinates, $\mathcal{C}$ is given by the equation*

$$x_1^2 + \cdots + x_p^2 - x_{p+1}^2 - \cdots - x_{p+q}^2 = 0\,. \tag{7.6}$$

When $\mathbb{K} = \mathbb{C}$, quadratic forms on V are classified by their rank (Corollary 3.16), therefore we obtain:

Theorem 7.96. *Let $\mathcal{C} \subseteq \mathcal{P} = \mathbf{P}(V)$ be a quadric in a complex projective space. There exists a unique integer $p \leqslant \dim V$ such that, in suitable homogeneous coordinates, $\mathcal{C}$ is given by*

$$x_1^2 + \cdots + x_p^2 = 0\,. \tag{7.7}$$

We call (7.6) [resp. (7.7)] the ***normal form*** of the quadric $\mathcal{C}$. Note that $\mathcal{C}$ is nondegenerate if and only if $p + q = \dim V$ [resp. $p = \dim V$].

Remark 7.97. Theorem 7.95 [resp. Theorem 7.96] can be described as a classification theorem because it implies that two quadrics differ by a projective transformation if and only if they have same signature [resp. rank]. See Exercise 7.13.

When $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$, we have seen that $\mathcal{P}$ is an smooth (analytic) manifold, and it is legitimate to wonder whether quadrics are submanifolds of $\mathcal{P}$. In general, quadrics can have singularities: for instance, the quadric in $\mathbb{R}\mathbf{P}^2$ defined by $q(x, y, z) = x^2 - y^2$ is the union of two intersecting lines–this is not a smooth manifold of $\mathbb{R}^2$. That being said, proper quadrics are always nice hypersurfaces:

[2]But according to US Marine Corps Major I. L. Holdridge: "Sometimes killing a fly with a sledge hammer is entirely appropriate. It doesn't make the fly any more dead, but the rest of the flies sure sit up and take notice".

[3]I guess this is more or less the proof that Berger had in mind for Exercise 14.8.1 in [Ber09].

Theorem 7.98. *Let $\mathcal{P} = \mathbf{P}(V)$ be a real or complex projective space. Any proper quadric in $\mathcal{P}$ is an analytic hypersurface, i.e. an analytic submanifold of codimension* 1.

Proof. Let $q \colon V \longrightarrow \mathbb{K}$ be the quadratic form defining $\mathcal{C}$ (unique up to scalar by Theorem 7.94) and B the associated symmetric bilinear form. The function $q \colon V \longrightarrow \mathbb{K}$ is analytic (because it is polynomial) and its differential at any point $v_0 \in V$ is the linear form $B(v_0, \cdot) \colon V \longrightarrow \mathbb{K}$. Since B is nondegenerate, $B(v_0, \cdot)$ is not the zero linear form unless $v_0 = 0$. Hence q is a submersion in restriction to $V - \{0\}$. By a standard theorem of differential geometry, which applies both in the real- and complex-analytic categories, it follows that $Z := q^{-1}(0) - \{0\}$ is an analytic hypersurface in V. Another standard theorem guarantees that since the action of $\mathbb{K}^*$ on $V - \{0\}$ is analytic, free, and proper[4], and preserves the hypersurface $Z \subseteq V$, it descends to an analytic hypersurface $\mathcal{C} \subseteq \mathcal{P}$ when taking the quotient. ■

Remark 7.99. As a by-product of the proof above, we obtain that the linear tangent space to $\{q = 0\}$ at a point $v_0 \in V - \{0\}$ is the kernel of the linear form $B(v_0, \cdot)$. When taking the projectivization, we obtain the projective hyperplane "tangent" to $\mathcal{C}$. Let us give an example: consider the conic $x^2 + y^2 - z^2 = 0$ in $\mathbb{R}\mathbf{P}^2$. The projective line tangent to $\mathcal{C}$ at $[x_0 : y_0 : z_0]$ is the line with equation $x_0 x + y_0 y - z_0 z$ in homogeneous coordinates. In any affine chart, the image of this line gives the affine tangent space to the conic. For instance, in the affine chart $z = 1$, $\mathcal{C}$ is the circle $x^2 + y^2 = 1$, and the tangent line at (x_0, y_0) has equation $x_0 x + y_0 y = 1$.

7.5.3 Real projective conics and quadrics

Conics

Assume $\mathcal{P}$ is a real projective plane ($\mathbb{K} = \mathbb{R}$ and $\dim \mathcal{P} = 2$). Theorem 7.95 implies that there are only two normal forms of nondegenerate conics up to sign:

$$x^2 + y^2 + z^2 = 0$$
$$x^2 + y^2 - z^2 = 0$$

The first conic is empty. Thus we are left with just one proper conic in normal form: the conic $\mathcal{C}$ with equation $x^2 + y^2 - z^2 = 0$. Note that this is the projectivized light cone of Minkowski space $\mathbb{R}^{2,1}$. What does $\mathcal{C}$ look like in an affine chart?

- If the line at infinity does not intersect $\mathcal{C}$, then the image of $\mathcal{C}$ in the affine chart is an ellipse. For instance, in the affine chart $z = 1$, the line at infinity $z = 0$ does not intersect $\mathcal{C}$, and the image of $\mathcal{C}$ is the circle $x^2 + y^2 = 1$.
- If the line at infinity intersects $\mathcal{C}$ once, then the image of $\mathcal{C}$ is a parabola. For instance, in the affine chart $z = x + 1$, the line at infinity $z = x$ intersects $\mathcal{C}$ at the point $[1 : 0 : 1]$, and the image of $\mathcal{C}$ is the parabola $y^2 = 1 + 2x$.

[4] We did not prove that the action of $\mathbb{K}^*$ on $V - \{0\}$ is proper, but it is a straightforward exercise of topology using the characterization for a locally compact group G acting on a Hausdorff space X: for all $x, y \in X$, there exists neighborhoods V_x, V_y such that $\{g \in G \mid g \cdot V_x \cap V_y \neq \emptyset\}$ is relatively compact [Bou71, TG III.31, Prop. 7].

- If the line at infinity intersects $\mathcal{C}$ twice, the image of $\mathcal{C}$ is a hyperbola. For instance, in the affine chart $y = 1$, the line at infinity $y = 0$ intersects $\mathcal{C}$ at the points $[1 \colon 0 \colon 1]$ and $[1 \colon 0 \colon -1]$, and the image of $\mathcal{C}$ is the hyperbola $x^2 - z^2 = -1$.

We leave it to the conscientious reader to prove carefully check these claims. The so-called "conic sections" mentioned above are illustrated in Figure 7.3.

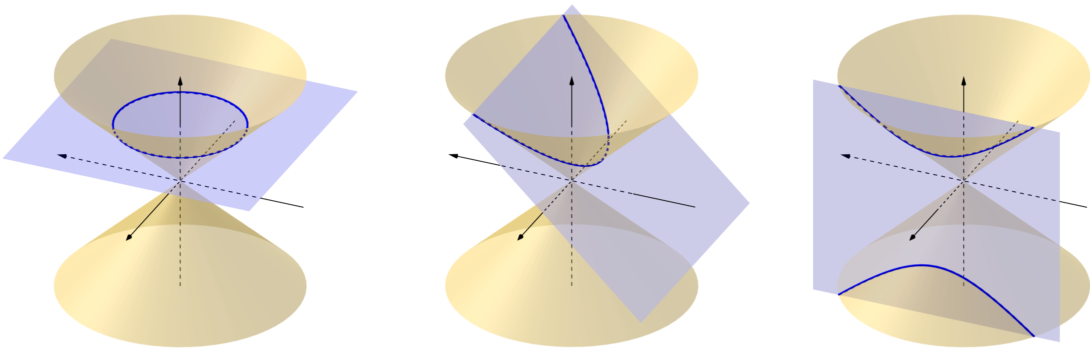


Figure 7.3: The intersection of the cone $C \colon x^2 + y^2 - z^2 = 0$ with the planes $z = 1$, $z = x + 1$, and $y = 1$ shows the image of the projective conic $\mathcal{C} = \mathbf{P}(C)$ in these affine charts: it is alternatively a circle, a parabola, and a hyperbola.

Quadric surfaces

Now let us increment the dimension: assume $\mathcal{P}$ has dimension 3. Proper quadrics in $\mathcal{P}$ are called ***quadric surfaces*** (they are smooth surfaces by Theorem 7.98). Theorem 7.95 implies that there are only three normal forms up to sign:

$$x^2 + y^2 + z^2 + t^2 = 0$$
$$x^2 + y^2 + z^2 - t^2 = 0$$
$$x^2 + y^2 - z^2 - t^2 = 0$$

The first quadric is empty. Thus we are left with two proper quadrics in normal form: $\mathcal{C}_1 \colon x^2 + y^2 + z^2 - t^2 = 0$ and $\mathcal{C}_2 \colon x^2 + y^2 - z^2 - t^2 = 0$. Note that $\mathcal{C}_1$ is the projectivized light cone of Minkowski space $\mathbb{R}^{3,1}$ (meanwhile $\mathcal{C}_2$ is the projectivized cone of $\mathbb{R}^{2,2}$). What do $\mathcal{C}_1$ and $\mathcal{C}_2$ look like in an affine chart H? This depends on the position of the plane at infinity $\mathcal{H} = \mathbf{P}(\vec{H})$:

- For $\mathcal{C}_1$: Depending on whether $\mathcal{H}$ intersects $\mathcal{C}_1$ in the empty set (e.g. $H \colon t = 1$), in a single point (e.g. $H \colon t = z + 1$), or in a proper conic (e.g. $H \colon z = 1$), the image of $\mathcal{C}_1$ in H is either an ***ellipsoid*** (e.g. $x^2 + y^2 + z^2 = 1$), an ***elliptic paraboloid*** (e.g. $x^2 + y^2 - 2z = 1$), or a ***hyperboloid of two sheets*** (e.g. $x^2 + y^2 - t^2 = -1$). See Figure 7.4.

- For $\mathcal{C}_2$: Depending on whether $\mathcal{H}$ intersects $\mathcal{C}_2$ in two lines (e.g. $H\colon t = y + 1$) or in a proper conic (e.g. $H\colon t = 1$), the image of $\mathcal{C}_1$ in H is either a ***hyperbolic paraboloid*** (e.g. $x^2 - z^2 - 2y = 1$) or a ***hyperboloid of one sheet*** (e.g. $x^2 + y^2 - z^2 = 1$). See Figure 7.5.

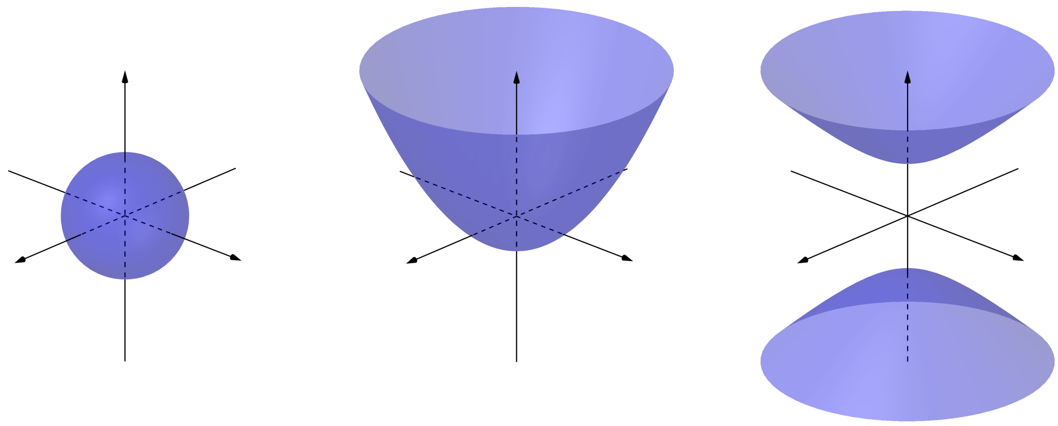


Figure 7.4: The ellipsoid $x^2 + y^2 + z^2 = 1$, the elliptic paraboloid $x^2 + y^2 - 2z = 1$, and the hyperboloid of two sheets $x^2 + y^2 - t^2 = -1$ are all images of the projective quadric $\mathcal{C}_1$ in different affine charts.

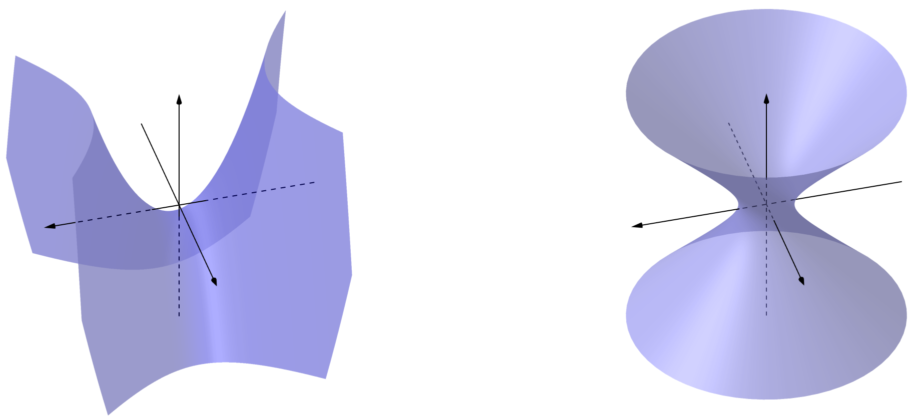


Figure 7.5: The hyperbolic paraboloid $x^2 - z^2 - 2y = 1$ and the hyperboloid of one sheet $x^2 + y^2 - z^2 = 1$ are both images of the projective quadric $\mathcal{C}_2$ in different affine charts.

Remark 7.100. One can check that the Gaussian curvature of $\mathcal{C}_1$ is everywhere positive in all three affine charts, while that of $\mathcal{C}_2$ is everywhere negative in both affine charts. This is not a coincidence: the sign of the Gaussian curvature is a projective invariant. See Exercise 7.17.

7.5.4 Projective completion of affine quadrics

Let H be an affine space over an infinite field $\mathbb{K}$.

Definition 7.101. A ***(affine) quadric*** in H is a subset $C \subseteq H$ that is the zero set of a polynomial function $f\colon H \to \mathbb{K}$ of degree 2. If $\dim H = 2$, a quadric is also called a ***conic***.

Example 7.102. The equation $x^2 - 3y^2 + 2xy - 6x + y - 7 = 0$ defines a conic in $\mathbb{R}^2$ (or in any affine plane equipped with an affine frame).

Example 7.103. Let $a, b, c > 0$. The quadric $\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1$ is an ***ellipsoid*** in $\mathbb{R}^3$.

Theorem 7.104. *Let $\mathcal{P}$ be a projective space. The image of a projective quadric $\mathcal{C} \subseteq \mathcal{P}$ in any affine chart $H \subseteq \mathcal{P}$ is an affine quadric $C \subseteq H$.*

*Conversely, let H be an affine space and $\mathcal{P}$ its projective completion. Any affine quadric $C \subseteq H$ is uniquely extended as a projective quadric $\mathcal{C} \subseteq \mathcal{P}$, called its **projective completion**.*

Proof. The first part of the theorem is the easier: a projective quadric $\mathcal{C}$ in $\mathcal{P} = \mathbf{P}(V)$ is given by the zero set of a homogeneous polynomial $f\colon V \to \mathbb{K}$ of degree 2. Any polynomial function on V restricts to a polynomial function of the same degree on H (called its ***dehomogenization***): this is easy to check by choosing a suitable system of coordinates. It follows that the image of $\mathcal{C}$ in H is the zero set of a polynomial function of degree 2. For the converse, we already did the work: see Theorem 7.91 and the discussion below the proof. ∎

An affine quadric is called *nondegenerate* or *proper* if its projective completion is.

Example 7.105. The parabola $C\colon y = x^2$ is an affine conic in the Euclidean plane $\mathbb{R}^2$. Its projective completion is the proper conic $\mathcal{C}\colon yz = x^2$ in $\mathbb{R}P^2$. Note that $\mathcal{C}$ is obtained from C by adding a single point at infinity: $[1\colon 0\colon 0]$. On the other hand:

- In an affine chart (e.g. $z = -y + 2$) where the line at infinity ($z = -y$) does not intersect $\mathcal{C}$, the image of $\mathcal{C}$ is an ellipse ($x^2 + (y-1)^2 = 1$).
- In an affine chart (e.g. $z = y + 2$) where the line at infinity ($z = y$) intersects $\mathcal{C}$ twice, the image of $\mathcal{C}$ is a hyperbola ($x^2 - (y+1)^2 = -1$).

This example illustrates that "moving" the line at infinity can change an ellipse into a hyperbola, transitioning through a parabola. This phenomenon can be visualized by intersecting a cone with different planes in $\mathbb{R}^3$: we have seen this (for a different conic) in Figure 7.3.

Example 7.106. It is shown in § 3.3 that pseudo-Euclidean spheres are proper quadrics.

We will study some features of affine and projective conics in the next section, under the pretext of solving a problem of perspective drawing. In particular, the center and axes of real affine conics will be discussed in § 7.6.3.

7.6 How to draw a wheel in perspective?

In this "bonus section", which can safely be skipped, we are going to investigate a problem of perspective geometry: how to draw a wheel? My friend Julien (the same who drew the cover!) asked me this question, or rather the more precise question posed in § 7.6.1. After thinking about it and finding some answers, I thought that a discussion of this problem would be a fitting way to conclude this chapter. It will give us the opportunity to experiment with affine and projective conics, expanding on the previous section.

7.6.1 The problem

Suppose you want to draw a wheel in perspective—maybe it is an elementary step in drawing a car. In the real 3-dimensional world, the wheel is a round circle, contained in a vertical plane; but in perspective this circle becomes an ellipse. How can one "find" and draw this ellipse? The easiest would be to know the center and the principal axes of the ellipse. We recall a mathematical definition of these notions in § 7.6.3, but they are easy to describe visually: the principal axes are the two axes of symmetry of the ellipse, and they meet at the center. The minor axis gives the shortest "diameter" of the ellipse, and the major axis gives the longest.

If you search for how to draw a wheel in perspective on the internet or in books that specialize in perspective drawing, you will quickly find images such as Figure 7.6[5]:

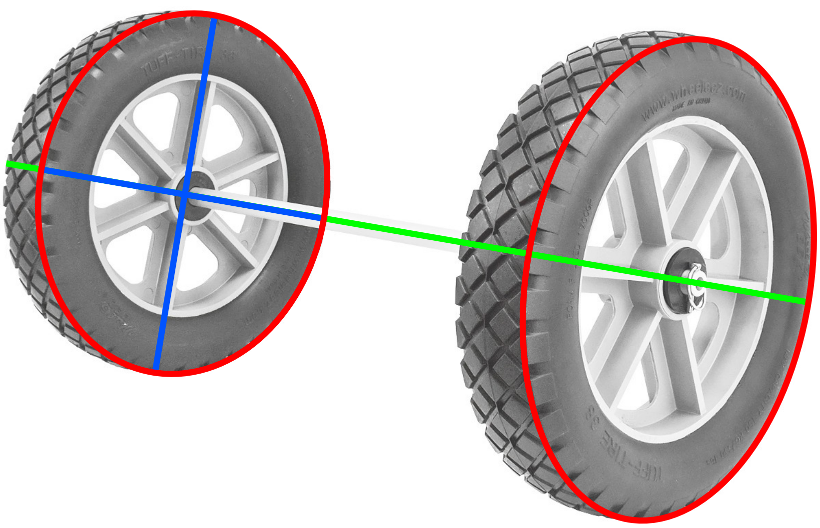


Figure 7.6: Wheels and axle

In this picture, the green line is the axis of rotation of the wheel (the axle), orthogonal to

[5] See e.g. https://courses.byui.edu/art110_new/art110/week01/minor.html for other examples of the minor axis apparently coinciding with the orthogonal direction to the circle, or https://youtu.be/LtIkcLJ129k for a tutorial video on how to draw wheels on car based on this principle.

the wheel in the 3-dimensional world; and the two blue lines are supposed to be the principal axes of the ellipse. This type of picture suggests that the minor axis of the ellipse coincides with the axis of rotation. Julien wondered if this was rigorously true:

Question 7.107. Does the minor axis of a wheel drawn in perspective always coincide with the axis of rotation?

According to internet resources or even serious books on perspective drawing, the answer is yes (see for instance [Ida10], [ske21], [Rob04; RB13], [Nor99], [Dob61]). In particular, this "fact" implies that the center of the ellipse lies on the axis of rotation, just like the center of the wheel. (Most sources do not go so far as saying that the two centers are equal, though.)

We are going to see that these facts are actually untrue: in general the minor axis does not coincide with the axis of rotation (nor is it parallel to it or has same vanishing point). As a matter of fact, the minor axis does not go through the center of the wheel. However, we shall prove that the center of the ellipse always lies on the axis of motion (the horizontal line contained in the wheel plane that goes through the center of the wheel).

7.6.2 Mathematical setup

Drawing in perspective means projecting the real world, modelled as a 3-dimensional Euclidean space E, to an affine plane D (the ***drawing plane***). Here we are not talking about the orthogonal projection, but a central projection as in Remark 7.21: we choose a point $O \in E$ not in D (the ***eye*** of the observer), and each point $M \in E$ is projected to the unique point M' defined as the intersection of the line OM with D: see Figure 7.7.

Note that the central projection is not defined on the plane parallel to D through O, although it extends to the projective completion of E as a projective linear map defined everywhere except at O. In restriction to any other plane, such as the plane containing the wheel, the central projection is called a *perspectivity*. See § 7.3.4 for details.

A slightly different point of view is to consider that the observer does not truly see the 3-dimensional world E, but only its projectivization, or its image on the drawing plane D. Indeed, two points that are aligned with the eye cannot be distinguished by the observer. In other words: the observer's representation of the world is the projective space $\mathbf{P}(E)$, or its image in the affine chart D. See § 7.2.2 for details on affine charts.

Consider Figure 7.8. According to the first viewpoint, it shows two different conics in perspectivity: the wheel (in the blue plane) and its projection (in the beige plane). The problem is: what is the effect of a perspectivity on the axes of a conic? According to the second viewpoint, Figure 7.8 shows a *single* projective conic (the yellow cone), and its image in two different affine charts. The question becomes: how to find the axes of a projective conic in an affine chart? Both approaches have their merits: the first is perhaps more natural, but the second is more simple and elegant.

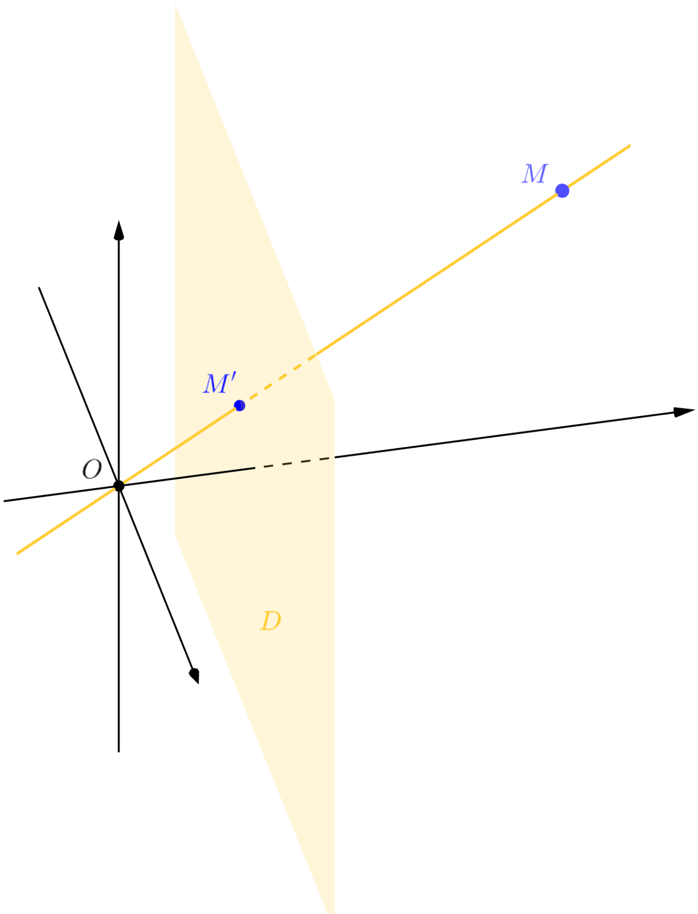


Figure 7.7: Perspective projection.

7.6.3 Center and axes of an affine conic

Center

Consider a conic C in an affine plane equipped with a frame. By definition, C is given by an equation $p(x, y) = 0$, where p is a polynomial function of degree 2 in two variables:

$$p(x, y) = a_{11}x^2 + a_{22}y^2 + 2a_{12}xy + 2a_{13}x + 2a_{23}y + a_{33}\,.$$

The polynomial coefficients are named to emphasize that p can be written

$$p(x, y) = \begin{bmatrix} x & y & 1 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}$$

i.e. $p(x, y) = {X_1}^{\mathsf{T}} A X_1$ where $X_1 = \begin{bmatrix} x & y & 1 \end{bmatrix}^{\mathsf{T}}$ and $A = \begin{bmatrix} a_{ij} \end{bmatrix}$ is a symmetric matrix. This expression makes it especially easy to write the homogenization of p, which defines the

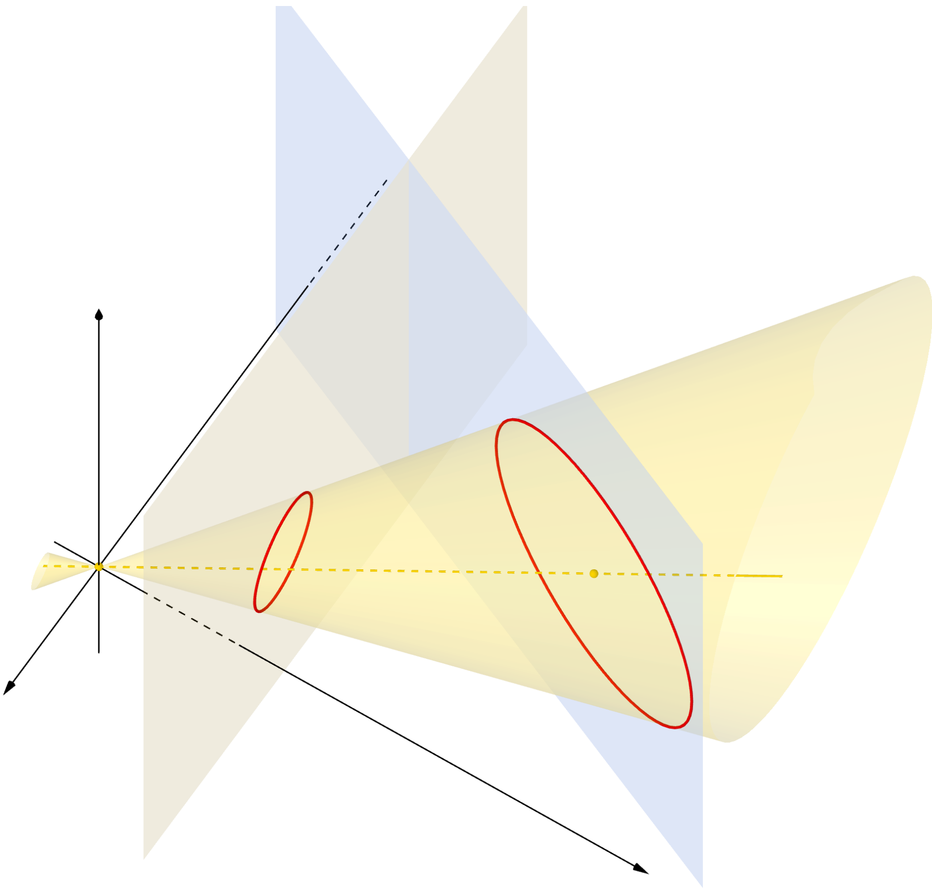


Figure 7.8: Depending on the point of view, this figure shows either: 1. Two affine conics in perspectivity, or 2. The image of a projective conic in two different affine charts.

projective completion $\mathcal{C}$ of the conic: $\hat{p}(x, y, z) = \hat{X}^{\mathsf{T}} A \hat{X}$ where $\hat{X} = \begin{bmatrix} x & y & z \end{bmatrix}^{\mathsf{T}}$. Thus A represents the symmetric bilinear form associated to $\mathbb{C}$. In order to define a proper conic, A must be invertible and of mixed signature.

Coming back to the affine version, we can write p as a sum of homogeneous polynomials of different degrees:

$$p(x, y) = p_2(x, y) + p_1(x, y) + p_0$$

where $p_2(x, y) = a_{11}x^2 + a_{22}y^2 + 2a_{12}xy$, $p_1 = 2a_{13}x + 2a_{23}y$, and $p_0 = a_{33}$. By Proposition 7.83, p_2 is a quadratic form and p_1 is a linear form (on $\mathbb{R}^2$).

Proposition 7.108. *The proper affine conic C is:*

- *An ellipse if and only if p_2 is positive or negative definite (signature $(2, 0)$ or $(0, 2)$).*
- *A hyperbola if and only if p_2 is indefinite (signature $(1, 1)$).*
- *A parabola if and only if p_2 is degenerate (signature $(1, 0)$ or $(1, 0)$).*

Proof. An easy application of Sylvester's law of inertia shows that, in suitable coordinates, C is the ellipse $x^2 + y^2 = 1$ if p_2 is positive or negative definite, the hyperbola $x^2 - y^2 = 1$ if p_2 is indefinite, and the hyperbola $x^2 - 2y = 0$ if p_2 is degenerate. (Other cases like $x^2 + y^2 = -1$, $x^2 - y^2 = 0$, $x^2 = 0$ are excluded because C is assumed proper.) ∎

The ***center*** of the conic is found by applying a translation $(x, y) \mapsto (x - x_0, y - y_0)$ that gets rid of the linear term. Indeed, if the equation $p_2(x, y) + p_1(x, y) + p_0 = 0$ has no linear term i.e. $p_1 = 0$, then it is invariant by the central symmetry $(x, y) \mapsto (-x, -y)$. This means that the origin of the frame (i.e. the point with coordinates (x_0, y_0) before applying the translation) is the center of symmetry of the conic. The center exists (and is unique) if (and only if) p_2 is nondegenerate, i.e. C is an ellipse or a hyperbola:

Proposition 7.109. *A proper affine conic admits a center if (and only if) it is not a parabola.*

Proof. Let us write $v = (x, y)$. The change of coordinates $v \mapsto v - c$ kills off the linear term in $p(v) = p_2(v) + p_1(v) + p_0$ provided that $2b(v, c) = p_1(v)$, where b is the symmetric bilinear form associated to p_2. Indeed, write $p_2(v - c) = p_2(v) - 2b(v, c) + p_2(c)$, etc. If b is nondegenerate, the map $c \mapsto b(\cdot, c)$ is an isomorphism between $V = \mathbb{R}^2$ and V^*, therefore there exists a unique $c \in V$ such that $b(\cdot, c) = \frac{1}{2}p_1$. ∎

Axes

Assume that the quadratic form p_2 is nondegenerate, i.e. the conic C is an ellipse or a hyperbola. By the previous discussion, up to a translation, the equation of C can be written $p_2(x, y) = 1$.

Let us still denote b the symmetric bilinear form associated to p_2. The spectral theorem says that b is diagonalizable in an orthonormal basis. In other words, after applying a suitable rotation to the coordinates, the equation of C becomes:

$$\lambda_1 x^2 + \lambda_2 y^2 = 1$$

where λ_1, λ_2 are the eigenvalues of b (i.e. the eigenvalues of the matrix $\left[a_{ij}\right]_{1\leqslant i,j\leqslant 2}$). The directions given by the x-axis and the y-axis in these coordinates are univocally defined: they are the two eigenspaces of b. By the definition, these lines are the ***(principal) axes*** of the conic. The spectral theorem ensures in particular that the axes are always orthogonal.

Remark 7.110. I deliberately left a small lie just above, can you find it?

If p_2 has signature $(2, 0)$, then we can put $\lambda_1 > \lambda_2 > 0$. Denoting $\lambda_1 = \frac{1}{a^2}$ and $\lambda_2 = \frac{1}{b^2}$ with $a, b > 0$, the equation of C is rewritten $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$. This is an ellipse with major axis a and minor axis b, which can be parametrized by $x = a\cos t$, $y = b\sin t$.

If p_2 has signature $(1, 1)$, then we can put $\lambda_1 > 0 > \lambda_2$. Denoting $\lambda_1 = \frac{1}{a^2}$ and $\lambda_2 = -\frac{1}{b^2}$, the equation of C is rewritten $\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1$. This is a hyperbola, which can be parametrized by $x = a\cosh t$, $y = \pm b\sinh t$.

Remark 7.111. In summary, the center is obtained by translating to kill off the linear term, and the axes by applying the spectral theorem to the quadratic term. Note that while the center is a purely affine notion, the axes rely on the Euclidean structure of the plane.

7.6.4 Center and axes of a projective conic?

Assume now that $\mathcal{C}$ is given as a projective conic in a projective plane $\mathcal{P} = \mathbf{P}(V)$. We assume that $V = E$ has a Euclidean structure, so that any affine plane $H \subseteq V$ inherits one. While the center and axes of $\mathcal{C}$ do not make sense in $\mathcal{P}$, it is reasonable to ask if one can predict the center and axes in any affine chart H. For the center, there is an elegant answer:

Proposition 7.112. *Let $\mathcal{C}$ be a proper conic in $\mathcal{P} = \mathbf{P}(V)$ and $B\colon V \times V \to \mathbb{R}$ the associated symmetric bilinear form. Let $H \subseteq V$ be an affine chart and $\vec{H}$ the underlying vector plane. The center of $\mathcal{C}$ in H is the point $c \in H$ such that $[c]$ is the B-orthogonal of $\vec{H}$.*

Remark 7.113. Despite being elementary, I was very pleased with Proposition 7.112 when I found it! It allows an elegant proof of Theorem 7.116.

Proof. Since B is nondegenerate and of mixed signature, up to sign it has signature $(2, 1)$. Let us assume that the restriction of B to $\vec{H}$ has signature $(2, 0)$, since this is the case we are interested in: the image of $\mathcal{C}$ in H is an ellipse. By Proposition 3.20, the B-orthogonal of $\vec{H}$ is a line L such that $\vec{H} \oplus L = V$. (It is instructive to examine the other cases: if B has signature $(1, 1)$ on $\vec{H}$, i.e. $\mathcal{C}$ traces a hyperbola on H, then the reader can show that $\vec{H} \oplus L = V$ is still true. What if B is degenerate on $\vec{H}$?)

Let (e_1, e_2, e_3) be a basis of V such that $e_1, e_2 \in \vec{H}$ and $e_3 = c$ is the intersection of L and H. In the associated coordinate system (x, y, z), the equation of $\vec{H}$ [resp. H] is $z = 0$ [resp. $z = 1$]. Since $\vec{H} \perp_B L$, the matrix A of the bilinear form B in this basis has a few zero entries:

$$A = \begin{bmatrix} a_{11} & a_{12} & 0 \\ a_{21} & a_{22} & 0 \\ 0 & 0 & a_{33} \end{bmatrix}$$

In other words, the equation of $\mathcal{C}$ in the homogeneous coordinates $[x\colon y\colon z]$ has no mixed terms containing z:

$$a_{11}x^2 + a_{22}y^2 + 2a_{12}xy + a_{33}z^2 = 0\,.$$

When dehomogenizing (just set $z = 1$) to obtain the equation of the conic in the affine frame $(c; e_1, e_2)$ of H, this translates to the absence of linear terms:

$$a_{11}x^2 + a_{22}y^2 + 2a_{12}xy + a_{33} = 0\,.$$

This proves that the origin of the frame, namely the point c, is the center of this conic. ■

What about the axes? The simple yet key observation is that $p_2(x, y) = a_{11}x^2 + a_{22}y^2 + 2a_{12}xy$ represents the restriction of B to $\vec{H}$. Thus the axes of the conic on H are simply given by applying the spectral theorem to this restriction.

Remark 7.114. Call (u_1, u_2) be the basis of $\vec{H}$ obtained by applying the spectral theorem to B in $\vec{H}$. Then (u_1, u_2, e_3) is a basis of V diagonalizing B. However, this is not the same as applying the spectral theorem to B in V; do you see why?

7.6.5 Resolution of the problem

Choosing coordinates

Let us come back to the wheel and develop the setup of § 7.6.2. Let us place the origin of E at the eye O (as in Figure 7.7), which allows us to identify E to a vector space.

We choose an orthonormal basis (i_0, j_0, k_0) of E adapted to the drawing plane D as follows. For k_0 we take the unit vector pointing upwards; it belongs to $\vec{D}$ since D is vertical. For j_0, we take the unit normal to D, oriented so that $\langle j_0, \overrightarrow{OM}\rangle > 0$ for all $M \in D$. More geometrically: call d_D the distance from the origin O to the plane D, given by $d_D = \|\overrightarrow{OO_D}\|$ where $O_D \in D$ is the closest point to O (its orthogonal projection). Then j_0 is the unit normal $j_0 = \frac{\overrightarrow{OO_D}}{d_D}$. Finally, i_0 is the unique vector such that (i_0, j_0, k_0) is orthonormal, i.e. $i_0 = k_0 \times j_0$ (cross-product).

The central projection on D is given by $\overrightarrow{OM'} = \lambda\overrightarrow{OM}$ with $\lambda = \frac{d_D}{\langle\overrightarrow{OM}, j_0\rangle}$ (check this). In the coordinate system associated to (i_0, j_0, k_0), the equation of D is $y = d_D$, and the projection is

$$(x, y, z) \mapsto \frac{d_D}{y}(x, y, z)\,. \tag{7.8}$$

(As expected, this is the restriction of a projective linear map $\hat{E} \longrightarrow \hat{D}$, remarkably simple in homogeneous coordinates: $[x\colon y\colon z\colon t] \mapsto [x\colon z\colon t]$.)

Of course, we could have similarly defined an orthonormal basis adapted to any other vertical plane, such as the wheel plane W. Call θ the oriented angle between D and W, so that the rotation of angle θ around the vertical axis pointing upwards sends $\vec{D}$ to $\vec{W}$. We denote $(i_\theta, j_\theta, k_\theta)$ the orthonormal frame of V adapted to W. In the previous coordinates, we have:

$$\begin{aligned} i_\theta &= (\cos\theta, \sin\theta, 0) \\ j_\theta &= (-\sin\theta, \cos\theta, 0) \\ k_\theta &= (0, 0, 1)\,. \end{aligned}$$

The ground (floor) is the horizontal plane $z = -h$ where h is the “altitude” of the eye O. For simplicity we take $h = 0$, but one could easily adapt what follows to keep track of h. We write the center of the wheel $c_W = (c_1, c_2, c_3 = R)$ where R is the radius. Since $c_W \in W$, we have $\langle c_W, j_\theta\rangle = c_2\cos\theta - c_1\sin\theta = d_W$ where d_W is the distance from O to W.

A picture

Figure 7.9 shows the wheel circle (in red) in the wheel plane and its projection on the drawing plane, as well as some notable lines:

- The ***axis of rotation*** in green: this is the line through the center of the wheel that is orthogonal to the wheel plane.
- The ***axis of motion*** in orange: this is the horizontal line through the center of the wheel that is contained in the wheel plane.
- In gray: the four lines tracing the vertical square in which the wheel is inscribed, the two diagonals of this square meeting at the center, the vertical line through the center.

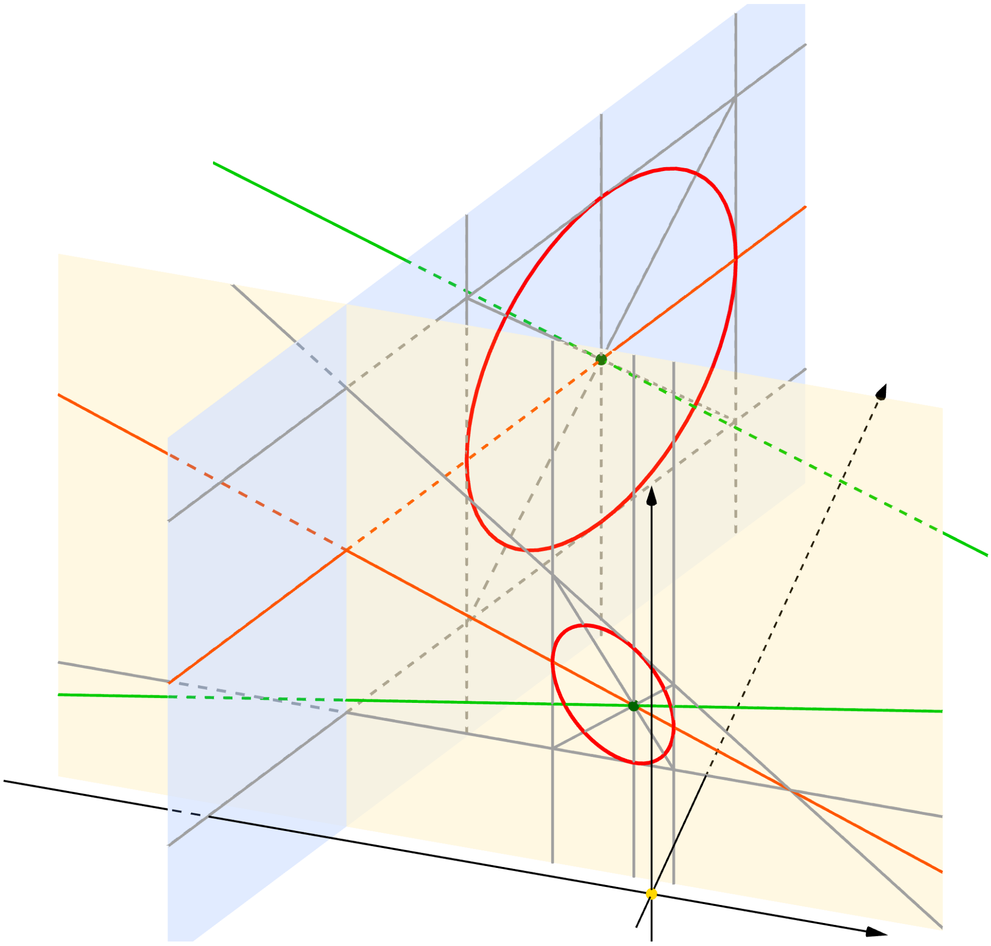


Figure 7.9: Projection of the wheel on the drawing plane.

Resolution

After all the preparation, the plan is clear:

1. Write the equation of the wheel: first as an affine conic in W, then homogenize to obtain the projective conic in $\mathbf{P}(E)$ represented by the yellow cone in Figure 7.8.
2. Find the center and axes of the conic in the drawing plane by applying § 7.6.3.

For the first step, we do not need coordinates: a point $v \in W$ is on the wheel if and only if $\|v - c_W\|^2 = R^2$, hence the polynomial function $p\colon W \longrightarrow \mathbb{R}$ defining our affine conic: $p(v) = \|v - c_W\|^2 - R^2$. To homogenize p into a quadratic form $q\colon V \longrightarrow \mathbb{R}$, we write $q(v) = \frac{1}{\lambda^2} q(\lambda v)$ and choose λ so that $\lambda v \in W$. We obtain $q(v) = \frac{1}{\lambda^2} p(\lambda v) = \frac{1}{\lambda^2}\left(\|\lambda v - c_W\|^2 - R^2\right)$. Substituting $\lambda = \frac{d_W}{\langle v, j_\theta\rangle}$ yields the expression of the quadratic form:

$$q(v) = \|v\|^2 - \frac{2}{d_W}\langle v, j_\theta\rangle\langle v, c_W\rangle + \frac{\|c_W\|^2 - R^2}{d_W^2}\langle v, j_\theta\rangle^2 .$$

It is easy to guess the polarization of q:

$$B(v, w) = \langle v, w\rangle - \frac{1}{d_W}\left[\langle v, j_\theta\rangle\langle w, c_W\rangle + \langle w, j_\theta\rangle\langle v, c_W\rangle\right] + \frac{\|c_W\|^2 - R^2}{d_W^2}\langle v, j_\theta\rangle\langle w, j_\theta\rangle . \tag{7.9}$$

For the second step, we need to:

(a) Find the center: find $c_D \in D$ such that $[c_D]$ is the B-orthogonal of $\vec{D}$.

(b) Find the axes: apply the spectral theorem to the restriction of $\vec{B}$ to $\vec{D}$.

These are basic exercises of linear algebra, and not the most interesting part of this section, but let us go ahead since we have come this far (also, all of projective geometry is linear algebra, is basically the point of this chapter).

(a) We look for $c_D = (x_0, y_0 = d_D, z_0)$ such that $B(v, c_D) = 0$ for all $v = (x, y = 0, z)$. Think of $B(v, c_D)$ as a linear function of x and z whose coefficients must vanish. This yields a system of two linear equations in x_0 and z_0, which is of course straightforward to solve. Let us spare the details and record the result:

Proposition 7.115. *The center of the ellipse on the drawing plane is $c_D = (x_0, y_0, z_0)$ with:*

$$\begin{aligned} x_0 &= d_D \frac{d_W(c_1\cos\theta - c_2\sin\theta) + (\|c\|^2 - 2R^2)\sin\theta\cos\theta}{d_W^2 + 2d_W c_1 \sin\theta + (\|c\|^2 - 2R^2)\sin^2\theta} \\ y_0 &= d_D \\ z_0 &= R\frac{d_D\cos\theta - (\sin\theta)x_0}{d_W} \end{aligned}$$

In contrast, the position of the center of the wheel on the drawing plane is, by (7.8), the point with coordinates $d_D(\frac{c_1}{c_2}, 1, \frac{R}{c_2})$. In general, the two are distinct, but note that when $\theta = 0$, we have $c_2 = d_W$, and both are $\frac{d_D}{d_W}(c_1, d_W, R)$. This is the situation where the wheel is parallel to the drawing plane, and we expect the projection of the wheel to be a circle.

(b) For the axes, we must find the eigenvectors of the symmetric bilinear form (7.9) restricted to $\vec{D}$, i.e. $y = 0$. In the coordinate system (x, z), the matrix of B is:

$$\begin{bmatrix} 1 + \frac{2c_1\sin\theta}{d_W} + \frac{\|c\|^2 - R^2}{d_W^2}\sin^2\theta & \frac{R\sin\theta}{d_W} \\ \frac{R\sin\theta}{d_W} & 1 \end{bmatrix}$$

Of course, we could easily find the two eigenvalues and eigenvectors of this matrix, thus giving an expression of the principal axes of the ellipse. This computation and its result would not be very insightful though, so let us save the space.

Instead, let us show a picture: Figure 7.10 is the same configuration as Figure 7.8, but only shows the drawing plane. Clearly, the axes of the ellipse (in blue) have nothing to do with the axis of rotation (in green). The minor axis does not even go through the center of the wheel.

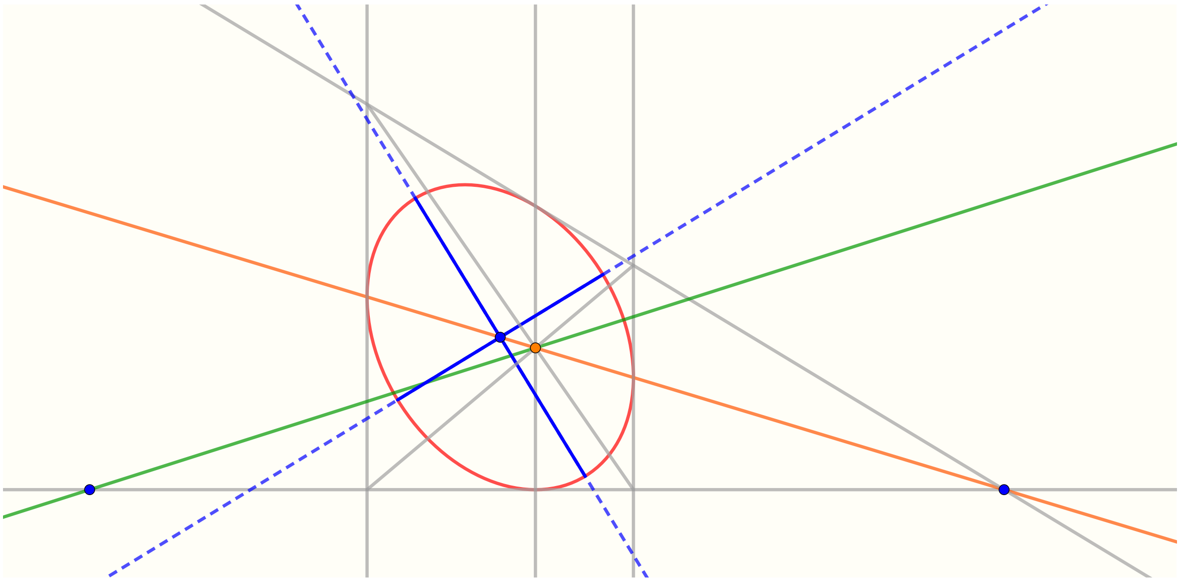


Figure 7.10: The wheel on the drawing plane.

In mathematics, answering *no* to a question like Question 7.107 ("is it always true that …") is often not very exciting, because it is enough to produce one counter-example. Let us instead try to prove a positive result: it seems on Figure 7.10 that the center of the ellipse belongs to the axis of motion (in orange). Is this always true?

7.6.6 A theorem of perspective geometry

Theorem 7.116. *When drawing a wheel in perspective, the center of the ellipse always lies on the image of the axis of motion. (In Figure 7.10: the blue axes intersect on the orange line.)*

Proof. The situation is as follows: we have a projective conic $\mathcal{C} \subseteq \mathbf{P}(E)$ and two vertical affine planes D and W in E. The centers of the conic on these two planes are c_D and c_W. We want to show that two vector planes $P, Q \subseteq E$ are equal:

- The vector plane containing the axis of motion, that is $P = [c_W] \oplus [i_\theta]$.
- The vector plane containing both centers, that is $Q = [c_W] \oplus [c_D]$.

(Indeed, if $P = Q$, then $[c_D] \subseteq P$, which in the affine chart D translates to c_D lying on the image of the axis of motion.)

The key observation is that $[c_W] = (\vec{W})^{\perp_B}$ and $[c_D] = (\vec{D})^{\perp_B}$, therefore $Q = (\vec{W} \cap \vec{D})^{\perp_B}$. The intersection of $\vec{W}$ and $\vec{D}$ is the vertical line $[k_0]$ (unless $\theta = 0$, in which case we saw that $[c_D] = [c_W]$). The result of this discussion is that we need to show that $P = [k_0]^{\perp_B}$, which amounts to $c_W \perp_B k_0$ and $i_\theta \perp_B k_0$. We already know that $c_W \perp_B k_0$ since $k_0 \in \vec{W}$. It remains to show that $B(i_\theta, k_0) = 0$, which is child's play with the expression of B (7.9). ■

Remark 7.117. In fact, the expression of the B-orthogonal of k_0 is quite simple: check that $B(v, k_0) = \left\langle v, k_0 - \frac{R}{d_W} j_\theta \right\rangle$ for all $v \in V$; it it follows that $k_0^{\perp_B}$ is the regular orthogonal of the vector $k_0 - \frac{R}{d_W} j_\theta$, which is the vector plane with equation:

$$z - \frac{R}{d_W}(y \cos\theta - x \sin\theta) = 0\,. \tag{7.10}$$

As I claimed, $i_\theta = (\cos\theta, \sin\theta, 0)$ satisfies this equation. But the point I want to make here is that (7.10) is the same as the equation giving z_0 in Proposition 7.115. This is not a coincidence: the two equations relating x_0 and z_0 express that $c_D \perp_B i_0$ and $c_D \perp_B k_0$.

7.6.7 Conclusion

Using nifty techniques from projective geometry, we saw how to calculate the center and axes of a wheel drawn in perspective. We debunked the idea that the minor axis coincides with the axis of rotation, but proved that the center of the ellipse lies on the axis of motion.

Unfortunately, our findings will be of little help to a real person who is trying to draw a wheel only equipped with a pencil and ruler (and maybe a compass). I think my friend Julien was a bit disappointed when I told him my solution ("remember the spectral theorem…") because he was expecting a more practical answer.

It is in fact possible to ask the precise question: on a 2D drawing, starting with the horizon, two vanishing points giving the directions parallel/orthogonal to the wheel, and the four lines tangent to the wheel as in Figure 7.11; can one find the center and axes of the ellipse by a straightedge and compass construction? In theory, the answer is yes: it is a famous theorem (due to Pierre-Laurent Wantzel in 1837 [Wan37]) that points and lines in the Euclidean plane are constructible if and only if they can be calculated by solving a finite number of linear and quadratic equations. Our work has shown that this is the case for the center and axes of the ellipse. The practical implementation of Wantzel's theorem is most likely horrible, though.

Remark 7.118. In the setup of § 7.6.5, it is easy to find the homogenized equations of:

- The axis of motion: $(R\sin\theta)x - (R\cos\theta)y + d_W z = 0$.
- The axis of rotation: $(R\cos\theta)x + (R\sin\theta)y - (c_1\cos\theta + c_2\sin\theta)z = 0$.

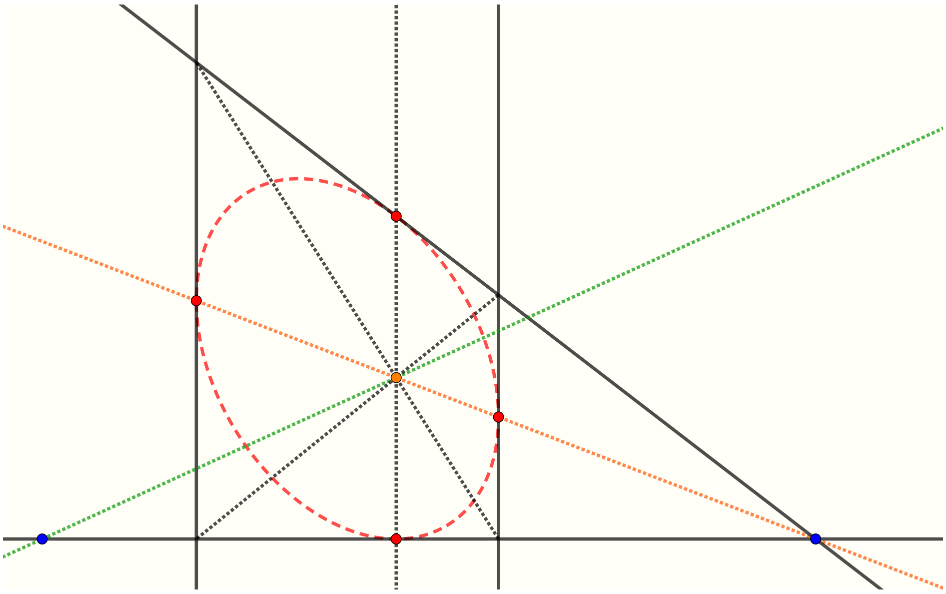


Figure 7.11: Drawing the wheel by hand?

For the equation in the drawing plane, just put $y = d_D$. Putting $z = 0$ yields the coordinates of the two vanishing points: $x_1 = d_D \cot\theta$ and $x_2 = -d_D \tan\theta$. Conversely, the positions of these two points allows one to recover d_D $(= \sqrt{-x_1 x_2})$ and $\tan\theta$ $(= \sqrt{-x_2/x_1})$. The values of $\cos\theta$ and $\sin\theta$ are derived from $\tan\theta$ by quadratic equations. The position of the center of the wheel gives us c_1, c_2, R. Finally, $d_W = c_2\cos\theta - c_1\sin\theta$. Thus all the parameters of the problem can be read off Figure 7.11, and the theorem of Wantzel is fully applicable.

Remark 7.119. Remark 7.118 confirms that the ellipse is determined by the four gray tangency lines and a fifth line through its center. This is reminiscent of the famous theorem (mentioned in Exercise 7.16) that a conic is uniquely determined by five tangency lines.

In practice, there are more useful tips to consider. First of all, it is easy to find the actual center of the wheel (not of the ellipse) by tracing the two diagonals of the square: see dotted gray lines in Figure 7.11. From the two vanishing points, we can then trace the axis of motion (dotted orange) and the axis of rotation (dotted green). The axis of motion and the vertical line through the center give the four points of tangency with the square. Pretty good! In order to freehand draw a decent ellipse, four points is not enough, but it is possible to find eight more by tracing more lines: see Figure 7.12. I learned this trick from [Dob61].

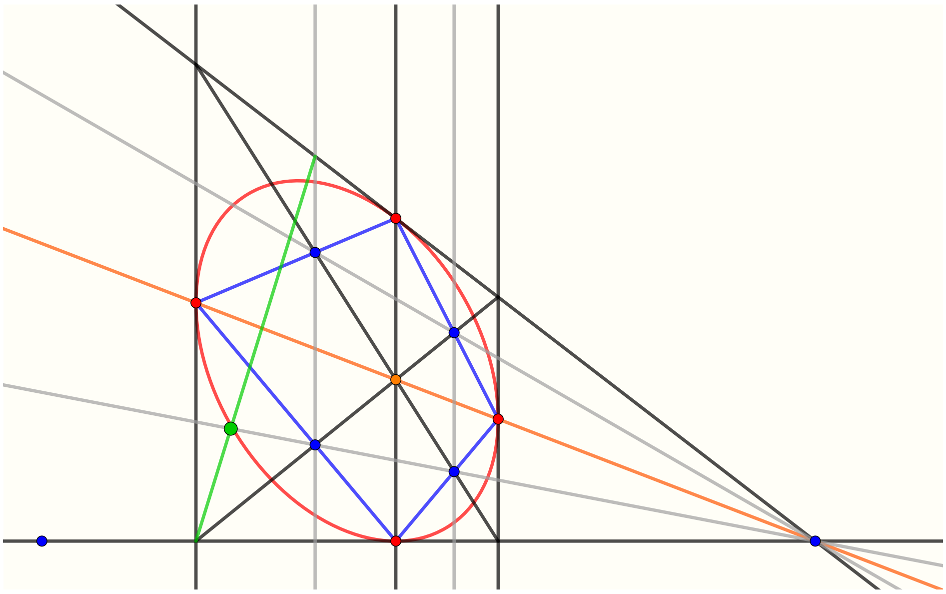


Figure 7.12: Finding eight more points on the ellipse: First draw the four new blue lines between the red points of tangency, then the four new gray lines between the blue points of intersection, and finally the eight possible green segments similar to the one that is shown.

7.7 Exercises

Exercise 7.1. Projective duality

Let V be a finite-dimensional vector space and $\mathcal{P} := \mathbf{P}(V)$ its projectivization.

(1) What is projective duality?

(2) For $x \in V$, denote δ_x the linear form on V^* defined by $\delta_x(\varphi) = \varphi(x)$. Prove that $x \mapsto \delta_x$ is an isomorphism between V and its bidual V^{**}. Derive that there is an isomorphism of projective spaces $\mathcal{P} \approx \mathcal{P}^{**}$. Prove Proposition 7.14:

Proposition . *Projective duality is a bijective correspondence between projective subspaces of $\mathcal{P}$ and projective subspaces of $\mathcal{P}^* = \mathbf{P}(V^*)$, and it is decreasing with respect to inclusion. Moreover, projective duality is involutive in the sense that [complete the sentence].*

(3) Let $\mathcal{P}$ be a projective plane. Prove that any two lines of $\mathcal{P}$ intersect at a unique point: first write a direct proof, then offer an alternate proof using projective duality.

Exercise 7.2. Axioms of projective geometry

Let $\mathbb{K}$ be any field and let V be a vector space over $\mathbb{K}$. Check that the projective space $\mathbf{P}(V)$ satisfies the Veblen–Young axioms of projective geometry (see § 7.1.4).

Exercise 7.3. Some elementary properties of homogeneous coordinates

Let $\mathcal{P}$ be a projective space of dimension n.

(1) Show that a choice of homogeneous coordinates on $\mathcal{P}$ amounts to the choice of a projective isomorphism $\mathcal{P} \approx \mathbb{K}\mathbf{P}^n$.

(2) Consider two systems of homogeneous coordinates $[x_k]$ and $[y_k]$. Show that there exists $P \in \mathrm{GL}(n+1, \mathbb{K})$ and $\lambda \in \mathbb{K}^\times$ such that $[x_1, \dots, x_{n+1}]^\mathsf{T} = \lambda P [y_1, \dots, y_{n+1}]^\mathsf{T}$.

(3) Explain why homogeneous coordinates on $\mathcal{P}$ induce "dual" homogeneous coordinates on $\mathcal{P}^*$. Show that the dual of the point $[a_1 : \dots : a_{n+1}]$ is the projective hyperplane with equation $a_1 x_1 + \dots + a_{n+1} x_{n+1} = 0$.

Exercise 7.4. Barycentric coordinates vs homogeneous coordinates

Let H be an affine space of dimension n. Denote $\vec{H}$ the underlying vector space.

(1) Let $A_1, \dots, A_N$ be points in H and let $\alpha_1, \dots, \alpha_N$ be weights (real numbers, not all zero). Denote $\alpha := \sum_{k=1}^N \alpha_k$.
   (a) Assume $\alpha \neq 0$. Show there exists a unique point $G \in H$, called ***barycenter***, such that $\sum_{k=1}^N \alpha_k \overrightarrow{GA_k} = \vec{0}$. Check that $\sum_{k=1}^N \alpha_k \overrightarrow{OA_k} = \alpha \overrightarrow{OG}$ for any $O \in H$.
   (b) If $\mathbb{K} = \mathbb{R}$, show that G is the unique minimizer of the function $F\colon H \to \mathbb{R}$ defined by $F(M) = \frac{1}{2} \sum_{k=1}^N \alpha_k \, d(M, A_k)^2$. *Hint: show that* $\operatorname{grad} F(M) = -\sum_{k=1}^N \alpha_k \overrightarrow{MA_k}$.

(c) Show that if $\alpha = 0$, then there is a well-defined vector $G \in \vec{H}$ such that $G = \sum_{k=1}^{N} \alpha_k \overrightarrow{OA_k}$ for any $O \in H$.

(d) Show that if H is an affine hyperplane in a vector space V, then $G = \frac{1}{\alpha} \sum_{k=1}^{N} \alpha_k A_k$ if $\alpha \neq 0$ and $G = \sum_{k=1}^{N} \alpha_k A_k$ if $\alpha = 0$. Show that $[G]$ is a well-defined element of $\mathbf{P}(V)$ and $[G] = \left[\sum_{k=1}^{N} \alpha_k A_k\right]$.

(2) Let $A_1, \ldots, A_{n+1}$ be points in H that are affinely independent, i.e. not contained in a hyperplane. Show that for all $M \in H$, there exists $(\alpha_1, \ldots, \alpha_{n+1})$ such that M is the barycenter of $\{(A_k, \alpha_k)\}$. Show that the tuple $(\alpha_1, \ldots, \alpha_{n+1})$ is unique up to multiplication by $\lambda \in \mathbb{K}^\times$. Such a $(n+1)$-tuple is called ***barycentric coordinates*** for M.

(3) Show that if H is an affine hyperplane in V, then $A_1, \ldots, A_{n+1} \in H$ are affinely independent if and only if $(A_1, \ldots, A_{n+1})$ is a basis of V. Show that for all $M \in H$, barycentric coordinates for M are the same thing as homogeneous coordinates for $[M]$.

(4) Conclude that barycentric coordinates in an affine space are homogeneous coordinates in the projective completion. Show that conversely, homogeneous coordinates in a projective space are barycentric coordinates in the appropriate affine patch.

Exercise 7.5. The complex projective line and the Hopf fibration

(1) Let us identify $\mathbb{R}^3 \approx \mathbb{C} \times \mathbb{R}$. Consider the stereographic projection from the North pole $s\colon S^2 - \{N\} \to \mathbb{C}$ (see e.g. § 9.3.1). Argue that s induces a homeomorphism $S^2 \approx \hat{\mathbb{C}}$. Check that $s(z,t) = \dfrac{z}{1-t}$ and $s^{-1}(z) = \left(\dfrac{2z}{1+|z|^2}, \dfrac{-1+|z|^2}{1+|z|^2}\right)$.

(2) Recall why $\mathbb{C}\mathbf{P}^1 \approx \hat{\mathbb{C}}$ (affine chart) and derive from the previous question that $\mathbb{C}\mathbf{P}^1 \approx S^2$ via the map $[z_1 : z_2] \mapsto \left(\dfrac{2z_1\overline{z_2}}{|z_1|^2+|z_2|^2}, \dfrac{|z_1|^2-|z_2|^2}{|z_1|^2+|z_2|^2}\right)$.

(3) Let $V = \mathbb{C}^2$ and $S = S^3 \subseteq V$ the unit sphere, i.e. $S = \{(z_1, z_2) \in V : |z_1|^2 + |z_2|^2 = 1\}$. Recall why the inclusion $S \to V$ induces a homeomorphism $S/\mathrm{U}(1) \xrightarrow{\sim} \mathbb{C}\mathbf{P}^1$.

(4) Let $\mathbb{R}^3 \hookrightarrow V$ (i.e. $\mathbb{R}^3 \times \{0\} \hookrightarrow \mathbb{R}^4$). One would like to say: each $\mathbb{C}$-vector line in V intersects S^2 once, therefore $\mathbb{C}\mathbf{P}^1 \approx S^2$, hence $S^3/\mathrm{U}(1) \approx S^2$. Does this work?

(5) Define instead $p(z_1, z_2) = (2z_1\overline{z_2}, |z_1|^2 - |z_2|^2)$. Show that p defines a map $S^3 \to S^2$ that passes to the quotient as a homeomorphism $S^3/\mathrm{U}(1) \approx S^2$. Is this map related to (2)?

(6) Show that the preimage of $X \subseteq S^2$ by $p\colon S^3 \to S^2$ is:
   - $X = \{\text{one point}\}$: $p^{-1}(X)$ is a circle in $\mathbb{R}^4$ (contained in S^3).
   - $X =$ the equator: $p^{-1}(X)$ is a flat torus, i.e. the product of two circles $C_1 \times C_2$.
   - more generally, $X =$ any circle of latitude: $p^{-1}(X)$ is also a flat torus.

   Conclude that S^3 is the reunion of a family of flat tori, plus two circles.

*Note: The map p is a smooth fiber bundle of S^3 over S^2 with fiber S^1, denoted $S^1 \to S^3 \to S^2$ and called **Hopf fibration**. Some consider it one of the most beautiful mathematical constructions.*

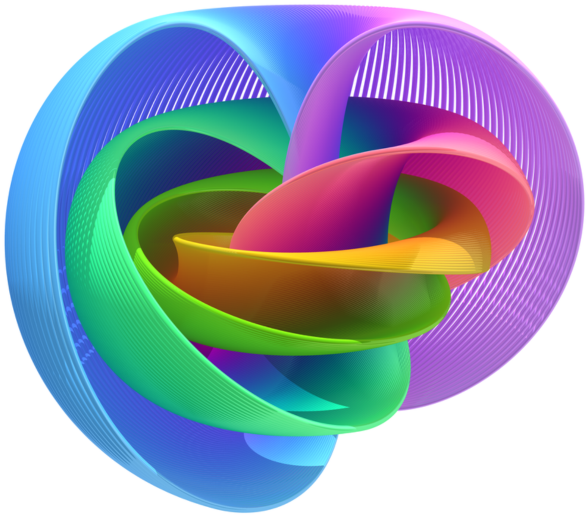


Figure 7.13: Hopf fibration. This picture shows the foliation of S^3 by flat tori under the stereographic projection $S^3 \xrightarrow{\sim} \widehat{\mathbb{R}^3}$. It is part of a beautiful animation created by Niles Johnson with `SageMath`: visit https://nilesjohnson.net/hopf.html.

Exercise 7.6. Homologies and elations

Let $\mathcal{P} = \mathbf{P}(V)$ be a projective space of dimension $\geqslant 2$. Let $f\colon \mathcal{P} \longrightarrow \mathcal{P}$ be a projective transformation and $F\colon V \longrightarrow V$ a linear lift of f.

(1) Show that f is a central projective transformation if and only if F admits an eigenspace of codimension 1.

(2) A central projective transformation f is called an ***elation*** if its center belongs to its axis and a ***homology*** otherwise. Show that f is a homology if and only if F is diagonalizable.

(3) Give the matrix of an elation in a suitable basis and explain how to find the center.

(4) What is the Jordan form of the matrix of an elation? Conclude that all elations are in the same conjugacy class. What about homologies?

Exercise 7.7. Central projective transformations (*)

Let $\mathcal{P} = \mathbf{P}(V)$ be a projective space of dimension $\geqslant 2$.

(1) Let $\mathcal{H} \subseteq \mathcal{P}$ be a hyperplane and let $O \in \mathcal{P}$ be a point not in $\mathcal{H}$. Show that for any points A, A' not in $\mathcal{H}$ such that O, A, A' are distinct and collinear, there exists a unique central projective transformation with center O that takes A to A'.

(2) Derive from the previous question that the central projective transformations with a given center and axis form a group isomorphic to $\mathbb{K}^\times$.

(3) Show that any perspectivity between hyperplanes of $\mathcal{P}$ extends to a central projective transformation of $\mathcal{P}$.

(4) Consider the configuration shown in Figure 7.2. Show that there exists a central projective transformation that takes A to A' etc. How unique is it? Conclude the equality of cross-ratios $[A, B, C, D] = [A', B', C', D']$.

Exercise 7.8. Pappus's theorem

Pappus's theorem, named after the Greek mathematician Pappus of Alexandria, is illustrated in Figure 7.14. It can be phrased as ([Cox89, p. 231]): *If the six vertices of a hexagon lie alternately on two lines, then the three points of intersection of pairs of opposite sides are collinear.*

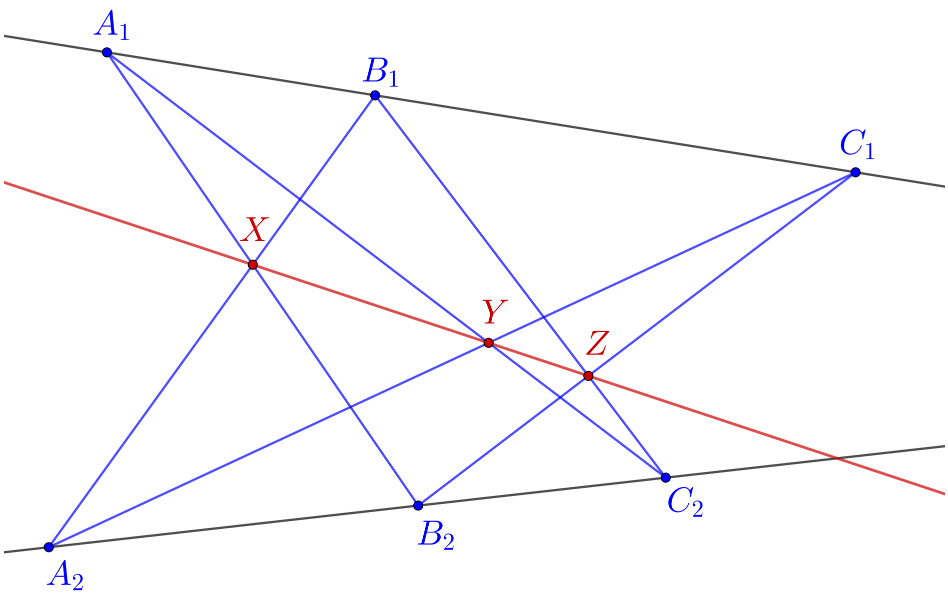


Figure 7.14: Pappus's theorem: the points X, Y, Z are collinear.

(1) Is this a theorem of affine or projective geometry?

(2) *Optional:* Show that Pappus's theorem is self-dual.

(3) Let us assume C_1, C_2, X are not collinear. Explain why we can assume that (C_1, C_2, X, A) is a projective frame. What are the homogeneous coordinates of C_1, C_2, X, and A?

(4) Show that one can write $B_1 = [p : 1 : 1]$, $Y = [1 : q : 1]$, $B_2 = [1 : 1 : r]$ with $p, q, r \in \mathbb{K}$. Considering the three lines intersecting at A_2, show that $rqp = 1$.

(5) Show that Z is the intersection of the three expected lines if and only if $rpq = 1$. Conclude. *Optional:* Show that Pappus's theorem holds if and only if $\mathbb{K}$ is commutative.

(6) *Optional:* Fix the proof if C_1, C_2, X are collinear.

Note: This proof is adapted from [Cox89, p. 236]. Pappus's original proof can be found in [Jon86] and is reproduced in [Wik21b]. The equivalence of Pappus's theorem with the commutativity of multiplication was proven by Hilbert in his Grundlagen *[Hil99]. I recommend [Ric11] for more perspective on Pappus's theorem (e.g., nine different proofs!).*

Exercise 7.9. Desargues's theorem

Desargues's theorem, named after a French mathematician, is illustrated in Figure 7.15. It says: *If two triangles are in perspective centrally, then they are in perspective axially.*

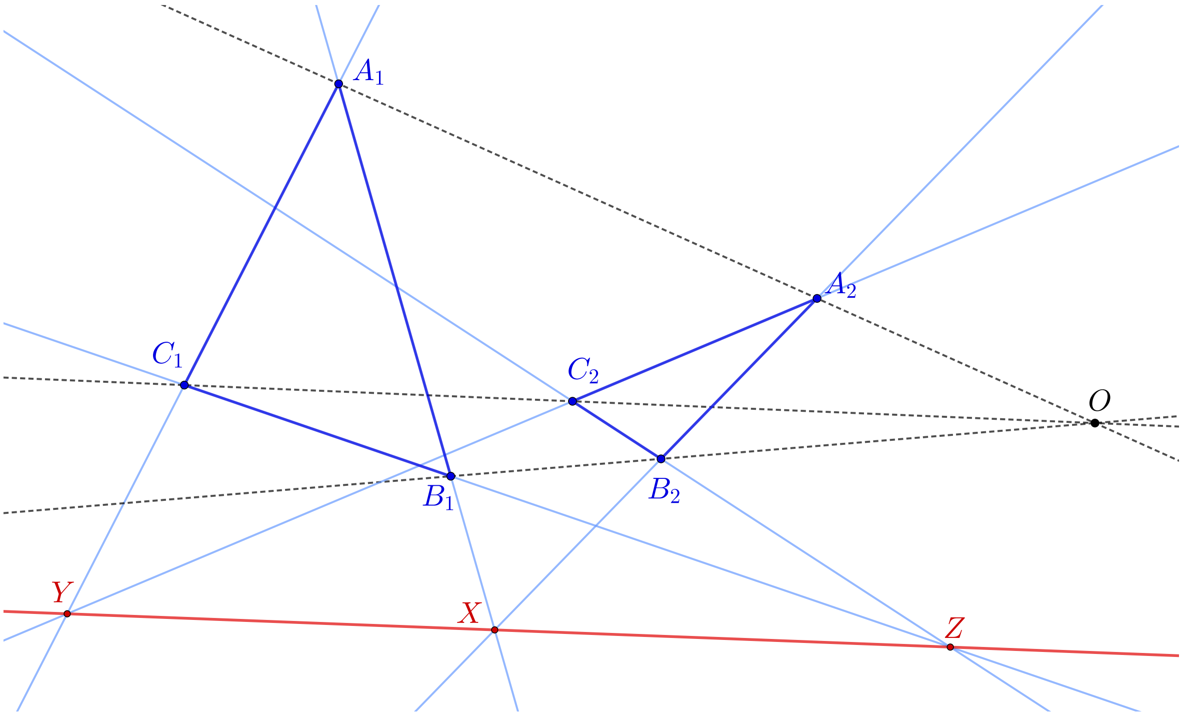


Figure 7.15: Desargues's theorem: the points X, Y, Z are collinear.

(1) We work in the projective plane $\mathcal{P} = \mathbf{P}(V)$ where V is a 3-dimensional vector space. Let us assume O, A_1, A_2 are distinct (why?). Recall why one can choose $O = [o]$, $A_1 = [a_1]$, $A_2 = [a_2]$ with $o = a_1 + a_2$. Write similarly $o = b_1 + b_2$ and $o = c_1 + c_2$.

(2) Show that $X = [x]$ where $x = a_1 - b_1 = b_2 - a_2$ and similar identities for Y and Z.

(3) Check that $x + y + z = 0$ and conclude.

(4) Show that the dual of Desargues's theorem is its converse. Does it also hold?

Note: This proof is adapted from [Bob19]. Desargues's theorem can also be derived from Pappus's: see [Cox89, p. 238]. It is a fundamental theorem of projective geometry because it characterizes projective spaces that are the projectivizations of vector spaces, see Remark 7.18.

Exercise 7.10. Formula for the cross-ratio

Let z_1, z_2, z_3, z_4 be four distinct points in $\hat{\mathbb{K}}$. Check that the map

$$f: z \mapsto \frac{(z - z_2)(z_1 - z_3)}{(z_1 - z)(z_3 - z_2)}$$

is a linear fractional transformation that maps z_1 to ∞, z_2 to 0, and z_3 to 1. Recover the formula for the cross-ratio.

Exercise 7.11. Cross-ratios and metrology

Consider the picture of Figure 7.16 (adapted from Wikipedia [Wik21a]). Denote by A, B, C, D, V the points in the 3-dimensional world, and by A', B', C', D', V' the points in the image. On the image, one can measure the lengths:

$$|A'B'| = 3\text{cm} \quad |B'C'| = 2\text{cm} \quad |C'D'| = 1\text{cm} \quad |D'V'| = 6\text{cm}$$

The goal is to determine the width $w = |BC|$ between the two buildings.

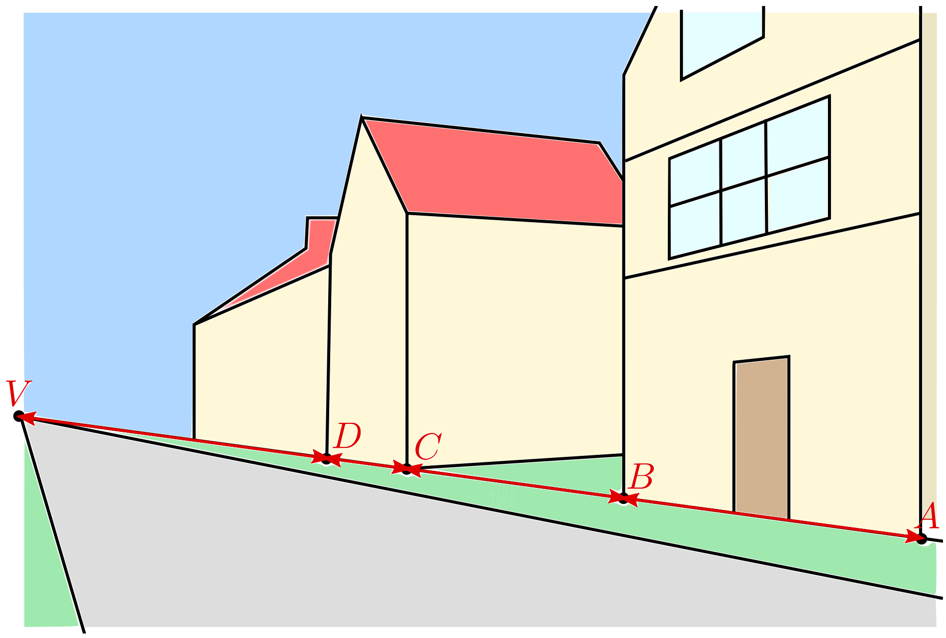


Figure 7.16: Application of cross-ratios to measure real-world dimensions.

(1) A naive attempt to find w consists in saying that lengths measured on the image are proportional to lengths measured in the real world. Does this work?

(2) Justify the equality of cross-ratios $[A, B, C, D] = [A', B', C', D']$. Given the widths of the adjacent houses $|AB| = 7\text{m}$ and $|CD| = 6\text{m}$, prove that $w = 8\text{m}$.

(3) Justify that $[A, B, C, V] = [A', B', C', V']$. Find w using only $|AB| = 7\text{m}$.

Exercise 7.12. Characterizing projective linear maps with cross-ratios

The goal of this exercise is to prove that a map $f : \mathcal{P} \to \mathcal{P}$ between projective spaces is projective linear if and only if it preserves cross-ratios.

(1) Recall why, if f is projective linear, then it preserves cross-ratios.
(2) Let us now prove the converse. First explain why the statement that "f preserves cross-ratios" implicitly assumes that f is a collineation.
(3) Prove the result when $\mathcal{P}$ is a projective line.
(4) Prove the result when $\dim \mathcal{P} > 1$. *Hint: use Theorem 7.63.*

Exercise 7.13. Classification of quadrics

Let V be a vector space. Denote $Q(V)$ the space of quadratic forms $q\colon V \to \mathbb{K}$.

(1) Show that $\mathrm{GL}(V)$ acts on $Q(V)$ by $f^*q \coloneqq q \circ f$. Can you describe this action in terms of matrices?
(2) Assume $\mathbb{K} = \mathbb{R}$. Show that two quadratic forms belong to the same $\mathrm{GL}(V)$-orbit if and only if they have same signature.
(3) Assume $\mathbb{K} = \mathbb{C}$. Show that two quadratic forms belong to the same $\mathrm{GL}(V)$-orbit if and only if they have same rank.
(4) Let $\mathcal{P} = \mathbf{P}(V)$ be a projective space. Show that $\mathrm{PGL}(V)$ has a natural action on the set of all quadrics in $\mathcal{P}$ and discuss the number of orbits when $\mathcal{K} = \mathbb{R}$ or $\mathbb{C}$.

Exercise 7.14. From a hyperboloid of two sheets to a sphere

Consider the hyperboloid $\mathcal{H}$ of two sheets with equation $x^2 + y^2 - z^2 = -1$ in $\mathbb{R}^3$.

(1) Show that by moving the plane at infinity $\partial_\infty\mathbb{R}^3$, the projective completion of $\hat{\mathcal{H}}$ can be seen as a sphere.
(2) Determine $\partial_\infty\mathcal{H}$ (the intersection of $\hat{\mathcal{H}}$ with the plane at infinity $\partial_\infty\mathbb{R}^3$). Can you describe why $\mathcal{H} \cup \partial_\infty\mathcal{H}$ is a topological sphere?

Exercise 7.15. Determinant quadric

Let $V = \mathcal{M}_{2\times 2}(\mathbb{K})$ denote the vector space of 2×2 matrices over a field $\mathbb{K}$.

(1) Show that the determinant function $\det\colon V \to \mathbb{K}$ is a quadratic form.
(2) Show that the set of non-invertible matrices defines a nondegenerate quadric in $\mathbf{P}(V)$. Find its normal form when $\mathbb{K} = \mathbb{R}$. *Optional: find an affine chart in which it is a hyperboloid of one sheet, and another where it is a hyperbolic paraboloid.*
(3) Show that $\mathrm{SL}(2, \mathbb{K})$ is an affine quadric in V. What is its projective completion? *Optional: when $\mathbb{K} = \mathbb{R}$, find an affine chart in which it is a hyperboloid of one sheet, and another where it is a hyperbolic paraboloid.*

Exercise 7.16. Dual conic

Let $\mathcal{P} = \mathbf{P}(V)$ be a real projective space of dimension 2 (or any other dimension). Let b be a nondegenerate symmetric bilinear form on V, i.e. a pseudo-inner product.

(1) Recall why b defines an isomorphism $\hat{b}\colon V \to V^*$ where $\hat{b}(v) := b(v, \cdot)$.

(2) For any subspace $W \subseteq V$, denote $W^{\perp_b} \subseteq V$ the b-orthogonal of W. Show that there is an induced map $\mathcal{Q} \mapsto \mathcal{Q}^{\perp_b}$ between subspaces of $\mathcal{P}$, which shares similar properties with projective duality. In fact, show that there is an isomorphism $\mathcal{Q}^{\perp_b} \xrightarrow{\sim} \mathcal{Q}^\circ$.

(3) Let $\mathcal{C} \subseteq \mathcal{P}$ denote the projective conic (or quadric) defined by b. Show that for any $p \in \mathcal{C}$, $p^{\perp_b}$ is a projective line (or hyperplane) that intersects $\mathcal{C}$ only at p. By definition, $p^{\perp_b}$ is the tangent line (or space) to $\mathcal{C}$ at p. *Optional: Show that in any affine chart A, $\mathcal{C}$ is a smooth submanifold of A and $p^{\perp_p}$ is the affine tangent space to $\mathcal{C}$ at p.*

(4) Show that b induces a pseudo-inner product b^* on V^*, hence a conic $\mathcal{C}^* \subseteq \mathcal{P}^*$. Show that under projective duality, points on $\mathcal{C}$ are mapped to tangent lines to $\mathcal{C}^*$.

(5) It is a famous theorem that given five points in general position on a projective plane, there exists a unique proper conic through all five. Being ***in general position*** means that no three are collinear. Using projective duality, show that given 5 lines in general position on a projective plane, there exists a unique proper conic tangent to all of them. (What does it mean for three or more lines to be in general position?)

(6) Look up *Pascal's hexagon theorem* and *Brianchon's hexagon theorem* and explain why they are dual theorems in the projective plane.

Exercise 7.17. Gaussian curvature of quadric surface (*)

Show that the sign of the Gaussian curvature of a surface is a projective invariant. Determine the sign of the Gaussian curvature of the quadric surfaces in normal form.

CHAPTER 8

The Klein model

Disclaimer: This chapter is a draft.

In this chapter we introduce and study the Klein model of hyperbolic space. This is a *projective* model: although it can simply be described as a disk (a ball in higher dimensions) with a special metric, it is best understood as a subset of the projective plane. In fact, the most natural definition of the Klein model makes it a special case of a *Cayley–Klein geometry*, which is a geometry that can be defined in the complement of a quadric in projective space. Remarkably, Euclidean geometry and elliptic geometry are also examples of Cayley–Klein geometries.

Historically, the Klein model was actually discovered by Eugenio Beltrami in 1868 ([Bel68a; Bel68b]), alongside what is now called the Poincaré models which we discuss in Chapter 10. While Beltrami described the model as a disk where chords are geodesics, Klein ([Kle71; Kle73]) showed its projective nature and gave the formula for the metric in terms of cross-ratios, inspired by work of Cayley [Cay59]. For a more detailed historical account, refer to [AP15].

8.1 Cayley–Klein geometries

In the complement of any quadric in projective space, one may define the Cayley–Klein "metric" using cross-ratios. Although we are mostly concerned with one case, namely the interior of an ellipsoid which will offer the Cayley–Klein model of hyperbolic space, it will be interesting to see that elliptic geometry can also be derived as a Cayley–Klein geometry, and even Euclidean geometry as a degenerate case. For a more extensive treatment, I recommend the paper [FS19]. Another good reference is the book [Ric11], which is very thorough.

8.1.1 The Cayley–Klein metric

Let $\mathcal{Q}$ be a quadric in a real projective space $\mathcal{P} = P(V)$ of dimension n. We denote $q\colon V \longrightarrow \mathbb{R}$ a quadratic form defining $\mathcal{Q}$, i.e. so that $\mathcal{Q}$ is the cone $\{q = 0\}$, and $b\colon V \times V \longrightarrow \mathbb{R}$ the associated symmetric bilinear form. In our setup, the quadric $\mathcal{Q}$ will be fixed. The following terminology is due to Cayley:

Definition 8.1. We shall call the quadric $\mathcal{Q} \subseteq \mathcal{P}$ the ***absolute***.

Example 8.2. When $\mathcal{Q}$ is of signature $(n, 1)$, it is called an ***ellipsoid***. By Sylvester's law of inertia, in suitable homogeneous coordinates $[X_1 : \ldots : X_{n+1}]$, $\mathcal{Q}$ is given by the equation

$$X_1^2 + \cdots + X_n^2 - X_{n+1}^2 = 0\,.$$

Note that $\mathcal{Q}$ does not intersect the hyperplane $X_{n+1} = 0$, therefore $\mathcal{Q}$ is contained in the affine chart $\mathcal{P} - \{X_{n+1} = 0\}$, and its equation in the inhomogeneous coordinates $x_k = \frac{X_k}{X_{n+1}}$ is:

$$x_1^2 + \cdots + x_n^2 - 1 = 0\,.$$

Thus we see in that $\mathcal{Q}$ is a sphere in such coordinates.

Typically, one could expect that $\mathcal{P} - \mathcal{Q}$ has two connected components determined by the sign of the quadratic form: $\Omega^+ := \{[x] \,:\, q(x) > 0\}$ and $\Omega^- := \{[x] \,:\, q(x) < 0\}$, e.g. the exterior and the interior of the ellipsoid.

Now let x, y be two points in $\mathcal{P} - \mathcal{Q}$, and consider the intersection of the line (xy) with the absolute $\mathcal{Q}$. In any affine chart, the points of intersection solve a polynomial equation of degree 2 in one variable, therefore there are three possibilities:

- There are two points of intersection I and J: the line (xy) is called ***hyperbolic***. For instance, this is the case when $\mathcal{Q}$ is an ellipsoid and $x, y \in \Omega^-$ are any two interior points as in Figure 8.1.
- There is one double point of intersection $I = J$: the line (xy) is called ***parabolic***.
- There are no points of intersection: the line (xy) is called ***elliptic***. In this case, one may still define I and J as complex points, that live in the complex projective space $\mathcal{P}^c := P(V \otimes \mathbb{C})$.

Remark 8.3. The case $(xy) \subseteq \mathcal{Q}$ is ruled out by the fact that $x, y \notin \mathcal{Q}$.

In all cases, one may take the cross-ratio:

$$c(x, y) := [x, y, J, I]\,.$$

This is a natural quantity to consider because it is a projective invariant, in particular it does not depend on the choice of coordinates on $\mathcal{P}$.

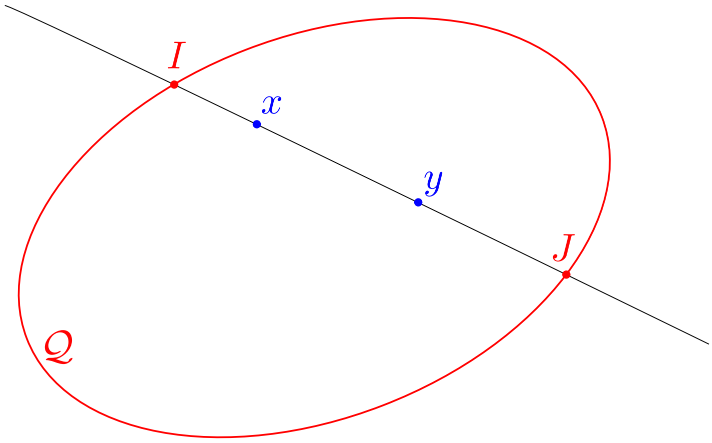


Figure 8.1: For any two x, y in the interior of the ellipsoid $\mathcal{Q}$, the projective line (xy) intersects $\mathcal{Q}$ in two distinct points I and J.

Proposition 8.4. *Let $l \subseteq \mathcal{P}$ be a projective line and consider the restriction c on $l - \mathcal{Q}$.*

- *If l is hyperbolic, then c is real-valued, and is positive on each component of $l - \mathcal{Q}$.*
- *If l is parabolic, then c is constant equal to* 1.
- *If l is elliptic, then c takes values in the unit circle* $\mathrm{U}(1) \subseteq \mathbb{C}$.

In all cases, for every $x, y, z \in l - \mathcal{Q}$:

$$c(x, y)c(y, z) = c(x, z)\,. \tag{8.1}$$

Proof. In the hyperbolic case and parabolic cases, c is real-valued by definition. Let us consider the hyperbolic case. The line l is a topological circle, therefore $l - \mathcal{Q} = l - \{I, J\}$ has two connected components. Choose any inhomogeneous coordinate on l, giving an identification $l \approx \hat{\mathbb{R}}$. Then the explicit formula for the cross-ratio (see Proposition 7.74) is:

$$c(x, y) = \frac{(J - x)(I - y)}{(J - y)(I - x)}\,. \tag{8.2}$$

We see with this formula that if x and y are in either component of $\hat{\mathbb{R}} - \{I, J\}$, then $c(x, y) > 0$.

For the parabolic case, since $I = J$, then $c(x, y) = [x, y, J, I] = 1$ by definition of the cross-ratio.

For the elliptic case, choose any inhomogeneous coordinate on l as before. We still have the formula (8.2), but now I and J are conjugate complex numbers: $J = \bar{I}$. Taking the modulus of (8.2) gives $c(x, y) = 1$.

Finally, the formula (8.1) is immediately checked using (8.2). ∎

In order to try and obtain a distance on (a connected component of) $\mathcal{P} - \mathcal{Q}$, it makes sense to take the logarithm of $c(x, y)$ in order to turn the multiplicative property (8.2) into an additive property.

Definition 8.5. The ***Cayley–Klein metric*** (or *Cayley–Klein pseudo-distance*) on $\mathcal{P} - \mathcal{Q}$ is the function defined by

$$d(x, y) := \frac{1}{2} |\ln c(x, y)|$$

where ln denotes the branch of the logarithm $\ln\colon \mathbb{C} - \{0\} \to \{z \in \mathbb{C}\colon \operatorname{Im}(z) \in (-\pi, \pi]\}$.

Indeed, taking the logarithm of equation Equation 8.1 and using the identity

$$\ln(ab) = \ln(a) + \ln(b) \tag{8.3}$$

when $a = c(x, y)$ and $b = c(y, z)$, we find that $\ln c(x, z) = \ln c(x, y) + \ln c(y, z)$. The triangle inequality for real numbers then yields

$$d(x, z) \leqslant d(x, y) + d(y, z)$$

Hence we essentially proved that d satisfies the triangle inequality in restriction to any line. However we have to be a little more careful, because the complex logarithm does not always verify the identity (8.3), in general it only holds up to a multiple of $i\pi$.

The following proposition is trivial to prove by definition of the Cayley–Klein metric, but is nevertheless important to note:

Proposition 8.6. *The Cayley–Klein metric is natural when restricting to projective subspaces: let $\mathcal{P}' \subseteq \mathcal{P}$ be a projective subspace, then the Cayley–Klein metric of $\mathcal{P}' - \mathcal{Q}'$ (where $\mathcal{Q}' := \mathcal{Q} \cap \mathcal{P}'$) is equal to the restriction of the Cayley–Klein metric of $\mathcal{P} - \mathcal{Q}$.*

Remark 8.7. Whenever $\mathcal{P}' = P(W) \subseteq \mathcal{P} = P(V)$ is a projective subspace, the restricted quadric $\mathcal{Q}' := \mathcal{Q} \cap \mathcal{P}'$ is a quadric in $\mathcal{P}'$: the associated quadratic form is simply the restriction of q to W.

8.1.2 Isometries and geodesics

Assume that the symmetric bilinear from b is nondegenerate. Recall that the subgroup of $\mathrm{GL}(V)$ that preserves b is denoted $\mathrm{O}(b)$ (or $\mathrm{O}(q)$). Clearly, $\mathrm{O}(b)$ preserves the quadric $\hat{\mathcal{Q}} := \{q = 0\}$ in V, and the decomposition of V into cones $V = \hat{\Omega}^+ \sqcup \hat{\mathcal{Q}} \sqcup \hat{\Omega}^-$ where $\hat{\Omega}^+ := \{v \in V : q(v) > 0\}$ and $\hat{\Omega}^- := \{v \in V : q(v) < 0\}$. Going to the quotient, we have that $\mathrm{PO}(b) \subseteq \mathrm{PGL}(V)$ preserves the quadric $\mathcal{Q}$ in $\mathcal{P} = P(V)$, and the decomposition of $\mathcal{P} = \Omega^+ \sqcup \mathcal{Q} \sqcup \Omega^-$. Since projective transformations preserve the cross-ratio, we clearly have:

Theorem 8.8. *The projective orthogonal group $\mathrm{PO}(b)$ acts on $\Omega^\pm$ by isometries with respect to the Cayley–Klein metric.*

Remark 8.9. The Cayley–Klein metric is not an genuine distance in general, but Theorem 8.8 still makes sense: it means that the action of $\mathrm{PO}(b)$ on $\Omega^\pm$ preserves d.

Remark 8.10. It is not too hard to show that conversely, any isometry of the Cayley–Klein metric coincides with the action of an element of $\mathrm{PO}(b)$, at least still assuming that b is nondegenerate. We will only prove it in the hyperbolic case i.e. signature $(n, 1)$ (see Theorem 8.36), relying on the analogous result for the hyperboloid (Theorem 5.7). Note that in the degenerate cases, one cannot hope that the statement is literally true, as shows the Euclidean case (signature $(1, 0)$) where the Cayley–Klein metric is identically zero.

Another fact that almost comes for free is that lines are geodesics for the Cayley–Klein metric, more precisely:

Definition 8.11. A ***chord*** in $\Omega^\pm$ is the intersection of a line in $\mathcal{P}$ with $\Omega^\pm$.

Theorem 8.12. *Chords are complete geodesics for the Cayley–Klein metric. More precisely:*

- *Hyperbolic chords are complete length-minimizing geodesics, in the sense that they can be parametrized as isometric curves $\gamma\colon \mathbb{R} \to \Omega^\pm$ (i.e. such that $d(\gamma(s), \gamma(t)) = |s - t|$ for all $s, t \in \mathbb{R}$).*
- *Elliptic chords are complete geodesics, in the sense that they can be parametrized as locally isometric curves (i.e. such that $d(\gamma(s), \gamma(t)) = |s-t|$ for all s, t sufficiently close). Moreover, they are closed geodesics.*
- *Parabolic lines are degenerate geodesics in a sense that can be made precise, in particular the distance between any two points on a parabolic line is zero.*

Remark 8.13. Note that any elliptic chord is equal to a whole projective line, so that it is a topological circle hence a closed geodesic (in particular it is not globally length-minimizing).

Proof. Let $c = l \cap \Omega^\pm$ be a chord. Assume that the line l is hyperbolic, so that $d(x, z) = d(x, y) + d(y, z)$ for any x, y, z that lie on c in this cyclic order. Let x_0 be any point on c and define $s(x) := \pm d(x_0, x)$ for any $x \in c$, where the sign is chosen so that $s(x) < 0$ when x is between I and x_0 and $s(x) > 0$ when x is between x_0 and J. It follows from the previous additive property that s is globally increasing along c, so that it gives a global coordinate on the chord. Moreover, one sees from (8.2) that $s(x) \longrightarrow \pm\infty$ when x approaches I or J, therefore $\gamma = s^{-1}$ is defined on $\mathbb{R}$. Finally, the fact that $d(\gamma(s), \gamma(t)) = |s - t|$ is again an immediate consequence of the additive property of the distance along c.

In the elliptic case, the additive property is only true if x, y, z are sufficiently close, but the proof is essentially the same.

In the parabolic case, it is clear that the distance between any two points on the line is zero. Let us leave the “sense that can be made precise” a mystery, but the example to have in mind is light-like geodesics in a pseudo-Riemannian manifolds. ∎

Remark 8.14. Again, it would be nice to prove that conversely, any geodesic for the Cayley–Klein metric is a projective line. We shall only do it in the hyperbolic case though (see § 8.2.4). As an exercise, the reader may prove the elliptic case (with the setup of § 8.1.5).

8.1.3 Cayley–Klein metrics in one dimension

Let us now examine the one-dimensional case more closely. Let $l = \mathcal{P} = P(V)$ be a projective line and let $Q \subseteq \mathcal{P}$ be a quadric as before, called the absolute, with associated quadratic form q and bilinear form b. As we have seen in § 8.1.1, Q consists of a pair of points I, J, possibly equal, possibly complex conjugate. Let us discuss these cases more precisely by looking at the signature of q.

Signature $(1, 1)$ *case.* (This is the case we are most interested in, which gives the Klein model.) By Sylvester's law of inertia, we can find coordinates (X_1, X_2) on V such that $q(X) = X_1^2 - X_2^2$. Therefore we see that $\mathcal{Q} = \{q = 0\}$ consists of two vector lines: $X_1 + X_2 = 0$ and $X_1 - X_2 = 0$, in other words Q consists of two points $I := [-1 \colon 1]$ and $J := [1 \colon 1]$. In the affine chart $\{X_2 \neq 0\}$ with coordinate $x = \frac{X_1}{X_2}$, this is $I = -1$ and $J = 1$. As expected, $\mathcal{P} - \mathcal{Q} \approx \hat{\mathbb{R}} - \{-1, 1\}$ consists of two connected components: $\Omega^- = \{|x| < 1\}$ and $\Omega^+ = \{|x| > 1\}$. Since the function c defined in (8.2) is positive on either connected components, the logarithm of c is the usual real logarithm, which satisfies (8.3). It follows that the Cayley–Klein metric (Definition 8.5) is a genuine distance on either connected components. Let us study it more precisely. The function c is given by

$$c(x, y) = \frac{(1 - x)(-1 - y)}{(1 - y)(-1 - x)}$$

so that

$$d(x, y) = \frac{1}{2} \left| \ln \frac{(1 + x)(1 - y)}{(1 - x)(1 + y)} \right| .$$

Let us consider the component Ω^-. There the factors $(1 + x)$, $(1 - y)$, $(1 - x)$, $(1 + y)$ are all positive, therefore

$$\begin{aligned} d(x, y) &= \left| \frac{1}{2} \ln \left(\frac{1 + x}{1 - x} \right) - \frac{1}{2} \ln \left(\frac{1 + y}{1 - y} \right) \right| \\ &= |\operatorname{artanh} x - \operatorname{artanh} y| \ . \end{aligned}$$

As we shall see in § 8.1.4, this expression makes it clear that Ω^- equipped with the Cayley–Klein metric is isometric to the hyperboloid model of hyperbolic space.

Remark 8.15. The component Ω^+ can be treated similarly. In fact, Ω^- and Ω^- are interchangeable: the fractional linear map $x \mapsto \frac{1}{x}$ is a projective transformation that exchanges the two. This symmetry is specific to dimension 1: the interior and exterior of higher-dimensional ellipsoids are not interchangeable: only the interior is convex.

Signature $(2, 0)$ *or* $(0, 2)$ *case.* Let us consider the $(2, 0)$ case; the $(0, 2)$ case is the same. Now in suitable coordinates we have $q(X) = X_1^2 + X_2^2$. The quadric Q is therefore empty, nevertheless we can define two imaginary points $I = [-i \colon 1]$ and $J = [i \colon 1]$, which correspond to the complex vector lines $X_1 + iX_2$ and $X_1 - iX_2$. In the affine chart $\{X_2 \neq 0\}$ with coordinate

$x = \frac{X_1}{X_2}$, this is $I = -i$ and $J = +i$, which live in $\mathcal{P}^c \approx \mathbb{C}P^1$ instead of $\mathcal{P} \approx \mathbb{R}P^1$. Now the function c is given by

$$c(x, y) = \frac{(i - x)(-i - y)}{(i - y)(-i - x)}$$

so that

$$d(x, y) = \frac{1}{2}\left|\ln \frac{(x + i)(y - i)}{(x - i)(y + i)}\right| .$$

We would like as before to expand this expression using the identity $\ln(ab) = \ln a + \ln b$, but we have to be a little careful: for arbitrary nonzero complex numbers a and b this is only true up to a multiple of $2i\pi$. With this in mind, we continue:

$$\begin{aligned} d(x, y) &= \frac{1}{2}\left|\ln\left(\frac{y - i}{y + i}\right) - \ln\left(\frac{x - i}{x + i}\right) - 2ik\pi\right| \\ &= \left|\frac{i}{2}\ln\left(\frac{y - i}{y + i}\right) - \frac{i}{2}\ln\left(\frac{x - i}{x + i}\right) + k\pi\right| \end{aligned}$$

that is

$$d(x, y) = |\operatorname{arccot} y - \operatorname{arccot} x + k\pi| \tag{8.4}$$

where k is the unique integer that makes $\operatorname{arccot} y - \operatorname{arccot} x + k\pi \in (-\frac{\pi}{2}, \frac{\pi}{2}]$, in other words (8.4) defines $d(x, y)$ uniquely as an element of $[0, \frac{\pi}{2}]$. As we shall see in § 8.1.5, this expression makes it clear that $\mathcal{P}$ equipped with the Cayley–Klein metric is isometric to the elliptic space $S^1/\{\pm 1\}$.

Signature $(1, 0)$ *or* $(0, 1)$ *case.* Let us consider the $(1, 0)$ case; the $(0, 1)$ case is the same. Note that this is a degenerate case: the bilinear form b is degenerate, so is the quadric Q (by definition). Now in suitable coordinates we have $q(X) = X_2^2$. The quadric $\mathcal{Q}$ is therefore reduced to the point $I = J = [1 : 0]$, which correspond to the vector lines $X_2 = 0$. In the affine chart $\{X_2 \neq 0\}$ with coordinate $x = \frac{X_1}{X_2}$, this is $I = J = \infty$. Now the function c is constant equal to 1, therefore $d(x, y) = 0$ for any x, y. Clearly d is not a distance on $\mathbb{R}$; nevertheless one may interpret this case as the Euclidean one. Indeed, we have seen in the previous chapter (see § 7.1.5) that the complement of a hyperplane (here a point) in a projective space is naturally an affine space, and it admits a natural (although not completely canonical) Euclidean structure.

8.1.4 Cayley–Klein model of hyperbolic space

This is the most important subsection of § 8.1 for us, since it gives the Klein model of hyperbolic space.

Let $\mathcal{P} = P(V)$ be an n-dimensional real projective space and let the absolute $\mathcal{Q} \subseteq \mathcal{P}$ be a quadric of signature $(n, 1)$. Such a quadric is called an ***ellipsoid***. As usual, we denote b and q the associated symmetric bilinear form and quadratic form.

Proposition 8.16. *The quadric $\mathcal{Q} \subseteq \mathcal{P}$ is a topological sphere of dimension $n-1$. $\mathcal{P} - \mathcal{Q}$ consists of two connected components: $\Omega^+ := \{[x] \,:\, q(x) > 0\}$ and $\Omega^- := \{[x] \,:\, q(x) < 0\}$. The component Ω^- is called the* ***interior*** *of the ellipsoid. It is a topological ball, it is convex, any line (xy) with $x, y \in \Omega^-$ is hyperbolic (it intersects $\mathcal{Q}$ in two distinct points).*

Let us introduce suitable coordinates to analyze the situation and prove Proposition 8.16 along the way. By Sylvester's law of inertia, in suitable homogeneous coordinates on V, the equation of $\mathcal{Q}$ is written

$$X_1^2 + \cdots + X_n^2 - X_{n+1}^2 = 0\,.$$

Note that $\mathcal{Q}$ does not intersect the hyperplane $X_{n+1} = 0$, therefore $\mathcal{Q}$ is contained in the affine chart $\mathcal{P} - \{X_{n+1} = 0\}$, and its equation in the inhomogeneous coordinates $x_k = \frac{X_k}{X_{n+1}}$ is:

$$x_1^2 + \cdots + x_n^2 - 1 = 0\,.$$

Thus we see that $\mathcal{Q}$ is an ellipsoid (it has the equation of a round sphere in the coordinates (x_i), but we do not have a Euclidean metric to distinguish between round spheres and other ellipsoids). The other claims of Proposition 8.16 are now immediate. In particular, the image of Ω^- in the affine chart $\mathcal{P} - \mathcal{Q} \to \{X_{n+1} = 1\} \approx \mathbb{R}^n$ is the unit ball, it is called the ***Beltrami–Klein disk*** (or ***Beltrami–Klein ball***).

Consider now the Cayley–Klein metric $d(x, y)$ on Ω^-. By the discussion carried out in in § 8.1.3, this is a genuine distance along any chord (intersection of Ω^- with a line) and it satisfies the additive property $d(x, y) + d(y, z) = d(x, z)$ whenever y is between x and z. It is tempting to say that d is a genuine distance on Ω^- and that the geodesics are the chords. Instead of proving it directly, we obtain this as a consequence of Theorem 8.18.

Recall from Chapter 5 that $\mathcal{H} \subseteq V$ denotes the hyperboloid

$$\mathcal{H} := \{v \in M \colon \langle v, v\rangle = -1\}\,.$$

and $\mathcal{H}^+$ is the upper sheet $\mathcal{H}^+ = \mathcal{H} \cap \{X_{n+1} > 0\}$. In Chapter 5, we saw that the induced metric on $\mathcal{H}^+$ from (V, b) is a Riemannian metric which makes $\mathcal{H}^+$ a model of hyperbolic space. Now, observe that there is an obvious way to identify Ω^- and $\mathcal{H}^+$, since Ω^- is the set of timelike lines in V, and each such line intersects $\mathcal{H}^+$ exactly once:

Definition 8.17. We denote $\psi \colon \mathcal{H}^+ \to \Omega^-$ the bijective map

$$\begin{aligned} \psi \colon \mathcal{H}^+ &\to \Omega^- \\ v &\mapsto [v]\,. \end{aligned}$$

The ***stereographic projection*** of the hyperboloid is the bijective map $\xi \colon \mathcal{H}^+ \to B$, where B is the unit ball in the affine hyperplane $\{X_{n+1} = 1\} \approx \mathbb{R}^n$, obtained by post-composing ψ with the affine chart $\varphi \colon \mathcal{P} - \mathcal{Q} \to \{X_{n+1} = 1\}$. Its image B is the Beltrami–Klein disk (see Figure 8.2).

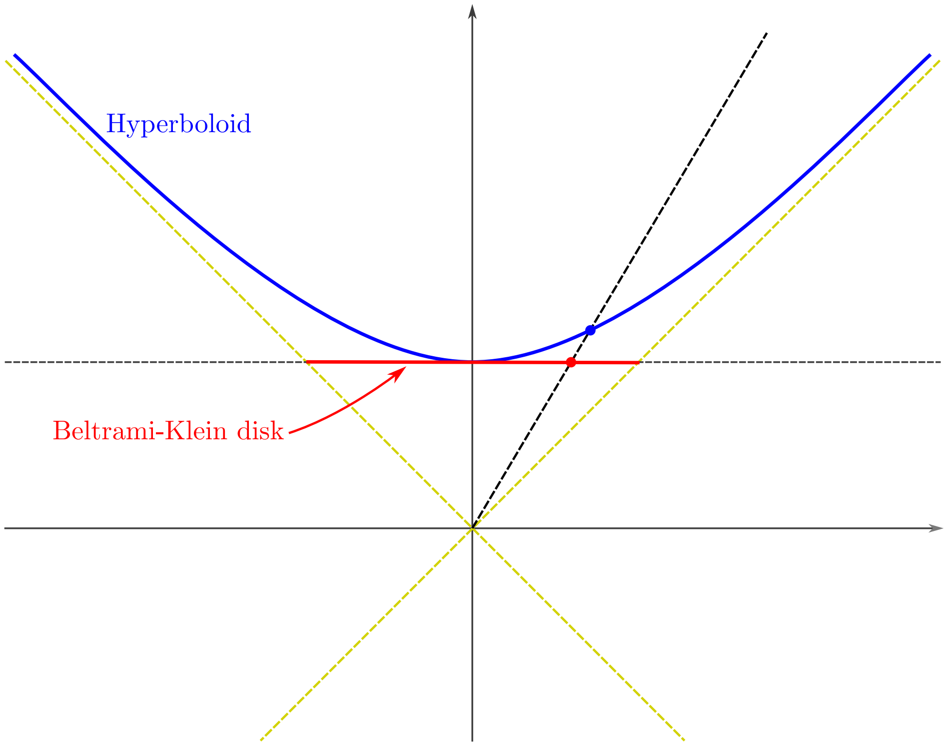


Figure 8.2: Stereographic projection of the hyperboloid to the Beltrami–Klein disk.

Theorem 8.18. *The map ψ is an isometry with respect to the hyperbolic distance $d_{\mathcal{H}}$ on $\mathcal{H}^+$ and the Cayley–Klein metric d_{CK} on Ω^-.*

Remark 8.19. Although we have not yet shown that the Cayley–Klein metric on Ω^- is a genuine distance, Theorem 8.18 means that $d_{\mathrm{CK}}(\psi(v), \psi(w)) = d_{\mathcal{H}}(v, w)$ for all $v, w \in \mathcal{H}^+$. (In addition, we know that ψ is bijective, so it deserves to be called a global isometry.) The fact that d_{CK} is a genuine distance is a corollary.

Proof. Let $v, w \in \mathcal{H}^+$, denote $x = \psi(v)$ and $y = \psi(w)$; we want to show that $d_{\mathrm{CK}}(x, y) = d_{\mathcal{H}}(v, w)$.

First we argue that it is enough to do the one-dimensional case, by simply restricting to the geodesic (vw) in $\mathcal{H}^+$, which is a one-dimensional hyperboloid. This amounts to intersecting $\mathcal{H}^+$ with a 2-dimensional vector subspace, therefore in $\mathcal{P}$ this corresponds to restricting to the projective line (xy). Indeed, on the one hand the Cayley–Klein metric is natural when restricting to projective subspaces (see Proposition 8.6); and on the other hand the hyperboloid is also natural when restricting to vector subspaces, (see Proposition 5.1): the restriction of the hyperbolic distance to a lower-dimensional hyperboloid is the hyperbolic distance on the lower-dimensional hyperboloid. (In Riemannian geometry, one says that lower-dimensional hyperboloids are ***totally geodesic***.)

We thus now assume that $\mathcal{P}$ is a projective line, and we can reinvest the work of § 8.1.3.

Choose coordinates such that the quadratic form is $q(X) = X_1^2 - X_2^2$, and denote $x = \frac{X_1}{X_2}$ the affine coordinate on $\mathcal{P} - \{X_2 = 0\}$. The quadric $\mathcal{Q}$ is the pair of points $I = -1$ and $J = 1$, the domain Ω^- is the interval $[-1, 1]$, and the Cayley–Klein metric is $d_{\mathrm{CK}}(x, y) = |\operatorname{artanh} x - \operatorname{artanh} y|$.

On the other hand, $\mathcal{H}^+$ is the upper arc of the hyperbola $X_1^2 - X_2^2 = -1$. It is parametrized by $\gamma(t) = (\sinh t, \cosh t)$, and in fact this is a unit geodesic (see Theorem 5.8). Let t_1 and t_2 be such that $v = \gamma(t_1)$ and $w = \gamma_{t_2}$. Since γ is a unit geodesic, we have $d_{\mathcal{H}}(v, w) = |t_1 - t_2|$. The points $x = \psi(v)$ is determined by $[x \colon 1] = [\sinh t_1 \colon \cosh t_1]$, so $x = \tanh t_1$. Similarly, $y = \tanh t_2$. Hence we find $d_{\mathrm{CK}}(x, y) = |\operatorname{artanh} x - \operatorname{artanh} y| = |t_1 - t_2| = d_{\mathcal{H}}(v, w)$ as desired. ∎

Corollary 8.20. *The Cayley–Klein metric on Ω^- may be written:*

$$d([u], [v]) = \operatorname{arcosh}\left(\frac{-b(u, v)}{\sqrt{q(u)q(v)}}\right) \tag{8.5}$$

Proof. The right-hand side of (8.5) is invariant by scaling u or v by positive numbers, therefore we may assume that $q(u) = q(v) = -1$. However in that case $\operatorname{arcosh}(-b(u, v))$ is the hyperbolic distance on $\mathcal{H}^+$ (see Theorem 5.12), so that (8.5) is precisely the statement of Theorem 8.18. ∎

As an immediate consequence of Theorem 8.18, we obtain:

Theorem 8.21. *The Cayley–Klein metric is a distance on Ω^-. It is induced by a complete Riemannian metric of constant sectional curvature -1.*

In other words, the previous theorem says that Ω^- equipped with the Cayley–Klein metric is a model of hyperbolic space. Being slightly pedantic, we call it the ***Cayley–Klein model*** or ***projective model***, to distinguish it from the ***Beltrami–Klein model*** which is the same model, except it is considered in an affine chart (where Ω^- becomes a disk).

Remark 8.22. Hilbert proposed a very elegant and elementary proof that the Cayley–Klein metric is a genuine distance that holds more generally on any proper convex set $\Omega \subseteq \mathcal{P}$. This generalization of the Cayley–Klein metric is called the ***Hilbert metric***. Hilbert's proof is reproduced by Papadopoulos in [Pap14, §5.6], and I sincerely encourage you to go and read it. The key ingredient is the invariance of the cross-ratio under perspectivities (central collineations).

8.1.5 Cayley–Klein model of elliptic space

Consider now the case where b is of signature $(n + 1, 0)$, in other words $(V, b) \approx \mathbb{R}^{n+1}$ is a Euclidean vector space. In this case, the quadric $\mathcal{Q}$ is empty, therefore the Cayley–Klein metric is defined on all $\mathcal{P}$. Notice that all lines are elliptic in this case.

Let $S = \{v \in V : q(v) = 1\}$ denote the unit sphere in (V, b). As in § 8.1.4, we would like to define the stereographic projection $\psi\colon S \to \mathcal{P}$, however note that each vector line in V intersects S twice, at two antipodal points $\pm v$. This is similar to the situation where each timelike line in Minkowski space intersects the hyperboloid $\mathcal{H}$ twice, except here there is no "upper sheet" of the sphere to resolve the issue. Instead, we have to define the stereographic projection as a map $\psi\colon S/\{\pm\,\mathrm{id}\} \to \mathcal{P}$, where $S/\{\pm\,\mathrm{id}\}$ is the set of pairs of antipodal points on the sphere.

Definition 8.23. We denote $\psi\colon S/\{\pm\,\mathrm{id}\} \to \mathcal{P}$ the bijective map

$$\begin{aligned} \psi\colon S/\{\pm\,\mathrm{id}\} &\to \mathcal{P} \\ \{\pm v\} &\mapsto [v]\,. \end{aligned}$$

Remark 8.24. In this setting, the ***stereographic projection*** of $S/\{\pm\,\mathrm{id}\}$ is the map $S/\{\pm\,\mathrm{id}\} \to \{X_{n+1} = 1\} \approx \mathbb{R}^n$, obtained by post-composing ψ with the affine chart $\mathcal{P} - \mathcal{Q} \to \{X_{n+1} = 1\}$. See Figure 8.3. Note that the stereographic projection is not defined on the $n-1$-dimensional sphere $S \cap \{X_{n+1} = 0\}$ (mod $\pm\,\mathrm{id}$), however it can be extended as a bijective map $S/\{\pm\,\mathrm{id}\} \to \{X_{n+1} = 1\} \cup \partial_\infty\{X_{n+1} = 1\} \approx \mathbb{R}P^n$.

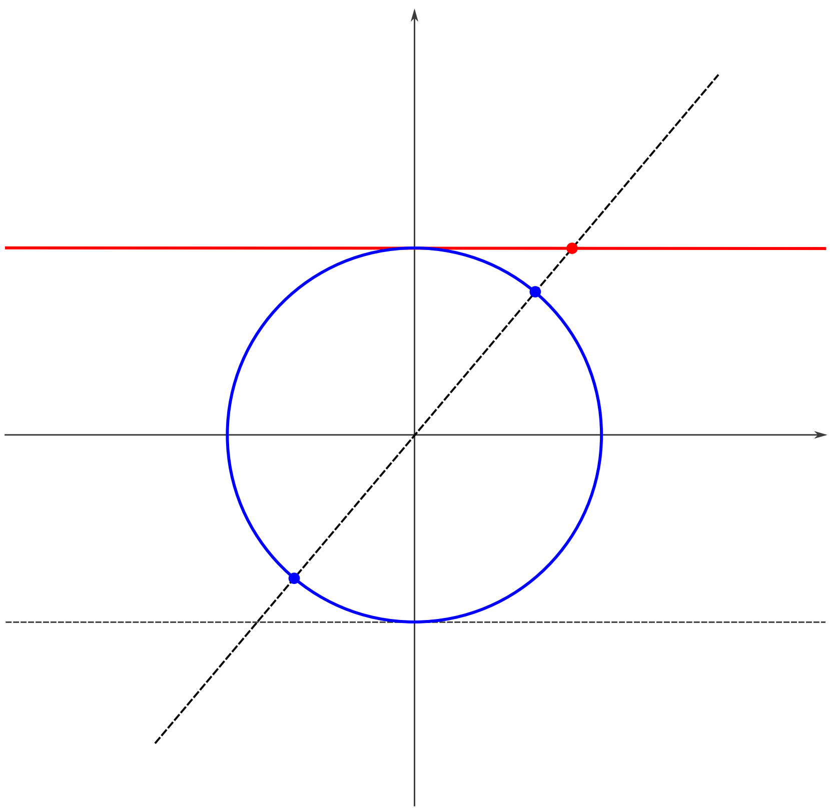


Figure 8.3: Stereographic projection of $S/\{\pm\,\mathrm{id}\}$.

Equip S with the Riemannian metric induced from the Euclidean metric on (V, b). As is well-known, this is a complete Riemannian metric of constant sectional curvature 1 (see Exercise 2.5). Since $v \mapsto -v$ is an isometry of S, the metric is well-defined on the quotient on $S/\pm 1$. We shall call the resulting Riemannian manifold $S/\{\pm \mathrm{id}\}$ the ***spherical model of elliptic space***.

Theorem 8.25. *The map ψ is an isometry with respect to the spherical distance d_S on $S/\{\pm \mathrm{id}\}$ and the Cayley–Klein metric d_{CK} on $\mathcal{P}$.*

Proof. The proof is completely analogous to that of Theorem 8.18; we leave it as an exercise to the reader (see Exercise 8.1). ∎

As an immediate consequence of the previous theorem, we obtain:

Theorem 8.26. *The Cayley–Klein metric is a distance on $\mathcal{P}$. It is induced by a complete Riemannian metric of constant sectional curvature* 1.

In other words, $\mathbb{R}P^n$ equipped with the Cayley–Klein metric is a model of elliptic space, which we naturally call the ***Cayley–Klein model*** or ***projective model***.

8.1.6 Cayley–Klein model of Euclidean space

Consider now the case where b is of signature $(1, 0)$, in particular it is degenerate. In suitable coordinates, the quadratic form is written

$$q(X) = X_{n+1}^2$$

therefore the degenerate quadric $\mathcal{Q}$ is the projective hyperplane $X_{n+1} = 0$. Note that in this case, all lines in $\mathcal{P} - \mathcal{Q}$ are parabolic, so the Cayley–Klein metric is constant equal to zero; it is not a Euclidean metric as one could hope. Nevertheless, we have already seen in the previous chapter (see § 7.1.5) that the complement $\mathcal{P} - \mathcal{Q}$ has a natural structure of an affine space. Like all affine spaces it admits a natural Euclidean structure, although it is not completely canonical: it depends on the choice of an inner product on the underlying vector space. In our case, this choice amounts to a Euclidean inner product b' on the kernel of b, so that $b + b'$ is a Euclidean inner product on V.

Let $\mathcal{E}^+ \subseteq V$ denote the affine hyperplane $X_{n+1} = 1$. Note that this is the upper sheet of the pseudosphere $\{q = 1\}$. The map ψ is now the bijective map $\psi\colon \mathcal{E}^+ \to \mathcal{P} - \mathcal{Q}$, given by $v \mapsto [v]$ as before. Note that it coincides with the inverse of the affine chart $\mathcal{P} - \mathcal{Q} \xrightarrow{\sim} \{X_{n+1} = 1\} \approx \mathbb{R}^n$, so the "stereographic projection" in this setting is the identity map $\mathcal{E}^+ \to \{X_{n+1} = 1\} \approx \mathbb{R}^n$. Equip $\mathcal{E}^+$ with the metric induced from $b + b'$. Then $\mathcal{E}^+$ is a complete Riemannian manifold of zero sectional curvature, in other words a model of Euclidean space. Using the stereographic projection ψ to transport the metric, we obtain that $\mathcal{P} - \mathcal{Q}$ is a model of Euclidean space, which we call the ***Cayley–Klein model*** or ***projective model***.

In summary, it is not the case that Euclidean geometry is obtained as a Cayley–Klein geometry, in the sense that the Euclidean metric is not a Cayley–Klein metric; nevertheless we can interpret the Cayley–Klein geometry associated to a degenerate quadric of signature $(1, 0)$ as a model of Euclidean geometry. In addition, in Exercise 8.2 it is shown that the Euclidean metric may be viewed as a degenerate elliptic metric.

8.1.7 Other Cayley–Klein geometries

Naturally, there are many more Cayley–Klein geometries, depending on the signature of quadric. Exploring these other examples is a fascinating program but beyond our scope, so we will be content with alluding to their existence.

Actually, as we saw in the Euclidean case, a Cayley–Klein geometry is not adequately defined by the Cayley–Klein metric in degenerate cases, nor even by the quadric alone. There are however more refined approaches to defining Cayley–Klein geometries. In [Ric11], a 2-dimensional Cayley–Klein geometry is defined by a "primal/dual" pair of conics, leading to seven types of Cayley–Klein geometries. A more general treatment is to define Cayley–Klein geometries as certain types of homogeneous spaces (spaces with a large group of symmetries), in the spirit of Klein's *Erlangen program*. This leads to nine 2-dimensional Cayley–Klein geometries: see e.g. [HOS00] (requires some knowledge of Lie theory!). I also recommend reading [FS19] for more geometric insights, especially on the Lorentzian geometries (Minkowski, de Sitter, anti de Sitter).

8.2 The Beltrami–Klein disk

8.2.1 Definition

Let us recap the setup of § 8.1.4. Let (V, b) be a Minkowski space of dimension $n + 1$. By choosing a suitable basis, we can identify $(V, b) \approx \mathbb{R}^{n,1}$. We denote $(X_1, \dots, X_{n+1})$ the associated coordinates. We denote $\mathcal{Q} \subseteq \mathbb{R}P^n$ the projectivized light cone, it is the projective quadric associated to b, called an ellipsoid. The open set $\Omega^- \subseteq \mathbb{R}P^n$ is the set of timelike vector lines, it is the interior of the ellipsoid, and a convex set in $\mathbb{R}P^n$. The Cayley–Klein metric $d_{\mathrm{CK}}$is a distance in Ω^-, making it a model of hyperbolic space. The image of this model under the usual affine chart $\varphi\colon P(V) - P(\{X_{n+1} = 0\}) \xrightarrow{\sim} \{X_{n+1} = 1\} \approx \mathbb{R}^n$ is called the *Beltrami–Klein disk*:

Definition 8.27. The ***Beltrami–Klein disk*** (or ***ball***) (B, d) is the unit ball $B \subseteq \mathbb{R}^n$ with the distance d that is the image of the Cayley–Klein model $(\Omega^-, d_{\mathrm{CK}}) \subseteq \mathbb{R}P^n$ under the affine chart φ.

We shall give an explicit expression of the distance d in the next subsection. As a corollary of Theorem 8.18, we obtain:

Theorem 8.28. *The Beltrami–Klein disk (B, d) is the isometric image of the hyperboloid model $(\mathcal{H}^+, d_{\mathcal{H}})$ under the stereographic projection $\xi \colon \mathcal{H}^+ \to B$.*

We will derive many features of the Beltrami–Klein disk from the hyperboloid using Theorem 8.28, especially the Riemannian metric. For now, we have:

Corollary 8.29. *The Beltrami–Klein disk is a model of hyperbolic space.*

8.2.2 Distance

By definition, the distance d in the Klein disk B is the image (the push-forward) of the Cayley–Klein metric d_{CK}, which is defined in terms of cross-ratios. Since the cross-ratio can be computed directly in any affine chart (due to its invariance under projective transformations), this distance can be defined directly in the Klein model:

Proposition 8.30. *Let $x, y \in B$. Denote by I and J the intersections of the straight line $l = (xy) \subseteq \mathbb{R}^n$ with ∂B, so that I, x, y, J are aligned in this order. Choose any affine frame on l, identifying it with $\mathbb{R}$. Then*

$$\begin{aligned} d(x, y) &= \frac{1}{2} \ln[x, y, J, I] \\ &= \frac{1}{2} \ln \frac{|Jx||Iy|}{|Jy||Ix|} \end{aligned}$$

where we denote $|AB|$ the Euclidean distance between two points $A, B \in \mathbb{R}^n$.

More explicitly, the distance can also be written:

Proposition 8.31.

$$d(x, y) = \operatorname{arcosh}\left(\frac{1 - \langle x, y \rangle}{\sqrt{(1 - \|x\|^2)(1 - \|y\|^2)}}\right) \tag{8.6}$$

where $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$ denote the Euclidean inner product and norm in $\mathbb{R}^n$.

Proof. Although Proposition 8.30 is an immediate application of Corollary 8.20 (also see Exercise 8.4), it is a good exercise to write the direct proof.

Let K be the Euclidean midpoint of I and J: see Figure 8.4. Let us choose the pair of points K and J as an affine chart on the line $l = (xy)$, giving an identification $(xy) \approx \mathbb{R}$: any point $m \in l$ is uniquely represented by a real coordinate λ such that $m = (1 - \lambda)K + \lambda J$. The coordinates of I, J, x, y are respectively:

$$\begin{aligned} \lambda_I &= -1 \\ \lambda_J &= 1 \\ \lambda_x &= \frac{\overline{Kx}}{\sqrt{1 - \|K\|^2}} \\ \lambda_y &= \frac{\overline{Ky}}{\sqrt{1 - \|K\|^2}} \end{aligned}$$

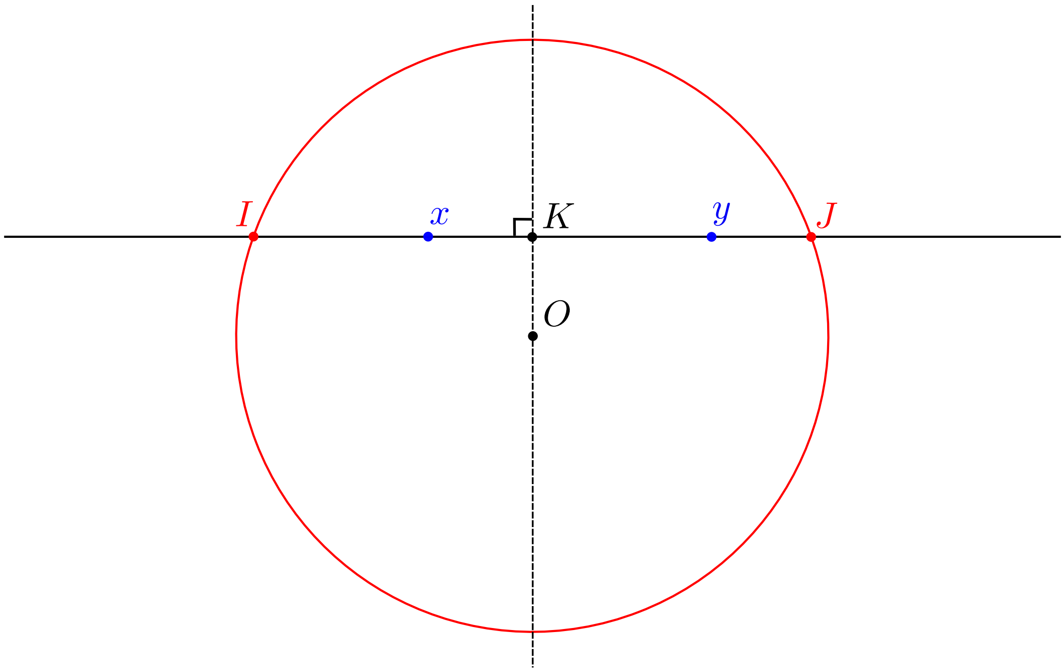


Figure 8.4: Calculating the distance $d(x, y)$ in the Beltrami–Klein disk.

where we denote $\overline{Kx}$ the signed distance between K and x, same for $\overline{Ky}$ (for instance, in Figure 8.4, $\overline{Kx} = -|Kx|$ and $\overline{Ky} = +|Ky|$). The expressions for λ_x and λ_y above can be found by noticing that $KI = \sqrt{1 - \|K\|^2}$ by the Pythagorean theorem.

Since the cross-ratio can be computed in any affine coordinates on a line, we may use these coordinates to compute the distance between x and y:

$$\begin{aligned} d(x, y) &= \frac{1}{2} |\ln[x, y, J, I]| \\ &= \frac{1}{2} \left| \ln \frac{(1 - \lambda_x)(-1 - \lambda_y)}{(1 - \lambda_y)(-1 - \lambda_x)} \right| . \end{aligned}$$

Let us manipulate this expression in view of obtaining (8.6):

$$\begin{aligned} d(x, y) &= \frac{1}{2} \left| \ln \frac{1 + \frac{\lambda_x - \lambda_y}{1 - \lambda_x \lambda_y}}{1 - \frac{\lambda_x - \lambda_y}{1 - \lambda_x \lambda_y}} \right| \\ &= \left| \operatorname{artanh} \frac{\lambda_x - \lambda_y}{1 - \lambda_x \lambda_y} \right| \\ &= \operatorname{arcosh} \frac{1 - \lambda_x \lambda_y}{\sqrt{(1 - \lambda_x^2)(1 - \lambda_y^2)}} . \end{aligned} \tag{8.7}$$

For the last equality, we used the identity $\operatorname{artanh} |t| = \operatorname{arcosh} \frac{1}{\sqrt{1 - t^2}}$ for $-1 < t < 1$.

To conclude, we compute:

$$\frac{1-\lambda_x\lambda_y}{\sqrt{(1-\lambda_x^2)(1-\lambda_y^2)}} = \frac{1-\|K\|^2-\overline{KxKy}}{(1-\|K\|^2-\overline{Kx}^2)(1-\|K\|^2-\overline{Ky}^2)} \tag{8.8}$$

By writing $x = K + (x-K)$ and $y = K + (y-K)$, we see that $\langle x,y\rangle = \|K\|^2 + \overline{KxKy}$, $\|x\|^2 = \|K\|^2 + \overline{Kx}^2$, and $\|y\|^2 = \|K\|^2 + \overline{Ky}^2$, so that (8.8) is rewritten

$$\frac{1-\lambda_x\lambda_y}{\sqrt{(1-\lambda_x^2)(1-\lambda_y^2)}} = \frac{1-\langle x,y\rangle}{(1-\|x\|^2)(1-\|y\|^2)}\,. \tag{8.9}$$

Inserting (8.9) into (8.7) yields the desired result. ■

8.2.3 Riemannian metric

The distance on the Beltrami–Klein disk is induced by a Riemannian metric, which can be computed as the pullback of the Riemannian metric on the hyperboloid under the inverse of the stereographic projection:

Proposition 8.32. *The Riemannian metric on the Beltrami–Klein disk $B \subseteq \mathbb{R}^n$ is given by*

$$\mathrm{d}s^2 = \frac{\mathrm{d}x_1^2+\cdots+\mathrm{d}x_n^2}{1-\|x\|^2} + \frac{(x_1\,\mathrm{d}x_1+\cdots+x_n\,\mathrm{d}x_n)^2}{(1-\|x\|^2)^2}$$

Proof. The inverse of the stereographic projection is the map

$$\begin{aligned}\xi^{-1}\colon B &\longrightarrow \mathcal{H}^+ \subseteq \mathbb{R}^{n,1}\\ x &\longmapsto \frac{\hat{x}}{\|\hat{x}\|}\end{aligned}$$

where we have denoted $\hat{x} = (x,1)$ and $\|\hat{x}\| = \sqrt{|q(\hat{x})|}$. In other words, this is:

$$\xi^{-1}\colon x \longmapsto \frac{(x,1)}{\sqrt{1-\|x\|^2}}$$

where $\|x\|$ now denotes the Euclidean norm of $x \in \mathbb{R}^n$. Recall that the metric on the hyperboloid is induced by the Minkowski metric

$$\mathrm{d}s^2 = \mathrm{d}X_1^2+\cdots+\mathrm{d}X_n^2-\mathrm{d}X_{n+1}^2\,.$$

The pullback metric on B under ξ^{-1} is obtained by replacing each $\mathrm{d}X_k$ by its expression of terms of the $\mathrm{d}x_k$'s. More precisely, the map ξ^{-1} is written

$$\begin{aligned}X_k &= \frac{x_k}{\sqrt{1-\|x\|^2}} \qquad \text{for } k\in\{1,\ldots,n\}\\ X_{n+1} &= \frac{1}{\sqrt{1-\|x\|^2}}\end{aligned}$$

therefore we find:

$$\begin{aligned} \mathrm{d}X_k &= \frac{\mathrm{d}x_k}{\sqrt{1-\|x\|^2}} + x_k(1-\|x\|^2)^{-3/2}\left(\sum_j x_j\,\mathrm{d}x_j\right) \qquad \text{for } k\in\{1,\dots,n\} \\ \mathrm{d}X_{n+1} &= (1-\|x\|^2)^{-3/2}\left(\sum_j x_j\,\mathrm{d}x_j\right) \end{aligned}$$

Taking the squares (symmetric product of one-forms):

$$\begin{aligned} \mathrm{d}X_k^2 &= \frac{\mathrm{d}x_k^2}{1-\|x\|^2} + \frac{x_k^2\left(\sum_j x_j\,\mathrm{d}x_j\right)^2}{(1-\|x\|^2)^3} + \frac{2x_k\,\mathrm{d}x_k\left(\sum_j x_j\,\mathrm{d}x_j\right)}{(1-\|x\|^2)^2} \qquad \text{for } k\in\{1,\dots,n\} \\ \mathrm{d}X_{n+1}^2 &= \frac{\left(\sum_j x_j\,\mathrm{d}x_j\right)^2}{(1-\|x\|^2)^3} \end{aligned}$$

Combining these, we find

$$\begin{aligned} \mathrm{d}s^2 &= \mathrm{d}X_1^2+\dots+\mathrm{d}X_n^2-\mathrm{d}X_{n+1}^2 \\ &= \frac{\sum_k \mathrm{d}x_k^2}{1-\|x\|^2} + \frac{\|x\|^2\left(\sum_j x_j\,\mathrm{d}x_j\right)^2}{(1-\|x\|^2)^3} + \frac{2\left(\sum_j x_j\,\mathrm{d}x_j\right)^2}{(1-\|x\|^2)^2} - \frac{\left(\sum_j x_j\,\mathrm{d}x_j\right)^2}{(1-\|x\|^2)^3} \\ &= \frac{\sum_k \mathrm{d}x_k^2}{1-\|x\|^2} + \frac{\left(\sum_j x_j\,\mathrm{d}x_j\right)^2}{(1-\|x\|^2)^2} \end{aligned}$$

as desired. ∎

Remark 8.33. We see from the expression that the Beltrami–Klein metric is not conformal to the Euclidean metric in B. More precisely, it is nowhere conformal except at the origin.

8.2.4 Geodesics

Since the stereographic projection $\mathcal{H}^+ \to B$ is a Riemannian isometry from the hyperboloid to the Beltrami–Klein disk, the (parametrized) geodesics on B are the images of the (parametrized) geodesics on $\mathcal{H}$. Ignoring the parametrization, recall that a geodesic on $\mathcal{H}$ is the intersection of $\mathcal{H}$ with a 2-plane in $\mathbb{R}^{n,1}$. In the projective model, this translates to the intersection of Ω^- with a projective line. In the Beltrami–Klein model, it thus translates to the intersection of B with a Euclidean straight line, in other words a chord.

Theorem 8.34. *The (unparametrized) geodesics in the Beltrami–Klein model are the chords, i.e. Euclidean straight line segments joining two points of ∂B.*

The curious reader will find an arclength parametrization of the geodesics (Exercise 8.8).

8.2.5 Isometries

We have seen in Theorem 8.8 that in the projective model, the projective orthogonal group $\mathrm{PO}(b)$ acts by isometries on Ω^- (and we promised to later prove that this is all the isometries). In the Beltrami–Klein disk, the action of $\mathrm{PO}(b)$ translates to the action of $\mathrm{PO}(n,1)$ on B by linear fractional transformations.

Example 8.35. Consider the Lorentz boost $M(t) \in \mathrm{SO}^+(2,1)$:

$$M(t) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cosh t & \sinh t \\ 0 & \sinh t & \cosh t \end{pmatrix}.$$

The projective linear action of $\pm M(t)$ on $\mathbb{R}P^2$ (preserving Ω^-) is given by

$$[X_1 : X_2 : X_3] \longmapsto [X_1 : (\cosh t)X_2 + (\sinh t)X_3 : (\sinh t)X_2 + (\cosh t)X_3]\,.$$

The linear fractional action of $\pm M(t)$ on $\mathbb{R}^2$ preserving the Beltrami–Klein disk $B \subseteq \mathbb{R}^2$ is given by:

$$(x_1, x_2) \longmapsto \left(\frac{x_1}{(\sinh t)x_2 + \cosh t}, \frac{(\cosh t)x_2 + (\sinh t)}{(\sinh t)x_2 + (\cosh t)} \right).$$

By the discussion above, we have:

Theorem 8.36. *The group of isometries of the Beltrami–Klein disk is* $\mathrm{PO}(n,1)$ *acting by linear fractional transformations. The subgroup of orientation-preserving isometries is* $\mathrm{PSO}(n,1)$.

Remark 8.37. Given any $f \in \mathrm{O}(n,1)$, exactly one element of the pair $\{f, -f\}$ is in $\mathrm{O}^+(n,1)$. It follows that there is an obvious isomorphism $\mathrm{PO}(n,1) \approx \mathrm{O}^+(n,1)$, and $\mathrm{PSO}(n,1) \approx \mathrm{SO}^+(n,1)$. It follows that $\mathrm{PO}(n,1)$ has two connected components, $\mathrm{PSO}(n,1)$ (orientation-preserving isometries) and the other one (orientation-reversing isometries).

The only nontrivial part of Theorem 8.36 that remains to prove is that any isometry of the Beltrami–Klein disk is given by the action of some element of $\mathrm{PO}(n,1)$. This follows from Theorem 5.6 and the following proposition:

Proposition 8.38. *The action of* $\mathrm{O}^+(n,1)$ *on* $\mathcal{H}^+$ *translates to the action of* $\mathrm{PO}(n,1)$ *on the Beltrami–Klein disk. More precisely, the stereographic projection* $\xi\colon \mathcal{H}^+ \longrightarrow B$ *conjugates the action of any* $f \in O^+(n,1)$ *on* $\mathcal{H}^+$ *(restricting the linear action on* $\mathbb{R}^{n,1}$*) to the projective linear (resp. fractional linear) action of* $\pm f$ *on* Ω^- *(resp. on* B*).*

The proof of Proposition 8.38 is essentially trivial and left to the reader: it is a matter of unraveling the definitions.

8.3 Exercises

Exercise 8.1.

Cayley–Klein model of elliptic space

Let (V, b) be a Euclidean vector space. We denote S the unit sphere in V.

(1) Prove Theorem 8.25: *The stereographic projection $S/\{\pm \mathrm{id}\} \longrightarrow P(V)$ is an isometry with respect to the spherical distance on $S/\{\pm \mathrm{id}\}$ and the Cayley–Klein metric on $P(V)$.*

(2) Show that the Cayley–Klein metric on $P(V)$ may be written:

$$d([u],[v]) = \arccos\left(\frac{b(u,v)}{\sqrt{b(u,u)b(v,v)}}\right).$$

Exercise 8.2.

Cayley–Klein model of Euclidean space

Let $\mathcal{P} = P(V)$ be a projective space of dimension n and let b be a symmetric bilinear form on V of signature $(1, 0)$. Let q denote the associated quadratic form and $\mathcal{Q} \subseteq \mathcal{P}$ the associated quadric.

(1) Let b_0 be a Euclidean inner product on $\ker b$. Show that $b_\varepsilon := \varepsilon^2 b_0 + b$ is a Euclidean inner product on V. Write the Cayley–Klein metric d_ε on $P(V)$ associated to b_ε using Exercise 8.1 (2). Derive the following expression in a suitable affine chart $\mathcal{P} - \mathcal{Q} \xrightarrow{\sim} \mathbb{R}^n$:

$$d_\varepsilon(x,y) = \arccos\left(\frac{1+\varepsilon^2\langle x,y\rangle}{\sqrt{(1+\varepsilon^2\|x\|^2)(1+\varepsilon^2\|y\|^2)}}\right).$$

(2) Show that, when $\varepsilon \longrightarrow 0$, the Cayley–Klein metric d_ε converges to the constant function $d_0 = 0$. Is this expected?

(3) Show that, when $\varepsilon \longrightarrow 0$, the "blown-up" Cayley–Klein metric $\frac{1}{\varepsilon} d_\varepsilon$ converges to a Euclidean metric on $\mathcal{P} - \mathcal{Q}$, which can be identified to b_0. Is this expected?

Exercise 8.3.

Hilbert metric

We have seen that the Cayley–Klein metric d is a distance in $\Omega \subseteq \mathbb{R}^n$ when Ω is the interior of an ellipsoid. Hilbert gave an elegant and elementary proof that applies more generally whenever Ω is a bounded convex open set. Your task is to go and read this proof in [Pap14, §5.6], and summarize it in a few lines.

Exercise 8.4.

Beltrami–Klein distance and stereographic projection

(1) Recall the expression of the hyperbolic distance $d_{\mathcal{H}}$ on the hyperboloid $\mathcal{H}^+ \subseteq \mathbb{R}^{n,1}$ and the distance d_{BK} on the Beltrami–Klein disk $B \subseteq \mathbb{R}^n$.

(2) Compute the image of the distance $d_{\mathcal{H}}$ on B under the stereographic projection. Recover that the stereographic projection is an isometry from the hyperboloid to the Beltrami–Klein disk.

Exercise 8.5.

Riemannian metric in the Beltrami–Klein disk

(1) Redo the calculation of the Riemannian metric in the Beltrami–Klein disk (preferably without looking at the lecture notes).

(2) Is the Beltrami–Klein metric conformal to the Euclidean metric in B?

Exercise 8.6.

Distance to origin

Check that the distance from the origin to a point x in the Beltrami–Klein disk $B \subseteq \mathbb{R}^n$ is given by $d(O, x) = \operatorname{artanh}(\|x\|)$, using three different arguments:

(1) Using the expression of the Cayley–Klein metric in terms of cross-ratios.

(2) Using the explicit expression of the distance (see Proposition 8.31).

(3) Using the Riemannian metric.

Exercise 8.7.

Circles in the Beltrami–Klein disk

A circle $C(x, R)$ in the 2-dimensional Beltrami–Klein disk (B, d) is the set of points at distance R from x. Show that any circle in the Beltrami–Klein disk is a Euclidean ellipse. Show an analogous result for higher-dimensional Beltrami–Klein disks.

Exercise 8.8.

Geodesics in the Beltrami–Klein disk

Find the expression of any parametrized geodesic in the Beltrami–Klein disk.

Exercise 8.9.

Isometries in the Beltrami–Klein disk

(1) Describe the action of $\mathrm{PO}(1,1)$ on the 1-dimensional Beltrami–Klein disk.

(2) Consider the matrix

$$R(t) = \begin{pmatrix} \cos t & -\sin t & 0 \\ \sin t & \cos t & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Show that $R(t) \in \mathrm{SO}(2,1)$ and describe its action on the 2-dimensional Beltrami–Klein disk.

(3) Show that any element of $\mathrm{PSO}(2,1)$ can be written $[L][R]$, for some Lorentz boost L and some $R = R(t)$. (We denote $[M]$ the element of PG associated to $M \in G$.) Recover the fact that $\mathrm{PSO}(2,1)$ is connected.

Part IV

Möbius geometry and the Poincaré models

Mathematics is the art of giving the same name to different things.

– Henri Poincaré[1]

[1] [Poi08]. It is amusing that Poincaré writes *la mathématique* as a singular noun in the original French text, contrary to the customary plural.

CHAPTER 9

Möbius transformations

Disclaimer: This chapter is a draft.

In this chapter, we review Möbius transformations, which can be either defined as conformal self-maps of S^n or $\widehat{\mathbb{R}^n}$, or as products of inversions through spheres. These are extremely important maps in hyperbolic geometry because they are the isometries of hyperbolic space in the Poincaré models, as we shall see in the next chapter. In a nutshell, the Poincaré models are conformal domains of $\mathbb{R}^n$, therefore their isometries will be conformal maps of $\mathbb{R}^n$, which are Möbius transformations. We shall also see that Möbius transformations are crucial in understanding the relations between different models of hyperbolic space. As it turns out, Möbius transformations play an even more special role in 2- and 3-dimensional hyperbolic geometry, where they are a key part of a striking connection between hyperbolic geometry and complex geometry. This small miracle is essentially due to the coincidence of Möbius transformations of the sphere S^2 with projective automorphisms of the complex projective line $\mathbb{C}P^1$.

Möbius transformations are named after the 19th century German mathematician August Ferdinand Möbius. He is best known for the discovery of the Möbius strip, but also made important contributions to projective geometry (e.g., he introduced homogeneous coordinates), where Möbius transformations play an important part.

I recommend [Bea95, Chap. 3, Chap. 4] as a complementary treatment of Möbius transformations: the coverage is slightly less extensive than these notes, but it contains more details and proofs.

9.1 Conformal maps

9.1.1 Similarities

Let $(V, \langle \cdot, \cdot \rangle)$ be a Euclidean vector space. We recall that a linear map $f : V \to V$ is called a ***similarity*** if it satisfies one of the equivalent conditions:

(i) f multiplies all distances by a constant factor. Equivalently, there exists $k > 0$ such that $\|f(x)\| = k\|x\|$ for all $x \in V$.

(ii) f can be written as the composition of a linear isometry (an element of $\mathrm{O}(V)$) and a homothety (an element of $\mathbb{R}^* \, \mathrm{id}_V$).

Linear similarities form a subgroup of $\mathrm{GL}(V)$, which one may sensibly denote $\mathbb{R}^* \, \mathrm{O}(V)$.

Remark 9.1. More generally, similarities refer to the affine version of the definition above: they are the maps $V \to V$ that multiply all distances by a constant factor, equivalently they are affine maps whose linear part is a linear similarity.

On the other hand, a linear map $f : V \to V$ is called ***conformal*** (or ***angle-preserving***) if it preserves unoriented angles between vectors:

$$\forall u, v \in V \quad \measuredangle(f(u), f(v)) = \measuredangle(u, v) \, .$$

Remark 9.2. What is an angle? One may define the unoriented angle between two nonzero vectors $u, v \in V$ as the real number $\theta = \measuredangle(u, v) \in [0, \pi]$ given by the formula

$$\langle u, v \rangle = \|u\| \, \|v\| \, \cos\theta \, .$$

Unless V is 2-dimensional, one cannot define a reasonable notion of oriented angles between vectors in V, so one may not define a notion of oriented angles-preserving map. It nevertheless makes sense to require that a map is angle-preserving and orientation-preserving.

Remark 9.3. By definition, if f is a linear conformal map, f must be injective: otherwise the angle $\measuredangle(f(u), f(v))$ would not always be well-defined. Thus f is an element of $\mathrm{GL}(V)$.

It turns out that linear similarities and linear conformal maps are the same:

Proposition 9.4. *A linear map $f : V \to V$ is conformal if and only if it is a similarity.*

Proof. Elementary and left to reader. ∎

9.1.2 Conformal maps of $\mathbb{R}^n$

Let $V = \mathbb{R}^n$, or more generally any Euclidean vector space, and let $\Omega \subseteq V$ be an open set. Let $f : \Omega \to W$ be a differentiable map, where W is another Euclidean space. For our purposes we will take $W = V$, but the reader should easily be able to generalize to any W. Let us assume that $\mathrm{d}f$ is always injective, in other words f is an immersion. (In our case where $V = W$, this amounts to saying that $\mathrm{d}f$ is always bijective, i.e. f is a local embedding.)

Consider two regular curves $\gamma_1 \colon I_1 \longrightarrow \Omega$ and $\gamma_2 \colon I_2 \longrightarrow \Omega$. By ***regular*** we mean that γ_i is differentiable and γ_i' does not vanish. Let $p \in \Omega$ be a point of intersection of the two curves: assume $p = \gamma_1(t_1) = \gamma_2(t_2)$. One can define the (unoriented) angle between γ_1 and γ_2 as the angle between their tangent vectors:

$$\measuredangle_p(\gamma_1, \gamma_2) := \measuredangle\left(\gamma_1'(t_1), \gamma_2'(t_2)\right) .$$

If f is an immersion, then the image curves $f \circ \gamma_1$ and $f \circ \gamma_2$ are regular curves in W that intersect at $f(p)$. One can again measure their angle of intersection. By definition, f is ***angle-preserving*** if, for any two regular curves γ_1 and γ_2 and for any point of intersection p,

$$\measuredangle_{f(p)}(f \circ \gamma_1, f \circ \gamma_2) = \measuredangle_p(\gamma_1, \gamma_2) .$$

(See Figure 9.1.)

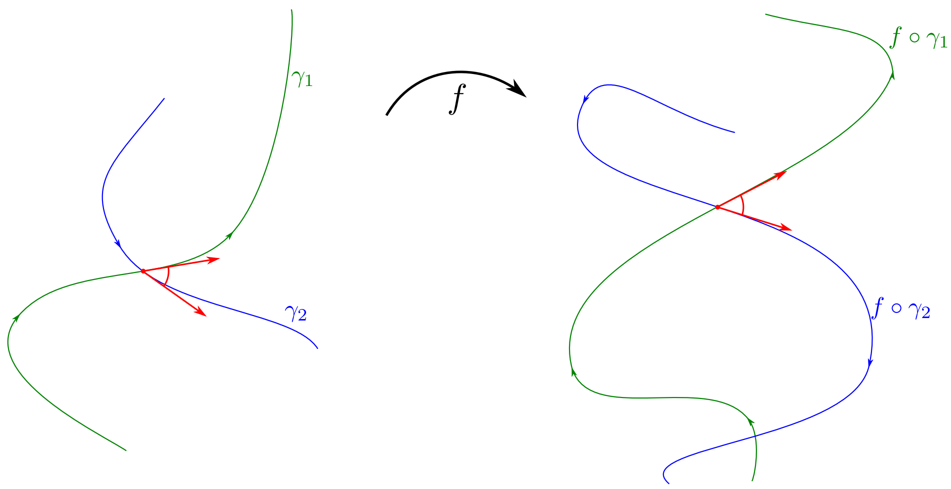


Figure 9.1: Angle-preserving map.

A synonym for *angle-preserving* is *conformal*:

Definition 9.5. A map $f \colon \Omega \subseteq V \longrightarrow W$ is called ***conformal*** if it is an angle-preserving immersion.

The next characterization is left to the reader as an exercise (Exercise 9.1):

Proposition 9.6. *Let $f \colon \Omega \subseteq V \longrightarrow W = V$. Then f is conformal if and only if f is differentiable and $\mathrm{d}f_x$ is a linear similarity for all $x \in \Omega$.*

In dimension 2, conformal maps are the same thing as locally injective holomorphic or anti-holomorphic maps:

Proposition 9.7. *Assume $V \approx \mathbb{C}$ be a 2-dimensional Euclidean vector space and $\Omega \subseteq V$ is an open connected subset. Then $f : \Omega \subseteq V \to V$ is conformal if and only if f is holomorphic or antiholomorphic and f' does not vanish.*

In higher dimensions, conformal maps are much more rigid, as shows the theorem of Liouville which will be given in § 9.2. Let us state a short version of this theorem here:

Theorem 9.8 (Liouville's conformal mapping theorem). *Let $f : \Omega \subseteq V \to V$ where $\dim V \geqslant 3$. Then f is conformal if and only if it is the restriction of a Möbius transformation of $\widehat{\mathbb{R}^n}$.*

9.1.3 Conformal maps of Riemannian manifolds

The definitions of § 9.1.2 swiftly generalize to the Riemannian setting. Let (M, g) be a Riemannian manifold. We recall that the Riemannian metric g is the data of an inner product $\langle \cdot, \cdot \rangle$ in each tangent space $\mathrm{T}_x M$. In particular, one can measure the angle of intersection of two regular curves just like in $\mathbb{R}^n$, by taking the angle between tangent vectors. Once again, a conformal map as defined as an angle-preserving map:

Definition 9.9. Let (M, g) and (N, h) be Riemannian manifolds. A ***conformal map*** $f : \Omega \subseteq M \to N$ is an angle-preserving immersion.

By definition, two Riemannian metrics g_1 and g_2 on M are called ***conformal*** (or ***conformally equivalent***) if there exists a positive function $\lambda : M \to \mathbb{R}_{>0}$ such that $g_1 = \lambda g_2$. This means that any point $x \in \mathrm{T}_x M$, the inner products g_1 and g_2 define the same angle between any two vectors in $\mathrm{T}_x M$. A ***conformal structure*** on M consists of a conformal class of metrics. The next proposition is elementary:

Proposition 9.10. *A differentiable map $f : \Omega \subseteq M \to N$ is conformal if and only if the pullback metric $f^* h$ is conformal to g.*

We leave the proof to Exercise 9.2 (elementary, yet a good exercise).

Definition 9.11. Let (M, g) be a Riemannian manifold. A ***conformal automorphism*** of M is a conformal diffeomorphism $f : M \to M$.

Remark 9.12. Definition 9.11 makes sense if M is only equipped with a conformal structure instead of a Riemannian metric.

Remark 9.13. By definition, any conformal map $M \to M$ is in particular a local diffeomorphism. If M is additionally compact and simply connected, then f must be a global diffeomorphism. Therefore, under these additional topological assumptions, the requirement that f is a diffeomorphism is superfluous in Definition 9.11. This is the case for instance when M is a topological sphere as in § 9.3.

9.2 Möbius transformations of $\widehat{\mathbb{R}^n}$

Let n be a positive integer. We denote $\widehat{\mathbb{R}^n} := \mathbb{R}^n \cup \{\infty\}$. Topologically, $\widehat{\mathbb{R}^n}$ is the one-point compactification of $\mathbb{R}^n$ and is homeomorphic to S^n via, for instance, the famous stereographic projection. We shall soon see that the stereographic projection is in fact a conformal equivalence between $\widehat{\mathbb{R}^n}$ and S^n.

Let us say that $S \subseteq \widehat{\mathbb{R}^n}$ is a ***(hyper)sphere*** if either $S \subseteq \mathbb{R}^n$ is a (hyper)sphere, or $S = P \cup \{\infty\}$ where $P \subseteq \mathbb{R}^n$ is an affine (hyper)plane.

Definition 9.14. $S \subseteq \widehat{\mathbb{R}^n}$ be a hypersphere. The ***inversion through*** S is the map $f : \widehat{\mathbb{R}^n} \longrightarrow \widehat{\mathbb{R}^n}$ defined as follows:

- If $S = S(a, r)$ is a hypersphere in $\mathbb{R}^n$, then f is defined on $\mathbb{R}^n - \{a\}$ by the property that $x' = f(x)$ if and only if x and x' lie on a same half-line starting at a, and the Euclidean distances ax and ax' are related by:
$$ax \cdot ax' = r^2 .$$
(See Figure 9.2.) f is continuously extended to $\widehat{\mathbb{R}^n}$ by $f(a) = \infty$ and $f(\infty) = a$.
- If $S = P \cup \{\infty\}$ where $P \subseteq \mathbb{R}^n$ is a hyperplane, then f is the Euclidean reflection through P on $\mathbb{R}^n$, extended to $\widehat{\mathbb{R}^n}$ by $f(\infty) = \infty$.

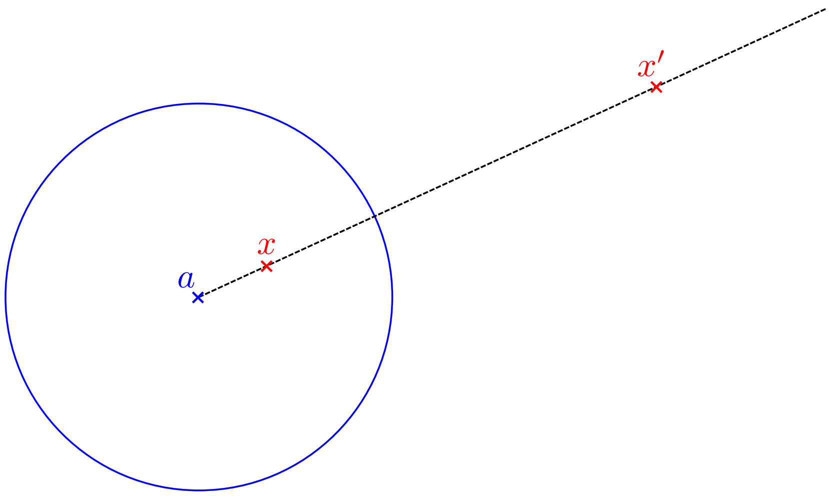


Figure 9.2: Inversion in a circle.

It is immediate to show from the definition that f is an involutive homeomorphism of $\widehat{\mathbb{R}^n}$, which fixes S and exchanges the two connected components of $\widehat{\mathbb{R}^n} - S$. It is also elementary to derive the analytic expression of the inversion in both cases (through a sphere or plane): see Exercise 9.4.

Definition 9.15. A map $f : \widehat{\mathbb{R}^n} \to \widehat{\mathbb{R}^n}$ is called a ***Möbius transformation*** if it can be written as a finite product of inversions.

We will denote $\text{Möb}(\widehat{\mathbb{R}^n})$ the group of Möbius transformations of $\widehat{\mathbb{R}^n}$ and $\text{Möb}^+(\widehat{\mathbb{R}^n})$ the subgroup of orientation-preserving elements. It is easy to see that it is an index 2 subgroup: there is a short exact sequence

$$1 \to \text{Möb}^+(\widehat{\mathbb{R}^n}) \to \text{Möb}(\widehat{\mathbb{R}^n}) \to \{\pm 1\} \to 1$$

where $\text{Möb}(\widehat{\mathbb{R}^n}) \to \{\pm 1\}$ is defined by assigning $+1$ [resp. -1] to an orientation-preserving (resp. reversing) Möbius transformation. Picking out any inversion $\tau \in \text{Möb}(\widehat{\mathbb{R}^n})$ yields a splitting of this short exact sequence via the isomorphism $\{\pm 1\} \xrightarrow{\sim} \{1, \tau\} \subseteq \text{Möb}(\widehat{\mathbb{R}^n})$.

Remark 9.16. It is quite common in the mathematics literature to impose that Möbius transformations are orientation-preserving, especially for $n = 2$. We do not make this restriction. We sometimes call $\text{Möb}(S^n)$ the ***full Möbius group*** and $\text{Möb}^+(S^n)$ the ***restricted Möbius group***.

Remark 9.17. We shall see in the next section that the Möbius group $\text{Möb}(\widehat{\mathbb{R}^n})$ is isomorphic to the Lie group $\text{PO}(n+1, 1)$, whence $\text{Möb}^+(\widehat{\mathbb{R}^n})$ is identified to $\text{PSO}(n+1, 1)$.

The central theorem of this section is:

Theorem 9.18. *Let $n \geqslant 2$ and let $f : \widehat{\mathbb{R}^n} \to \widehat{\mathbb{R}^n}$. The following are equivalent:*

(i) f is a Möbius transformation.

(ii) f preserves (unsigned) cross-ratios.

(iii) f is bijective and sphere-preserving, in the sense that it sends any sphere of lower dimension of $\widehat{\mathbb{R}^n}$ to a sphere.

(iv) f can be expressed as

$$f(x) = b + \frac{\alpha A(x-a)}{|x-a|^{\varepsilon}} \tag{9.1}$$

where $a, b \in \mathbb{R}^n$, $\alpha \in \mathbb{R}$, $A \in \text{O}(n)$, and $\varepsilon \in \{0, 2\}$.

(v) f is a conformal automorphism.

Remark 9.19. To make sense of (ii), we need to define cross-ratios in $\widehat{\mathbb{R}^n}$. Let a, b, c, d be four distinct points in $\mathbb{R}^n$. Let us define the (unsigned) cross-ratio as

$$[a, b, c, d] = \frac{|ac||bd|}{|bc||ad|}$$

where we take the Euclidean distances. This expression can be extended when one of the points is ∞ by ignoring the factors containing it. Note that, when a, b, c, d are collinear, their cross-ratio coincides up to sign with their cross-ratio as four points on a real projective line as defined in § 7.4.3.

Remark 9.20. To be precise in (v), we should say what it means for an immersion $f : \widehat{\mathbb{R}^n} \to \widehat{\mathbb{R}^n}$ to be conformal at the point ∞. One way to define it is to say that for some/any inversion g through a sphere $S(a, r) \subseteq \mathbb{R}^n$, the composition $f \circ g$ is conformal at a. Similarly, at a point x_0 where $f(x_0) = \infty$, one can make sense of f being conformal at x_0 by requiring that for some/any inversion g through a sphere $S(a, r) \subseteq \mathbb{R}^n$, the composition $g \circ f$ is conformal at x_0.

We shall not prove theorem Theorem 9.18, but let us give a few insights. The proof of (i) $\Leftrightarrow$ (ii) is surprisingly simple, see [Bea95, Theorem 3.2.7]. The fact that inversions satisfy (iii), (iv), and (v) can be checked by direct computation. Clearly, these properties are stable under finite composition. Proving that conversely, (iii) implies (i) requires some tricks, but it is not too difficult. The fact that (iv) implies (i) may be seen as a variation of the important theorem of linear algebra that any orthogonal transformation is a finite product of reflections. It remains to show that (v) implies one of the other characterizations, which is the truly hard part of the theorem. When $n = 2$, the result can be proven using complex analysis (see § 9.5 for the derivation of the Möbius group in that case). When $n \geqslant 3$, the result is a special case of Liouville's theorem below.

Theorem 9.21 (Liouville's conformal mapping theorem)**.** *Let $f : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}^n$ with $n \geqslant 3$. Then f is conformal if and only if f can be written as in* (9.1).

Proving Liouville's theorem essentially amounts to solving a PDE, a higher-dimensional version of the Cauchy–Riemann equations. As can be expected, this is a hard task. We shall not provide a proof, which is more or less difficult depending on the regularity assumption on f: a proof avoiding functional analysis can be written for f of class C^3, but the theorem is known to hold more generally for f in the Sobolev space $\mathrm{W}^{1,n}_{\mathrm{loc}}(\Omega, \mathbb{R}^n)$. We refer to [IM01] for a detailed account. Let us mention that, while $\mathrm{W}^{1,n}_{\mathrm{loc}}(\Omega, \mathbb{R}^n)$ does not include all differentiable functions, it is not hard to show that any conformal map is automatically $\mathrm{W}^{1,n}_{\mathrm{loc}}$: see [Dap]. I also recommend reading Danny Calegari's blog post [Cal13] for a sketch of proof with geometric insight.

9.3 Möbius transformations of S^n

9.3.1 Stereographic projection

There are several versions of the stereographic projection of a sphere to a plane. Let us consider the most standard one: given the unit sphere centered at the origin $S^n \subseteq \mathbb{R}^{n+1}$, we project the sphere S^n to the hyperplane $x_{n+1} = 0$ from the "North pole" $N = (0, \ldots, 0, 1)$. See Figure 9.3.

The stereographic projection is a homeomorphism $s : S^n - \{N\} \to \mathbb{R}^n$ that can be extended as a homeomorphism $S^n \to \widehat{\mathbb{R}^n}$ by setting $s(N) = \infty$. It is elementary to derive its analytic

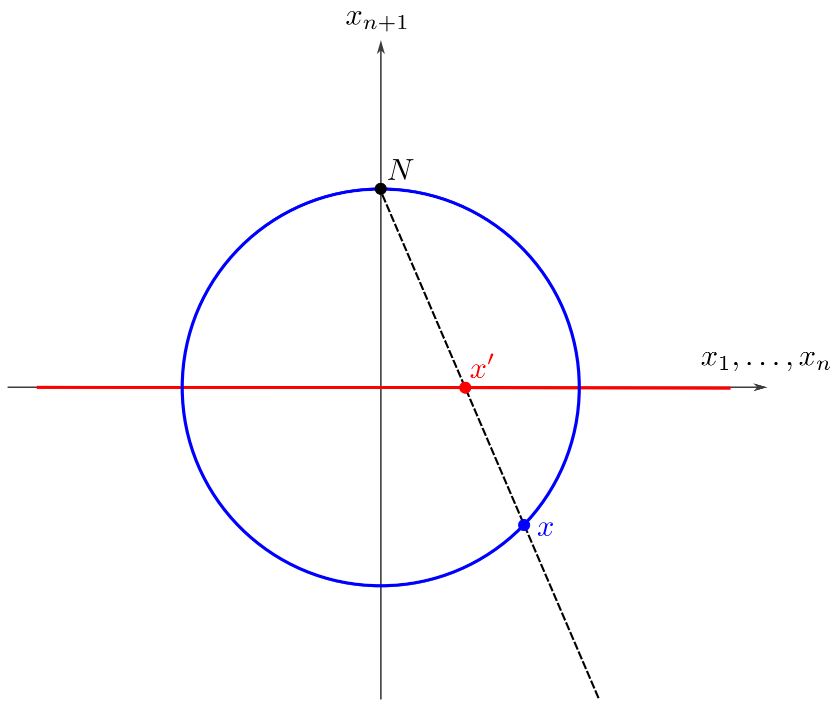


Figure 9.3: Standard stereographic projection $S^n - \{N\} \to \mathbb{R}^n$.

expression: write $x' - N = \lambda(x - N)$ where $x' = s(x)$. Examining the last component gives $0 - 1 = \lambda(x_{n+1} - 1)$, so $\lambda = \frac{1}{1-x_{n+1}}$. We thus find:

$$x'_k = \frac{x_k}{1 - x_{n+1}}$$

for all $k \in \{1, \ldots, n\}$. We easily recognize from this expression that the stereographic projection is the restriction to S^n of an inversion of $\widehat{\mathbb{R}^{n+1}}$ (figure out the details in Exercise 9.7):

Proposition 9.22. *The stereographic projection $s \colon S^n \to \widehat{\mathbb{R}^n}$ is the restriction to S^n of the inversion of $\widehat{\mathbb{R}^{n+1}}$ through the sphere $S(a, r)$ with $a = N$ and $r^2 = 2$.*

In particular, s is the restriction of a Möbius transformation, therefore it is conformal.

Corollary 9.23. *The stereographic projection $s \colon S^n \to \widehat{\mathbb{R}^n}$ is a conformal diffeomorphism.*

Remark 9.24. In Corollary 9.23, it is understood that the conformal structure of S^n is induced by $\mathbb{R}^{n+1}$. This coincides with the conformal structure on S^n underlying the spherical metric, since this metric is also induced by the Euclidean metric of $\mathbb{R}^{n+1}$.

9.3.2 Möbius transformations

Since the stereographic projection is the restriction of a Möbius transformation of $\widehat{\mathbb{R}^{n+1}}$, it is sphere-preserving: it sends spheres (of lower dimensions) in S^n to spheres. By definition, a map $S^n \to S^n$ is called an inversion if it is conjugate to an inversion of $\widehat{\mathbb{R}^n}$ by the stereographic projection. Thus inversions of S^n are conformal involutions and their fixed point sets are hyperspheres.

Let us define a Möbius transformation of S^n as a map $S^n \to S^n$ that can be written as a finite product of inversions. Using Theorem 9.18 we immediately obtain the characterization:

Theorem 9.25. *Let $n \geqslant 2$ and let $f : S^n \to S^n$. The following are equivalent:*

(i) f is a Möbius transformation (finite product of inversions).

(ii) f is conjugate to a Möbius transformation of $\widehat{\mathbb{R}^n}$ by the stereographic projection.

(iii) f is a sphere-preserving bijection.

(iv) f is a conformal automorphism.

Naturally, we denote $\text{Möb}(S^n)$ the group of Möbius transformations of S^n and $\text{Möb}^+(S^n)$ the index 2 subgroup of orientation-preserving elements. Clearly, the stereographic projection conjugates $\text{Möb}(S^n)$ and the $\text{Möb}(\mathbb{R}^n)$. In particular, they are isomorphic Lie groups.

9.3.3 Projective point of view

The projective point of view consists as seeing the sphere as a projective quadric. This will enable us to identify the Möbius group of S^n as the group of its projective automorphisms.

Consider Minkowski space $V = \mathbb{R}^{n+1,1}$. Recall that the projectivized light cone $P(\{\langle v, v\rangle = 0\})$ is a projective quadric $\mathcal{Q} \subseteq P(V)$ called an ellipsoid, whose equation in homogeneous coordinates is

$$X_1^2 + \cdots + X_{n+1}^2 - X_{n+2}^2 = 0\,.$$

In the affine chart $\varphi : P(\{X_{n+2} \neq 0\}) \xrightarrow{\sim} \{X_{n+2} = 1\} \approx \mathbb{R}^{n+1}$ with coordinates $x_k = \frac{X_k}{X_{n+2}}$, the equation of the ellipsoid is

$$x_1^2 + \cdots + x_{n+1}^2 - 1 = 0$$

so $\mathcal{Q}$ is identified to the unit sphere S^n in $\mathbb{R}^{n+1}$. Clearly, the Lorentz group $\mathrm{O}(n+1, 1)$ acts on V preserving the light cone, therefore the projective Lorentz group $\mathrm{PO}(n+1, 1)$ acts on $P(V)$ preserving the ellipsoid $\mathcal{Q}$.

Theorem 9.26. *Let $n \geqslant 2$. The identification $S^n \approx \mathcal{Q}$ given by the inverse of the affine chart φ yields isomorphisms*

$$\text{Möb}(S^n) \approx \mathrm{PO}(n+1, 1)$$
$$\text{Möb}^+(S^n) \approx \mathrm{PSO}(n+1, 1)\,.$$

Remark 9.27. Recall that we also have $\mathrm{PO}(n+1, 1) \approx \mathrm{O}^+(n+1, 1)$ and $\mathrm{PSO}(n, 1) \approx \mathrm{SO}^+(n+1, 1)$ (the latter is called the restricted Lorentz group). Since $\mathrm{SO}^+(n, 1)$ is connected (see § 4.2), it follows that $\text{Möb}^+(S^{n-1})$ is the identity component of $\text{Möb}(S^{n-1})$.

We do not give the detailed proof of Theorem 9.26, but here is the idea: both the Möbius transformations of S^{n-1} and its projective automorphisms can be characterized by the property that they are sphere-preserving, in the sense that they send spheres (of lower dimensions) to spheres. For the projective automorphisms, this characterization is a variation of Theorem 7.63. For Möbius transformations, it is part of Theorem 9.25. A variation of this proof using the cross-ratios preserving property might also be possible.

As a consequence of Theorem 9.26 and the discussion of § 9.3.2, we obtain:

Theorem 9.28. *We have isomorphisms:*

$$\begin{aligned} \mathrm{Möb}(\widehat{\mathbb{R}^n}) &\approx \mathrm{Möb}(S^n) \approx \mathrm{PO}(n+1,1) \approx \mathrm{O}^+(n+1,1) \\ \mathrm{Möb}^+(\widehat{\mathbb{R}^n}) &\approx \mathrm{Möb}^+(S^n) \approx \mathrm{PSO}(n+1,1) \approx \mathrm{SO}^+(n+1,1) \end{aligned}$$

Remark 9.29. In Theorem 9.28, we have isomorphisms of Lie groups: they are both homomorphisms of groups and diffeomorphisms of smooth finite-dimensional manifolds.

9.4 Möbius transformations of H^n and B^n

9.4.1 Möbius transformations of H^n

Consider the natural inclusion $\widehat{\mathbb{R}^{n-1}} \subseteq \widehat{\mathbb{R}^n}$ given by $(x_1, \dots, x_{n-1}) \mapsto (x_1, \dots, x_{n-1}, 0)$. Note that the complement $\widehat{\mathbb{R}^n} - \widehat{\mathbb{R}^{n-1}} = \mathbb{R}^n - \mathbb{R}^{n-1}$ consists of two half-spaces, which we denote H^n_+ and H^n_- according to the sign of the last coordinate.

Clearly, any Möbius transformation of $\widehat{\mathbb{R}^n}$ that preserves $\widehat{\mathbb{R}^{n-1}}$ restricts to a Möbius transformation of $\widehat{\mathbb{R}^{n-1}}$. Moreover, it must either preserve or exchange H^n and H^-, since these are the connected components of $\widehat{\mathbb{R}^n} - \widehat{\mathbb{R}^{n-1}}$. Conversely, we have:

Theorem 9.30. *Let $n \geqslant 2$. Any Möbius transformation of $\widehat{\mathbb{R}^{n-1}}$ uniquely extends to a Möbius transformation of $\widehat{\mathbb{R}^n}$ that preserves each of the two connected components of $\widehat{\mathbb{R}^n} - \widehat{\mathbb{R}^{n-1}}$.*

Remark 9.31. If one does not insist that each of the two components of $\widehat{\mathbb{R}^n} - \widehat{\mathbb{R}^{n-1}}$ are preserved, then there are two possible extensions, which differ by the inversion through the hyperplane $\widehat{\mathbb{R}^{n-1}} \subseteq \widehat{\mathbb{R}^n}$.

Using the conformal equivalence $S^n \approx \widehat{\mathbb{R}^n}$ given by the stereographic projection, we obtain the equivalent form of the previous theorem:

Theorem 9.32. *Let $n \geqslant 2$. Any Möbius transformation of $S^{n-1} \subseteq S^n$ uniquely extends to a Möbius transformation of S^n that preserves each of the two connected components of $S^n - S^{n-1}$.*

Proof. We use the projective point of view explained in § 9.3.3. We see S^{n-1} as a projective quadric in $\mathbb{R}^{n,1}$. Consider the inclusion $\mathbb{R}^{n,1} \to \mathbb{R}^{n+1,1}$ given by $(x_1, \dots, x_{n+1}) \mapsto$

$(0, x_1, \dots, x_{n+1})$. This induces an inclusion between the projective spaces, which restricts to an inclusion $S^n \longrightarrow S^{n+1}$. Up to a change of coordinates, this is the same as the inclusion in the statement of theorem.

It is easy to check that the "obvious" inclusion of $\mathrm{PO}(n, 1)$ in $\mathrm{PO}(n+1, 1)$ given the diagonal embedding

$$\begin{aligned} \mathrm{O}(n, 1) &\longrightarrow \mathrm{O}(n+1, 1) \\ M &\longmapsto \begin{bmatrix} 1 & 0 \\ 0 & \boxed{M} \end{bmatrix} \end{aligned}$$

provides an suitable extension of any $M \in \mathrm{O}(n, 1)$. Conversely, any suitable extension of M must be of the form

$$\hat{M} = \begin{bmatrix} x & 0 \\ v & \boxed{M} \end{bmatrix}$$

with $v \in \mathbb{R}^n$ and $x \in \mathbb{R}$, but the condition that $\hat{M} \in O(n+1, 1)$ enforces $v = 0$ and $x^2 = 1$ (we leave this computation as an easy exercise). Finally, the fact that $\hat{M}$ preserves each component of $S^n - S^{n-1}$ rules out $x = -1$. ■

Now let us examine the half-space $H^n := H^n_+$. The topological boundary of H^n in $\widehat{\mathbb{R}^n}$ is $\partial H^n = \widehat{\mathbb{R}^{n-1}}$. Observe that an inversion of $\widehat{\mathbb{R}^n}$ through a sphere S preserves H^n if and only if S is orthogonal to ∂H^n. Let us call ***Möbius transformation of*** H^n any map $f : H^n \longrightarrow H^n$ that can be written as a product of such inversions. We have the characterization:

Theorem 9.33. *Let $n \geqslant 2$ and $f : H^n \longrightarrow H^n$. The following are equivalent:*

(i) f is a Möbius transformation.

(ii) f is the restriction of a (unique) Möbius transformation of $\widehat{\mathbb{R}^n}$ that preserves H^n.

(iii) f is a conformal automorphism.

Proof. It is clear that (i) implies (ii) by definition. The converse is more tricky, we admit it. The fact that (ii) and (iii) are equivalent follows from Liouville's theorem when $n \geqslant 3$, and from direct analysis in the case $n = 2$ (see § 9.5). ■

Note that in particular, any Möbius transformation H^n extends continuously to the boundary $\partial H^n = \widehat{\mathbb{R}^{n-1}}$, and the boundary map is a Möbius transformation of $\widehat{\mathbb{R}^{n-1}}$. The uniqueness of this boundary map is merely due to continuity, and its existence to the previous theorem. Conversely, given a Möbius transformation of $\widehat{\mathbb{R}^{n-1}}$, Theorem 9.30 guarantees that it extends to a Möbius transformation of H^n. Let us record this:

Theorem 9.34. *Let $n \geqslant 2$. Any Möbius transformation of H^n extends continuously to the boundary $\partial H^n = \widehat{\mathbb{R}^{n-1}}$, and the boundary map is a Möbius transformation of $\widehat{\mathbb{R}^{n-1}}$. Conversely, any Möbius transformation $f \in \text{Möb}(\widehat{\mathbb{R}^{n-1}})$ is the boundary map of a unique Möbius transformation $\hat{f} \in \text{Möb}(H^n)$, called the **Poincaré extension** of f.*

Corollary 9.35. *Let $n \geqslant 2$. We have the isomorphisms:*

$$\begin{aligned} \text{Möb}(H^n) &\approx \text{Möb}(\widehat{\mathbb{R}^{n-1}}) \approx \text{PO}(n,1) \approx \text{O}^+(n,1) \\ \text{Möb}^+(H^n) &\approx \text{Möb}^+(\widehat{\mathbb{R}^{n-1}}) \approx \text{PSO}(n,1) \approx \text{SO}^+(n,1) \end{aligned}$$

9.4.2 Cayley transform and Möbius transformations of B^n

Let $n \geqslant 2$ and consider the open unit ball $B^n \subseteq \mathbb{R}^n$. Its topological boundary is $\partial B^n = S^{n-1}$. The story we told for H^n and ∂H^n can be repeated for B^n and ∂B^n. Indeed, the two are conformally equivalent via a Möbius transformation of $\widehat{\mathbb{R}^n}$.

Consider the stereographic projection $s\colon S^{n-1} \longrightarrow \widehat{\mathbb{R}^{n-1}}$ from the "South pole" P with coordinates $(0,\ldots,0,-1)$. Similarly to the stereographic projection from the North pole studied in § 9.3.1, s extends as an inversion of $\widehat{\mathbb{R}^n}$, namely the inversion through the sphere $S(a,r)$ with $a = P$ and $r^2 = 2$. We leave it as an easy exercise to the reader to argue that this inversion maps $\widehat{\mathbb{R}^{n-1}}$ to S^{n-1} and H^n to B^n, and conversely. However, this map is orientation-reversing, so instead let us consider the composition

$$c \coloneqq \tau \circ s$$

where τ is the inversion (reflection) through the hyperplane $\widehat{\mathbb{R}^{n-1}}$, which clearly preserves B^n. We thus have:

Theorem 9.36. *The map c is an orientation-preserving Möbius transformation of $\widehat{\mathbb{R}^n}$ that restricts to a conformal equivalence $H^n \longrightarrow B^n$, called the **Cayley transform**.*

It is straightforward to derive the expression of the Cayley transform:

$$x'_k = \frac{2x_k}{1+\|x\|^2+2x_n} \quad \text{for } k \in \{1,\ldots,n-1\} \qquad x'_n = \frac{\|x\|^2-1}{1+\|x\|^2+2x_n}$$

We can now use the Cayley transform to transport the situation of H^n to B^n, following § 9.4.1 step by step.

Theorem 9.37. *Let $n \geqslant 2$. Any Möbius transformation of S^{n-1} uniquely extends to a Möbius transformation of $\widehat{\mathbb{R}^n}$ that preserves each of the two connected components of $\widehat{\mathbb{R}^n} - S^{n-1}$.*

An inversion of $\widehat{\mathbb{R}^n}$ through a sphere S preserves B^n if and only if S is orthogonal to $\partial B^n = S^{n-1}$ (be careful: this does not amount to saying that the center of S lies on S^{n-1}). Let us call ***Möbius transformation of*** B^n any map $f\colon B^n \longrightarrow B^n$ that can be written as a product of such inversions. We have the characterization:

Theorem 9.38. *Let $n \geqslant 2$ and $f\colon B^n \longrightarrow B^n$. The following are equivalent:*

(i) f is a Möbius transformation of B^n.

(ii) f *is conjugate to a Möbius transformation of* H^n *by the Cayley transform.*

(iii) f *is the restriction of a (unique) Möbius transformation of* $\widehat{\mathbb{R}^n}$ *that preserves* B^n*.*

(iv) f *is a conformal automorphism.*

In particular, any Möbius transformation of B^n extends continuously to the boundary $\partial B^n = S^{n-1}$, and the boundary map is a Möbius transformation of S^{n-1}. Conversely, given a Möbius transformation of S^{n-1}, it uniquely extends to a Möbius transformation of B^n:

Theorem 9.39. *Let* $n \geqslant 2$*. Any Möbius transformation of* B^n *extends continuously to the boundary* $\partial B^n = S^{n-1}$*, and the boundary map is a Möbius transformation of* S^{n-1}*. Conversely, any Möbius transformation* $f \in \mathrm{Möb}(S^{n-1})$ *is the boundary map of a unique Möbius transformation* $\hat{f} \in \mathrm{Möb}(B^n)$*, called the* ***Poincaré extension*** *of* f*.*

Corollary 9.40. *Let* $n \geqslant 2$*. We have the isomorphisms:*

$$\begin{aligned} \mathrm{Möb}(B^n) \approx \mathrm{Möb}(S^{n-1}) \approx \mathrm{PO}(n,1) \approx \mathrm{O}^+(n,1) \\ \mathrm{Möb}^+(B^n) \approx \mathrm{Möb}^+(S^{n-1}) \approx \mathrm{PSO}(n,1) \approx \mathrm{SO}^+(n,1) \end{aligned}$$

9.5 Möbius transformations of $\hat{\mathbb{C}}$, $\mathbb{D}$, and $\mathbb{H}$

The 2-dimensional case is special, because the theorem of Liouville (Theorem 9.21) no longer holds for an arbitrary open set $\Omega \subseteq \mathbb{R}^2$. On the other hand, the possibility to use complex numbers and complex analysis opens new perspectives. By a fortunate coincidence, we will see that the conformal automorphisms of $\Omega \subseteq \widehat{\mathbb{R}^2} \approx \hat{\mathbb{C}}$ are indeed Möbius transformations when $\Omega = \widehat{\mathbb{R}^2}$, $\Omega = B^2$, and $\Omega = H^2$. As a result, the theory of Möbius transformations that we developed in the previous sections is still valid in the 2-dimensional case when working on these domains.

From now on, we identify $\mathbb{R}^2 = \mathbb{C}$ and $\widehat{\mathbb{R}^2} = \hat{\mathbb{C}}$, and we denote $\mathbb{D} = B^2$ the unit disk in $\mathbb{C}$ and $\mathbb{H} = H^2$ the upper half-plane.

9.5.1 Holomorphic and conformal maps on domains of $\mathbb{C}$

Let $f\colon \Omega \subseteq \mathbb{C} \to \mathbb{C}$ where Ω is an open set. We recall that f is called ***holomorphic*** if it is complex-differentiable at every $z_0 \in \Omega$. Equivalently, f is real-differentiable at z_0 and its derivative $\mathrm{d}f_{z_0}$ is $\mathbb{C}$-linear (this amounts to the so-called ***Cauchy–Riemann equations***). We also say that f is ***antiholomorphic*** if it is differentiable at every $z_0 \in \Omega$, and its derivative $\mathrm{d}f_{z_0}$ is $\mathbb{C}$-antilinear ($\mathrm{d}f(iv) = -i\,\mathrm{d}f(v)$). It is left as an easy exercise to the reader to show that f is antiholomorphic if and only if the complex conjugate $\bar{f}$ is holomorphic.

The relation between conformal maps and holomorphic maps in real dimension 2 is entirely explained by the following elementary lemma of linear algebra:

Lemma 9.41. *Let $L\colon \mathbb{R}^2 \to \mathbb{R}^2$ be a nonzero linear map. Identify $\mathbb{R}^2 \approx \mathbb{C}$.*

If L is orientation-preserving ($\det L > 0$), then

$$L \text{ is a similarity} \quad \Leftrightarrow \quad L \text{ is } \mathbb{C}\text{-linear} \quad \Leftrightarrow \quad L(z) = az \quad (a \in \mathbb{C}^*)$$

If L is orientation-reversing ($\det L < 0$), then

$$L \text{ is a similarity} \quad \Leftrightarrow \quad L \text{ is } \mathbb{C}\text{-antilinear} \quad \Leftrightarrow \quad L(z) = a\bar{z} \quad (a \in \mathbb{C}^*)$$

Proof. Elementary and left to the reader as an exercise. ∎

Corollary 9.42. $f\colon \Omega \subseteq \mathbb{C} \to \mathbb{C}$. *We have:*

$$f \text{ is conformal} \quad \Leftrightarrow \quad f \text{ is } \pm\text{-holomorphic and } f' \text{ does not vanish}\,.$$

*(We call ±-**holomorphic** a function that is holomorphic or antiholomorphic.)*

9.5.2 Möbius transformations of $\hat{\mathbb{C}}$

The extended complex line $\hat{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$ is called the ***Riemann sphere***, and it can be identified to the complex projective line $\mathbb{C}\mathbf{P}^1$ via the standard affine chart $[z_1 : z_2] \mapsto \frac{z_1}{z_2}$. Under this identification, the group of projective automorphisms of $\mathbb{C}P^1$, which is the projective linear group $\mathrm{PGL}(2, \mathbb{C})$, acts on $\hat{\mathbb{C}}$ by fractional linear transformations $z \mapsto \frac{az+b}{cz+d}$. For more details, see § 7.4.

Theorem 9.43. *A map $f\colon \hat{\mathbb{C}} \to \hat{\mathbb{C}}$ is an Möbius transformation if and only if it is fractional linear (orientation-preserving case) or its conjugate is fractional linear (orientation-reserving case).*

Proof. See Exercise 9.9. ∎

The fact that Theorem 9.18 and Theorem 9.25 hold when $n = 2$, even though Liouville's theorem does not, follows from the next theorem, whose proof relies on complex analysis.

Theorem 9.44. *Let $f\colon \hat{\mathbb{C}} \to \hat{\mathbb{C}}$. The following are equivalent:*

(i) f is an orientation-preserving Möbius transformation.

(ii) f is a fractional linear transformation.

(iii) f is orientation-preserving conformal automorphism.

(iv) f is a complex automorphism (a biholomorphism $\hat{\mathbb{C}} \to \hat{\mathbb{C}}$).

Remark 9.45. We have seen that a map $f\colon \Omega \subseteq \mathbb{C} \to \mathbb{C}$ is orientation-preserving conformal if and only if f is holomorphic and f' does not vanish. We can extend this property to maps $f\colon \Omega \subseteq \hat{\mathbb{C}} \to \hat{\mathbb{C}}$, but first we need to extend the notions of holomorphicity and conformality for such maps. We have already seen how to define conformality in Remark 9.20. One can similarly define holomorphicity by composing with the map $z \mapsto \frac{1}{z}$. More formally, $\hat{\mathbb{C}}$ (or $\mathbb{C}P^1$) can naturally be equipped with a structure of one-dimensional complex manifold (also known as Riemann surface), which is the right setting for holomorphicity.

Proof. The equivalence (i) $\Leftrightarrow$ (ii) immediately follows from Theorem 9.43. The equivalence (i) $\Leftrightarrow$ (iv) essentially follows from the compatibility between the extension of the notion of holomorphicity and conformality at ∞ (see Remark 9.45). The skeptical or meticulous reader may take this equivalence as a definition of complex automorphism. Let us now prove that (ii) $\Leftrightarrow$ (iii).

If f is fractional linear, then it is easy to check that it is orientation-preserving and conformal. Essentially, this boils down to the fact that holomorphic maps (with non-vanishing derivative) are orientation-preserving and conformal. Even quicker, we can use (i): since f is a Möbius transformation, it is conformal.

Conversely, assume that f is an orientation-preserving conformal automorphism. Since fractional linear transformations act transitively on $\hat{\mathbb{C}}$, we may assume that $f(\infty) = \infty$ by precomposing f with a fractional linear transformation (for instance, take $z \mapsto \frac{wz+1}{z+w}$ where $w = f^{-1}(\infty)$). Thus f restricts to an entire function $\mathbb{C} \to \mathbb{C}$, moreover $\lim |f(z)| = +\infty$ when $|z| \to +\infty$. It is a classical exercise of complex analysis that this forces f to be polynomial. Indeed, the function $g\colon z \mapsto f(\frac{1}{z})$ has a pole at $z = 0$ (since $|g(z)| \to +\infty$ when $z \to 0$), therefore the Laurent series of g has finitely many nonzero coefficients of negative degree, i.e. the power series of f has finitely many nonzero coefficients. Since f is bijective it must have exactly one zero, therefore it has degree 1 by the fundamental theorem of algebra. Thus $f(z) = az + b$ for some $a \in \mathbb{C}^*$ and $b \in \mathbb{C}$, in particular f is fractional linear. ■

Corollary 9.46. *The natural identifications $S^2 \approx \hat{\mathbb{C}} \approx \mathbb{C}P^1$ induce isomorphisms:*

$$\text{Möb}^+(S^2) \approx \text{Aut}(\hat{\mathbb{C}}) \approx \text{Aut}(\mathbb{C}P^1)$$

where $\text{Aut}(\hat{\mathbb{C}})$ is the group of complex (i.e. conformal orientation-preserving) automorphisms of the Riemann sphere $\hat{\mathbb{C}}$ and $\text{Aut}(\mathbb{C}P^1)$ is the group of projective transformations of $\mathbb{C}P^1$.

Recall that we also have isomorphisms $\text{Möb}^+(S^2) \approx \text{PSO}(3,1)$ and $\text{Aut}(\mathbb{C}P^1) \approx \text{PGL}(2,\mathbb{C})$ (acting projective linearly on $\mathbb{C}P^1$ or fractional linearly on $\hat{\mathbb{C}}$), therefore we obtain the "accidental" isomorphism of Lie groups:

$$\text{PSO}(3,1) \approx \text{PGL}(2,\mathbb{C})\,.$$

Remark 9.47. Let us mention that there is also an accidental isomorphism of complex Lie groups $\text{PGL}(2,\mathbb{C}) \approx \text{SO}(3,\mathbb{C})$.

9.5.3 Möbius transformations of $\mathbb{D}$

Which subgroup of $\mathrm{PGL}(2,\mathbb{C})$ leaves the unit disk $\mathbb{D}$ invariant when acting on the Riemann sphere $\hat{\mathbb{C}}$? To answer this question, it is useful to work in $\mathbb{C}\mathbf{P}^1$. In homogeneous coordinates, the disk $\mathbb{D}$ can be written: $\mathbb{D} = \{[z_1 : z_2] \mid |z_1|^2 - |z_2|^2 < 0\}$. Indeed, this is clearly equivalent to $|z|^2 < 1$ where $z = \frac{z_1}{z_2}$. Consider the Hermitian symmetric form on $\mathbb{C}^2$:

$$\begin{aligned} h\colon \mathbb{C}^2 \times \mathbb{C}^2 &\to \mathbb{C} \\ ((z_1, z_2), (z_1', z_2')) &\mapsto z_1\overline{z_1'} - z_2\overline{z_2'} \end{aligned}$$

and denote $q(z_1, z_2) = h((z_1, z_2), (z_1, z_2)) = |z_1|^2 - |z_2|^2$ the associated quadratic form. The signature of h as a Hermitian symmetric form is $(1,1)$, in fact its matrix in the canonical basis of $\mathbb{C}^2$ is

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

The subgroup of $\mathrm{GL}(2,\mathbb{C})$ leaving h invariant is denoted $\mathrm{U}(h)$ or simply $\mathrm{U}(1,1)$. Let us also introduce the group $\mathrm{SU}(1,1)$ of elements of $\mathrm{U}(1,1)$ with determinant 1. In terms of matrices (see Exercise 9.10):

$$\begin{aligned} \mathrm{U}(1,1) &= \{M \in \mathcal{M}_{2\times 2}(\mathbb{C}) \mid M^{\mathsf{T}} H \bar{M} = H\} \\ &= \{uA \mid |u| = 1, A \in \mathrm{SU}(1,1)\} \end{aligned}$$

$$\begin{aligned} \mathrm{SU}(1,1) &= \{M \in \mathrm{SL}(2,\mathbb{C}) \mid M^{\mathsf{T}} H \bar{M} = H\} \\ &= \left\{ \begin{bmatrix} a & b \\ \bar{b} & \bar{a} \end{bmatrix} \in \mathcal{M}_{2\times 2}(\mathbb{C}) \mid |a|^2 - |b|^2 = 1 \right\} \end{aligned}$$

Clearly, the projective action of $\mathrm{U}(1,1)$ on $\mathbb{C}P^1$ preserves $\mathbb{D} = \mathbf{P}\{q < 0\}$, since $\mathrm{U}(1,1)$ preserves q by definition. Conversely, any projective transformation of $\mathbb{C}P^1$ preserving $\mathbb{D}$ is induced by some element of $\mathrm{U}(1,1)$, see Exercise 9.10. Also, note that since any element $M \in \mathrm{U}(1,1)$ may be written $M = uA$ with $A \in \mathrm{SU}(1,1)$, the projective action of M and A coincide, and the inclusion $\mathrm{SU}(1,1) \subseteq \mathrm{U}(1,1)$ induces an isomorphism $\mathrm{PSU}(1,1) \approx \mathrm{PU}(1,1)$.

Theorem 9.48. *Let $f\colon \mathbb{D} \to \mathbb{D}$. The following are equivalent:*

(i) f is an orientation-preserving Möbius transformation of $\mathbb{D}$.

(ii) f is a fractional linear transformation that preserves $\mathbb{D}$.

(iii) f is a fractional linear transformation induced by some element of $\mathrm{SU}(1,1)$.

(iv) f is an orientation-preserving conformal automorphism.

(v) f is a complex automorphism (a biholomorphism $\mathbb{D} \to \mathbb{D}$).

Proof. The equivalence (i) $\Leftrightarrow$ (ii) follows from the general case Theorem 9.38 and from the characterization of Möbius transformations of $\hat{\mathbb{C}}$ as fractional linear transformations

(Theorem 9.44). The equivalence (ii) $\Leftrightarrow$ (iii) follows from the discussion above the theorem. The equivalence (iv) $\Leftrightarrow$ (v) is Corollary 9.42.

Finally, let us show that (iii) $\Leftrightarrow$ (v). It is clear that (iii) $\Rightarrow$ (v), since fractional linear transformations are holomorphic and bijective. It remains to show that conversely, any complex biholomorphism of $\mathbb{D}$ is fractional linear. This is the hardest part of the theorem, which requires some basic knowledge of complex analysis.

So, let $f\colon \mathbb{D} \longrightarrow \mathbb{D}$ be a biholomorphism. After composing f with a fractional linear transformation (that preserves $\mathbb{D}$), we can assume that $f(0) = 0$. Specifically, one may post-compose f with $z \mapsto \frac{z-a}{1-\bar{a}z}$ where $a = f(0)$. Conclude with the lemma of Schwarz (see e.g. [Ahl78, Theorem 13]) that $f(z) = uz$ for some $u \in \mathbb{C}$ with $|u| = 1$. [If you are unfamiliar with the lemma of Schwarz, the argument is essentially as follows: apply the maximum principle to the function $g(z) = \frac{f(z)}{z}$, which can be holomorphically extended at $z = 0$ by $g(0) = f'(0)$. By applying the maximum principle to g on the disk $D(0, r)$ with $r \longrightarrow 1$, we obtain that $|g| \leqslant 1$ on $\mathbb{D}$. On the other hand, switching f and f^{-1} if necessary, we have $|g'(0)| \geqslant 1$. By the maximum principle, g is constant.] In particular, $f(z) = uz$ is fractional linear. ■

Corollary 9.49. *We have isomorphisms:*

$$\mathrm{M\ddot{o}b}^+(B^2) \approx \mathrm{Aut}(\mathbb{D}) \approx \mathrm{PSU}(1,1)\,.$$

9.5.4 Möbius transformations of $\mathbb{H}$

We have seen that in general the Cayley transform is the conformal equivalence $c\colon H^n \longrightarrow B^n$ which can be described as $c = \tau \circ s$, where τ is the reflection through the $x_n = 0$ hyperplane and s is the inversion through the sphere $S(a, r)$ where $a = (0, \dots, 0, -1)$ and $r^2 = 2$. In the case $n = 2$, using the complex variable z, we find $\tau(z) = \bar{z}$ and $s(z) = -i\frac{\bar{z}-i}{\bar{z}+i}$, which gives us the expression of the Cayley transform and its inverse:

$$\begin{aligned} c\colon \mathbb{H} &\longrightarrow \mathbb{D} & \qquad c^{-1}\colon \mathbb{D} &\longrightarrow \mathbb{H} \\ z &\longmapsto i\,\frac{z-i}{z+i} & z &\longmapsto -i\,\frac{z+i}{z-i} \end{aligned}$$

Note that c is induced by the linear map $C\colon \mathbb{C}^2 \longrightarrow \mathbb{C}^2$ with matrix

$$C = \begin{bmatrix} i & 1 \\ 1 & i \end{bmatrix}.$$

Our discussion from the previous subsection (Möbius transformations of $\mathbb{D}$) can be transported to $\mathbb{H}$ via the Cayley transform.

Consider the Hermitian form $\tilde{h} = C^* h$ associated to the matrix

$$\tilde{H} = C^{\mathrm{T}} H \bar{C} = \begin{bmatrix} 0 & 2i \\ -2i & 0\,. \end{bmatrix}$$

This is a Hermitian form on $\mathbb{C}^2$ of signature $(1, 1)$, with associated quadratic form

$$\begin{aligned}\tilde{q}(z_1, z_2) &= 2i(z_1\bar{z_2} - z_2\bar{z_1}) \\ &= -\operatorname{Im}(z_1\bar{z_2})\end{aligned}$$

As expected, the locus $\{\tilde{q} < 0\}$ in $\mathbb{C}^2$ is the cone over $\mathbb{H} = \{\operatorname{Im}(z) > 0\} \subseteq \mathbb{C}\mathbf{P}^1$, since $\operatorname{Im}(z) = \frac{\operatorname{Im}(z_1\bar{z_2})}{|z_2|^2}$ for $z = \frac{z_1}{z_2}$.

The subgroup of $\mathrm{GL}(2, \mathbb{C})$ preserving $\{\tilde{q} < 0\}$ is $\mathbb{C}^* \, \mathrm{U}(\tilde{h}) = \mathbb{C}^* \, \mathrm{SU}(\tilde{h})$, where $\mathrm{U}(\tilde{h})$ [resp. $\mathrm{SU}(\tilde{h})$] is the subgroup of $\mathrm{GL}(2, \mathbb{C})$ [resp. $\mathrm{SL}(2, \mathbb{C})$] preserving $\tilde{h}$. In terms of matrices:

$$\begin{aligned}\mathrm{SU}(\tilde{h}) &= \{M \in \mathrm{SL}(2, \mathbb{C}) \mid M^{\mathsf{T}}\tilde{H}\bar{M} = \tilde{H}\} \\ &= \mathrm{SL}(2, \mathbb{R})\end{aligned}$$

Indeed, we leave it to the reader as an easy exercise to check that for $M \in \mathrm{SL}(2, \mathbb{C})$, $\bar{M}^{\mathsf{T}}\tilde{H} = \tilde{H}M^{-1}$ if and only if M has real coefficients.

Remark 9.50. Alternatively, we can write $\mathrm{SU}(\tilde{h}) = C^{-1}\,(\mathrm{SU}(1, 1))\,C$, and one can prove that $C^{-1}\,(\mathrm{SU}(1, 1))\,C = \mathrm{SL}(2, \mathbb{R})$ by direct computation.

By transporting Theorem 9.48 via the Cayley transform, we obtain:

Theorem 9.51. *Let $f \colon \mathbb{H} \longrightarrow \mathbb{H}$. The following are equivalent:*

(i) f is an orientation-preserving Möbius transformation of $\mathbb{H}$.

(ii) f is a fractional linear transformation that preserves $\mathbb{H}$.

(iii) f is a fractional linear transformation induced by an element of $\mathrm{SL}(2, \mathbb{R})$.

(iv) f is an orientation-preserving conformal automorphism.

(v) f is a complex automorphism (a biholomorphism $\mathbb{H} \longrightarrow \mathbb{H}$).

Corollary 9.52. *We have isomorphisms:*

$$\mathrm{Möb}^+(H^2) \approx \mathrm{Aut}(\mathbb{H}) \approx \mathrm{PSL}(2, \mathbb{R})\,.$$

Since we also have isomorphisms $\mathrm{Möb}^+(B^2) \approx \mathrm{Möb}^+(H^2) \approx \mathrm{Möb}^+(S^1) \approx \mathrm{PSO}(2, 1)$, we obtain the "accidental" isomorphisms of Lie groups:

$$\mathrm{PSU}(1, 1) \approx \mathrm{PSL}(2, \mathbb{R}) \approx \mathrm{PSO}(2, 1)\,.$$

9.6 Exercises

Exercise 9.1.

Characterization of conformal maps of $\mathbb{R}^n$.

Let V, W be Euclidean vector spaces and $\Omega \subseteq V$ be an open set. Consider an immersion $f\colon \Omega \to V$.

(1) Let γ_1 and γ_2 be two regular curves in Ω that intersect at $p \in \Omega$. Denote v_i the tangent vector to γ_i at p. Show that $f \circ \gamma_1$ and $f \circ \gamma_2$ are two regular curves in W that intersect at $f(p)$, and that the tangent vector to γ_i at $f(p)$ is $\mathrm{d}f(v_i)$.

(2) Prove Proposition 9.6: *Let $f\colon \Omega \subseteq V \to W = V$. Then f is conformal if and only if f is differentiable and $\mathrm{d}f_x$ is a linear similarity for all $x \in \Omega$.*

(3) Prove Proposition 9.7: *$f\colon \Omega \subseteq \mathbb{C} \to \mathbb{C}$ is conformal if and only if f is holomorphic or antiholomorphic and f' does not vanish.* (This question requires basic knowledge of holomorphic functions.)

Exercise 9.2.

Characterization of conformal maps between Riemannian manifolds

Let (M, g) and (N, h) be Riemannian manifolds.

(1) Let $f\colon V \to W$ be a linear map between vector spaces. For any bilinear form b on W, we define the bilinear form f^*b on V by $f^*b(u, v) := b(f(u), f(v))$. Show that if b is an inner product, f^*b is an inner product if and only if f is injective.

(2) Let $f\colon (V, \langle\cdot,\cdot\rangle_V) \to (W, \langle\cdot,\cdot\rangle_W)$ be a linear map between Euclidean vector spaces. Show that f is angle-preserving if and only if there exists $\lambda \in \mathbb{R}_{>0}$ such that $f^*\langle\cdot,\cdot\rangle_W = \lambda\langle\cdot,\cdot\rangle_V$.

(3) Let $f\colon (M, g) \to (N, h)$ be a differentiable map between Riemannian manifolds. How do you define the pullback f^*h? Show that f is conformal if and only if f^*h is conformal to g.

Exercise 9.3.

Full vs restricted Möbius group

Denote $\text{Möb}^+(S^n)$ the restricted Möbius group of S^n, consisting of orientation-preserving Möbius transformations.

(1) Show that $\text{Möb}^+(S^n)$ is an index 2 normal subgroup of $\text{Möb}(S^n)$.

(2) Show that $\text{Möb}^+(S^n)$ is the identity component of $\text{Möb}(S^n)$.

(3) Show the same results for $\text{Möb}^+(B^n) < \text{Möb}(B^n)$ and $\text{Möb}^+(\widehat{\mathbb{R}^n}) < \text{Möb}(\widehat{\mathbb{R}^n})$.

Exercise 9.4.

Inversions

(1) Let $S = S(a, r)$ be the sphere of center a and radius r in $\mathbb{R}^n$. What is its Cartesian equation? Show that the inversion through S has the expression:

$$f(x) = a + \frac{r^2}{\|x-a\|^2}(x-a)\,.$$

(2) Let $P \subseteq \mathbb{R}^n$ be an affine hyperplane. Denote v a nonzero normal vector and $\lambda \in \mathbb{R}$ such that $x_0 = \lambda v$ belongs to P (why is λ well-defined?). Show that the Cartesian equation of P is $\langle x - x_0, v\rangle = 0$. Show that the inversion through P has the expression:

$$f(x) = x - 2\langle x - x_0, v\rangle \frac{v}{\|v\|^2}\,.$$

(3) Show that the results of (2) may be obtained by taking the limit of (1) with $a = x_0 + tv$ and $r = t\|v\|$ when $t \longrightarrow +\infty$.

(4) Recover the result that any finite product of inversions may be written

$$f(x) = b + \frac{\alpha A(x-a)}{|x-a|^{\varepsilon}}$$

where $a, b \in \mathbb{R}^n$, $\alpha \in \mathbb{R}$, $A \in \mathrm{O}(n)$, and $\varepsilon \in \{0, 2\}$.

Exercise 9.5.

More inversions

(1) Show that any translation $\mathbb{R}^n \longrightarrow \mathbb{R}^n$ can be written as a product of two reflections. Could you expect such a result?

(2) Show that any linear similarity $\mathbb{R}^n \longrightarrow \mathbb{R}^n$ can be written as a product of two inversions. Could you expect such a result?

Exercise 9.6.

Möbius transformations vs Euclidean similarities

Show that the subgroup of $\text{Möb}(\widehat{\mathbb{R}^n})$ fixing ∞ is isomorphic to the group of affine similarities of $\mathbb{R}^n$.

Exercise 9.7.

Stereographic projection

(1) Recover the expression of the standard stereographic projection $s\colon S^n \longrightarrow \widehat{\mathbb{R}^n}$.

(2) Recover that the stereographic projection is the restriction to S^n of an inversion of $\widehat{\mathbb{R}^{n+1}}$. Derive that s is a conformal equivalence.

(3) Recover that s is conformal by direct computation: compute the pullback Riemannian metric s^*g on $S^n - \{N\}$, where g is the Euclidean metric on $\mathbb{R}^n$.

Exercise 9.8.

Poincaré extension

(1) Find the Poincaré extension of an inversion of $\widehat{\mathbb{R}^n}$.

(2) Write a new proof of the existence of the Poincaré extension of a Möbius transformation. Can you extend your argument to also prove uniqueness?

Exercise 9.9.

Möbius transformations of $\hat{\mathbb{C}}$

The goal of this exercise is to show Theorem 9.43: *A map $f\colon \hat{\mathbb{C}} \longrightarrow \hat{\mathbb{C}}$ is an Möbius transformation if and only if it is fractional linear (orientation-preserving case) or its conjugate is fractional linear (orientation-reserving case).*

(1) Argue that it is enough to show that f is an orientation-preserving Möbius transformation if and only if it is fractional linear.

(2) (a) Show that the inversion through the sphere $S(a, r)$ can be written $f(z) = a + \frac{r^2}{\bar{z}-\bar{a}}$.

(b) Show that the inversion through the line with normal vector v going through the point $z_0 = \lambda v$ can be written $f(z) = 2z_0 - \frac{v}{\bar{v}}\bar{z}$.

(c) Show that the composition of any two inversions is fractional linear. Conclude that any Möbius transformation of $\hat{\mathbb{C}}$ is fractional linear.

(3) (a) Show that any fractional linear transformation may be written as a composition of maps of the form: $z \mapsto z + b$ where $b \in \mathbb{C}$, $z \mapsto az$ where $a \in \mathbb{C}^*$, and $z \mapsto \frac{1}{z}$.
(b) Show that the three maps of the previous question may be written as a product of inversions.
(c) Conclude that any fractional linear transformation is a Möbius transformation of $\hat{\mathbb{C}}$.

Exercise 9.10.

The group $\mathrm{PSU}(1,1)$

(1) Recall the definition of $\mathrm{SU}(1,1)$ and show that

$$\mathrm{SU}(1,1) = \left\{ \begin{bmatrix} a & b \\ \bar{b} & \bar{a} \end{bmatrix} \in \mathcal{M}_{2\times 2}(\mathbb{C}) \mid |a|^2 - |b|^2 = 1 \right\}$$

(2) Show that $\mathrm{U}(1,1) = \{uA \mid |u| = 1, A \in \mathrm{SU}(1,1)\}$. Derive that $\mathrm{PU}(1,1) \approx \mathrm{PSU}(1,1)$.

(3) Show that the action of any element of $\mathrm{U}(1,1)$ by fractional linear transformation can be written

$$z \mapsto u\frac{z-a}{1-\bar{a}z}$$

where $|u| = 1$ and $a \in \mathbb{D}$.

(4) Recover from the previous question that the action of $\mathrm{U}(1,1)$ on $\hat{\mathbb{C}}$ preserves $\mathbb{D}$.

(5) Prove that conversely, a fractional linear transformation preserving $\mathbb{D}$ coincides with the action of an element of $\mathrm{U}(1,1)$.

(6) Recall why $\mathrm{Möb}^+(\mathbb{D}) \approx \mathrm{Aut}(\mathbb{D}) \approx \mathrm{PSU}(1,1)$.

Exercise 9.11.

The group $\mathrm{PSL}(2,\mathbb{R})$

(1) Recover by direct proof that the Cayley transform $c(z) = i\,\frac{z-i}{z+i}$ defines a biholomorphism from $\mathbb{H}$ to $\mathbb{D}$.

(2) Recover by direct proof that the fractional linear action of $M \in \mathrm{SL}(2,\mathbb{C})$ on $\hat{C}$ preserves $\mathbb{H}$ if and only if M has real coefficients.

(3) Recover by direct proof that $\mathrm{SL}(2,\mathbb{R}) = C^{-1}\,(\mathrm{SU}(1,1))\,C$ where $C = \begin{bmatrix} i & 1 \\ 1 & i \end{bmatrix}$. Recall the connection between this result and the previous question.

(4) Show that there are natural "inclusions"

$$\mathrm{PSL}(2,\mathbb{R}) \hookrightarrow \mathrm{PGL}(2,\mathbb{R}) \hookrightarrow \mathrm{PGL}(2,\mathbb{C})$$
$$\mathrm{PSL}(2,\mathbb{R}) \hookrightarrow \mathrm{PSL}(2,\mathbb{C}) \xrightarrow{\sim} \mathrm{PGL}(2,\mathbb{C})$$

How would you describe the difference between $\mathrm{PSL}(2,\mathbb{R})$ and $\mathrm{PGL}(2,\mathbb{R})$?

Exercise 9.12.

The one-dimensional case

Throughout the chapter, we discussed conformal maps and Möbius transformations of $\widehat{\mathbb{R}^n}$, S^n, H^n, B^n for $n \geqslant 2$. What about the case $n = 1$? Work out as many details as possible about what still works and what breaks.

CHAPTER 10

The Poincaré models

Disclaimer: This chapter is a draft.

In this chapter we present the Poincaré ball model and the Poincaré half-space model of hyperbolic geometry. These are conformal models, meaning that they can be defined as Euclidean domains equipped with a metric that is conformally equivalent to the Euclidean metric.

Alternatively, the Poincaré ball model may be obtained from the hyperboloid model studied in Chapter 5 via a stereographic projection, and the half-space model may be derived via a Möbius transformation called the Cayley transform. We will use these relations to showcase the essential features of these models.

Historically, both Poincaré models of the hyperbolic plane were discovered by Eugenio Beltrami in 1868 ([Bel68a; Bel68b]), alongside the Beltrami–Klein model which we discussed in Chapter 8[1]. Poincaré rediscovered the half-plane and disk models in 1882 and revealed the connection between 2-dimensional hyperbolic geometry and complex geometry, especially Fuchsian groups and automorphic functions [Poi82].

[1] Beltrami also discovered the pseudosphere in [Bel68a], which we prefer to call tractricoid: see Exercise 2.6.

10.1 The Poincaré ball model

10.1.1 Stereographic projection of the hyperboloid

Let $n \geqslant 2$ be an integer (we could also allow $n = 1$ for most of this chapter). Embed $\mathbb{R}^n$ in Minkowski space $\mathbb{R}^{n,1}$ in the obvious way:

$$\begin{aligned} \mathbb{R}^n &\longrightarrow \mathbb{R}^{n,1} \\ (x_1, \ldots, x_n) &\longmapsto (x_1, \ldots, x_n, 0) . \end{aligned}$$

Consider the point $S = (0, \ldots, 0, -1) \in \mathbb{R}^{n,1}$ (the "South pole"). Let us denote $\mathcal{H}^+ \subseteq \mathbb{R}^{n,1}$ the hyperboloid as in Chapter 5 and $B^n \subseteq \mathbb{R}^n$ the Euclidean unit ball. We call ***stereographic projection*** of the hyperboloid from the point S the map $s\colon \mathcal{H}^+ \longrightarrow B$ such that for every $x \in \mathcal{H}^+$ and $x' \in B^n$, the points S, x', x are collinear if and only if $x' = s(x)$. See Figure 10.1.

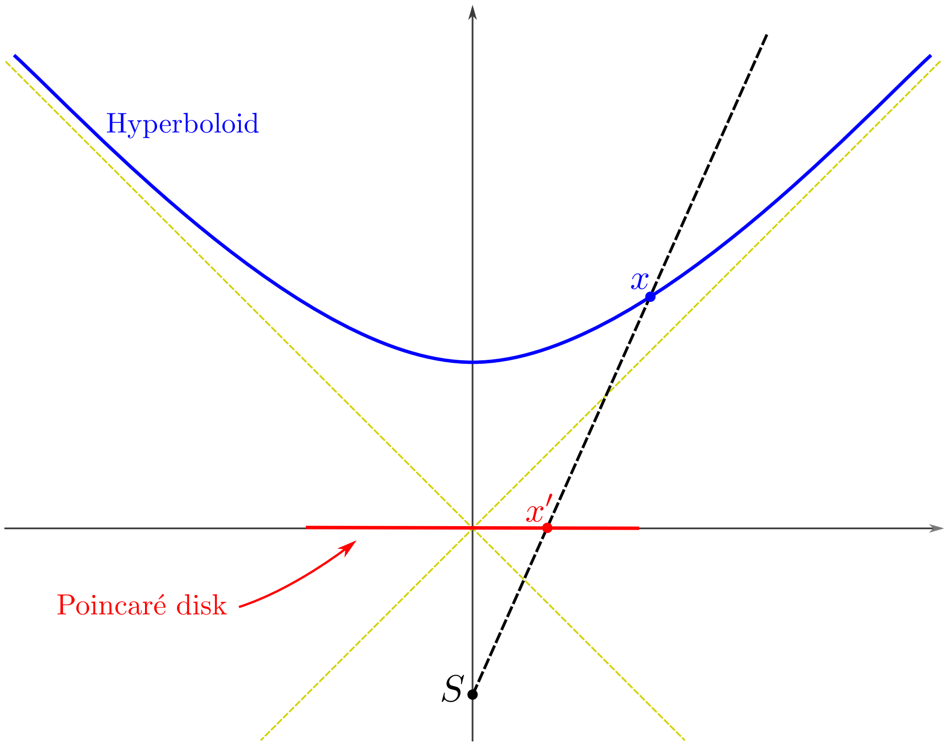


Figure 10.1: Stereographic projection of the hyperboloid to the Poincaré disk.

It is elementary to compute the analytic expression of the map s, and thereby prove that it is well-defined and bijective: writing $(x' - S) = \lambda(x - S)$ yields $\lambda = \frac{1}{1+x_{n+1}}$ by examining the last coordinate. Therefore we find:

$$x'_k = \frac{x_k}{1 + x_{n+1}}$$

for $k \in \{1, \ldots, n\}$. In order to find the inverse, one can find the expression of λ in terms of x' by writing that $\langle x, x \rangle = -1$ (since $x \in \mathcal{H}^+$), which yields $\lambda = \frac{1-\|x'\|^2}{2}$. Therefore we find:

$$x_k = \frac{2x_k'}{1 - \|x'\|^2} \ (k \in \{1, \ldots, n\}) \qquad x_{n+1} = \frac{1}{\lambda} - 1 = \frac{1 + \|x'\|^2}{1 - \|x'\|^2} .$$

In particular, we see that the stereographic projection s is a smooth (even real-analytic) diffeomorphis from the hyperboloid $\mathcal{H}^+$ to the ball B^n.

Definition 10.1. The ***Poincaré ball*** (or ***Poincaré disk***) (B^n, g_{B^n}) is the image of the hyperboloid $(\mathcal{H}^+, g_{\mathcal{H}^+})$ by the stereographic projection $s \colon \mathcal{H}^n \to B^n$.

This definition means that we use the stereographic projection to transport the geometry of the hyperboloid to the unit ball. Technically, it is enough to transport the Riemannian metric, since all other geometric features follow: distance, geodesics, isometries, etc. It follows immediately from its definition that the Poincaré ball is a model of hyperbolic space:

Theorem 10.2. *The Poincaré ball (B^n, g_{B^n}) is a complete, simply-connected Riemannian manifold of constant sectional curvature -1.*

Remark 10.3. We have seen several different stereographic projections in this course. Their common feature is that they are all projections to a (hyper)plane by drawing lines from a single point. The stereographic projection of the hyperboloid to the Poincaré ball is especially similar to the stereographic projection of the hyperboloid to the Klein ball (see Figure 8.2). Nevertheless, the Poincaré ball and the Klein ball are significantly different models. See Exercise 10.3.

10.1.2 Riemannian metric

By definition, the hyperbolic metric (also called Poincaré metric) g_{B^n} is the pullback of the hyperbolic metric $g_{\mathcal{H}^+}$ on the hyperboloid by $s^{-1} \colon B^n \to \mathcal{H}^+$. We leave it as an exercise (Exercise 10.1) to derive its explicit expression:

$$\mathrm{d}s^2 = 4\,\frac{\mathrm{d}x_1^2 + \cdots + \mathrm{d}x_n^2}{(1 - \|x\|^2)^2} .$$

Remark 10.4. Of course, we could have defined the Poincaré ball by giving the Riemannian metric above, and then proved that it is isometric to the hyperboloid via stereographic projection.

We immediately note that $g_{B^n} = f g_0$, where g_0 is the Euclidean metric in B^n and $f(x) = \frac{4}{(1-\|x\|^2)^2}$ is a smooth function on B^n. This shows that the Poincaré metric is conformally equivalent to the Euclidean metric in B^n (see § 9.1.3). In short, we say that the Poincaré ball is a *conformal model* of hyperbolic space.

Remark 10.5. Note that $\lim_{\|x\| \to 1} f(x) = +\infty$: the conformal factor blows up as one approaches the boundary of the ball. This is expected because the hyperbolic metric in B^n is complete (unlike the Euclidean metric), therefore point of ∂B^n should be infinitely far away.

10.1.3 Distance

The distance function on the Poincaré ball can be computed directly as the pullback of the distance on the hyperboloid:

Proposition 10.6. *The distance in the Poincaré ball is given by*

$$d(x, y) = \operatorname{arcosh}\left(1 + \frac{2\|x - y\|^2}{(1 - \|x\|^2)(1 - \|y\|^2)}\right).$$

Proof. Since the stereographic projection $s\colon \mathcal{H}^+ \to B^n$ is a Riemannian isometry, it is also a metric isometry for the induced distances. Thus one can compute the distance on B^n as the pullback of the distance on $\mathcal{H}^n$: we have $d_{B^n} = \left(s^{-1}\right)^* d_{\mathcal{H}^n}$. Concretely:

$$\begin{aligned} d_{B^n}(x, y) &= d_{\mathcal{H}^n}\left(s^{-1}(x), s^{-1}(y)\right) \\ &= \operatorname{arcosh}\left(-\langle s^{-1}(x), s^{-1}(y)\rangle\right) \end{aligned}$$

The conclusion quickly follows from inputting the explicit expressions of $s^{-1}(x)$ and $s^{-1}(y)$, namely

$$s^{-1}(x) = \left(\frac{2x}{1 - \|x\|^2}, \frac{1 + \|x\|^2}{1 - \|x\|^2}\right)$$

and similarly for $s^{-1}(y)$, and writing the Minkowski inner product. ∎

Remarkably, the metric can be rewritten almost like a Cayley–Klein metric. Let $x, y \in B^n$ be any two distinct points. As we shall see in § 10.1.5, the geodesic through x and y is a Euclidean circle arc, which intersects the sphere ∂B^n orthogonally in two points. Call the two boundary points I and J as in Figure 10.2. We have seen in the previous chapter (Remark 9.19)

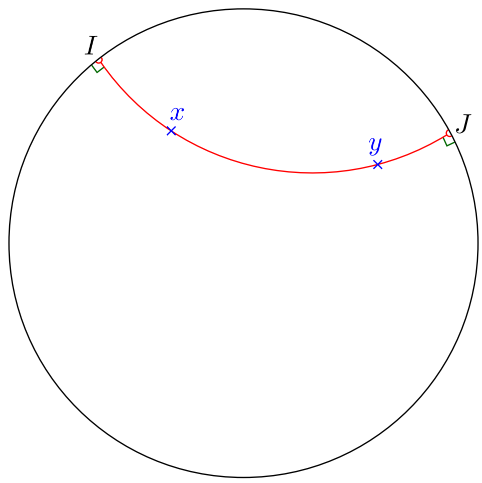


Figure 10.2: Geodesic in the Poincaré disk.

that one can define the (unsigned) cross-ratio of any 4-tuple of distinct points in $\mathbb{R}^n$. We claim:

Proposition 10.7. *The distance in the Poincaré ball is given by*

$$\begin{aligned} d(x,y) &= \ln[x,y,J,I] \\ &= \ln \frac{|Jx||Iy|}{|Jy||Ix|} \end{aligned} \tag{10.1}$$

It is a striking "coincidence" that the distance in the Poincaré ball can be written in such a similar fashion as the distance in the Klein ball (see Proposition 8.30)! Note however two differences: 1. There is a factor $\frac{1}{2}$ in the Cayley–Klein distance that does not appear here, and 2. The points I and J are different here, and the fours points I, x, y, J are not collinear in $\mathbb{R}^n$.

Proof. We shall see in § 10.1.4 that the isometries of the Poincaré ball are the Möbius transformations of the ball. Since Möbius transformations preserve cross-ratios (see Theorem 9.18), without loss of generality we can assume that $x = 0$ by choosing an isometry that maps x to 0 (recall that isometries act transitively on hyperbolic space). Since geodesics through the origin are diameters (see § 10.1.5), the geodesic through x and y is a diameter $[I, J]$. We thus have $|I_x| = 1$, $|Jx| = 1$, $|Iy| = 1 + r$, $|Jy| = 1 - r$ where $r = \|y\|$. Therefore

$$\begin{aligned} \ln \frac{|Jx||Iy|}{|Jy||Ix|} &= \ln \frac{1+r}{1-r} \\ &= 2 \operatorname{artanh} r \,. \end{aligned}$$

On the other hand, by Proposition 10.6 we have

$$\begin{aligned} d(x,y) &= \operatorname{arcosh}\left(1 + \frac{2r^2}{1-r^2}\right) \\ &= 2 \operatorname{arcosh} \frac{1}{\sqrt{1-r^2}} \\ &= 2 \operatorname{artanh} r \,. \end{aligned}$$

We used the identities: $\operatorname{arcosh}(2x^2 - 1) = 2 \operatorname{arcosh} x$ and $\operatorname{arcosh} \frac{1}{\sqrt{1-x^2}} = \operatorname{artanh} x$. ∎

10.1.4 Isometries

In the previous chapter, we introduced Möbius transformations of the ball B^n.

Theorem 10.8. *The group of isometries of the Poincaré ball is exactly the Möbius group of the ball:*

$$\begin{aligned} \operatorname{Isom}(B^n, g_{B^n}) &= \text{Möb}(B^n) \\ \operatorname{Isom}^+(B^n, g_{B^n}) &= \text{Möb}^+(B^n) \end{aligned}$$

Proof. It is enough to prove $\operatorname{Isom}(B^n, g_{B^n}) = \text{Möb}(B^n)$, the second identity follows immediately. Let us prove the mutual inclusion:

$\mathrm{Isom}(B^n, g_{B^n}) \subseteq \mathrm{Möb}(B^n)$: Since isometries are conformal, any $f \in \mathrm{Isom}(B^n, g)$ is a conformal automorphism of (B^n, g_{B^n}). Since g is conformally equivalent to the Euclidean metric g_0, f is also a conformal automorphism of (B^n, g_0). By Theorem 9.38, f is a Möbius transformation of B^n.

$\mathrm{Isom}(B^n, g_{B^n}) \supseteq \mathrm{Möb}(B^n)$: Since the Möbius group is generated by inversions, it is enough to prove that any inversion is an isometry. This can be checked by direct computation. Alternatively, since Möbius transformations preserves (unsigned) cross-ratios (Theorem 9.18), they preserve the distance (10.1). Conclude by remembering that distance-preserving maps and Riemannian isometries are the same. ■

The next theorem follows immediately from Theorem 9.39.

Theorem 10.9. *Any isometry of (B^n, g_{B^n}) uniquely extends continuously to $\partial B^n = S^{n-1}$, and the boundary map is a Möbius transformation of S^{n-1}. Conversely, any Möbius transformation $f \in \mathrm{Möb}(S^{n-1})$ extends to a unique isometry $\hat{f} \in \mathrm{Isom}(B^n, g_{B^n})$ called the* ***Poincaré extension*** *of f.*

Corollary 10.10. *We have isomorphisms:*

$$\mathrm{Isom}(B^n, g_{B^n}) \approx \mathrm{Möb}(S^{n-1}) \approx \mathrm{PO}(n,1)$$
$$\mathrm{Isom}^+(B^n, g_{B^n}) \approx \mathrm{Möb}^+(S^{n-1}) \approx \mathrm{PO}^+(n,1)$$

In dimension 2, the Poincaré disk $B^2 = \mathbb{D}$ can be identified as a subset of $\hat{\mathbb{C}}$, and the orientation-preserving Möbius group of $\mathbb{H}$ is identified to $\mathrm{PSU}(1,1)$ acting by fractional linear transformations. This is also the group of complex automorphisms of $\mathbb{D}$. (See § 9.5.3 for details.)

Corollary 10.11. *The group of orientation-preserving isometries of the Poincaré disk is:*

$$\mathrm{Isom}^+(B^2, g_{B^2}) \approx \mathrm{Aut}(\mathbb{D}) \approx \mathrm{PSU}(1,1)\,.$$

In dimension 3, the boundary S^2 of Poincaré ball B^3 can be identified to $\hat{\mathbb{C}}$ by stereographic projection, or to $\mathbb{C}P^1$ by the standard affine chart. Any isometry of B^3 is uniquely determined by its extension to the boundary, which is a Möbius transformation of $S^2 \approx \hat{\mathbb{C}} \approx \mathbb{C}P^1$. We have seen in § 9.5.2 that the orientation-preserving Möbius group of S^2 is identified to $\mathrm{PSL}(2,\mathbb{C})$ acting by fractional linear transformations on $\hat{\mathbb{C}}$ or by projective transformations of $\mathbb{C}P^1$, and that this is also the group of complex automorphisms of $\hat{\mathbb{C}}$.

Corollary 10.12. *The group of orientation-preserving isometries of the 3-dimensional Poincaré ball is:*

$$\mathrm{Isom}^+(B^3, g_{B^3}) \approx \mathrm{Aut}(\hat{\mathbb{C}}) \approx \mathrm{PGL}(2,\mathbb{C})\,.$$

10.1.5 Geodesics

Theorem 10.13. *The (unparametrized) geodesics of the Poincaré ball (B^n, g_{B^n}) are the intersections of B^n with circles in $\widehat{\mathbb{R}}^n$ that are orthogonal to $\partial B^n = S^{n-1}$.*

Remark 10.14. A circle in $\widehat{\mathbb{R}}^n$ is either a Euclidean circle in $\mathbb{R}^n$, or $l \cup \{\infty\}$ where l is a straight line in $\mathbb{R}^n$. Therefore geodesics of the Poincaré ball are either arcs of Euclidean circles orthogonal to S^{n-1} (geodesics not going through the origin), or diameters (geodesics through the origin). See Figure 10.3 for a few geodesics in the Poincaré disk ($n = 2$).

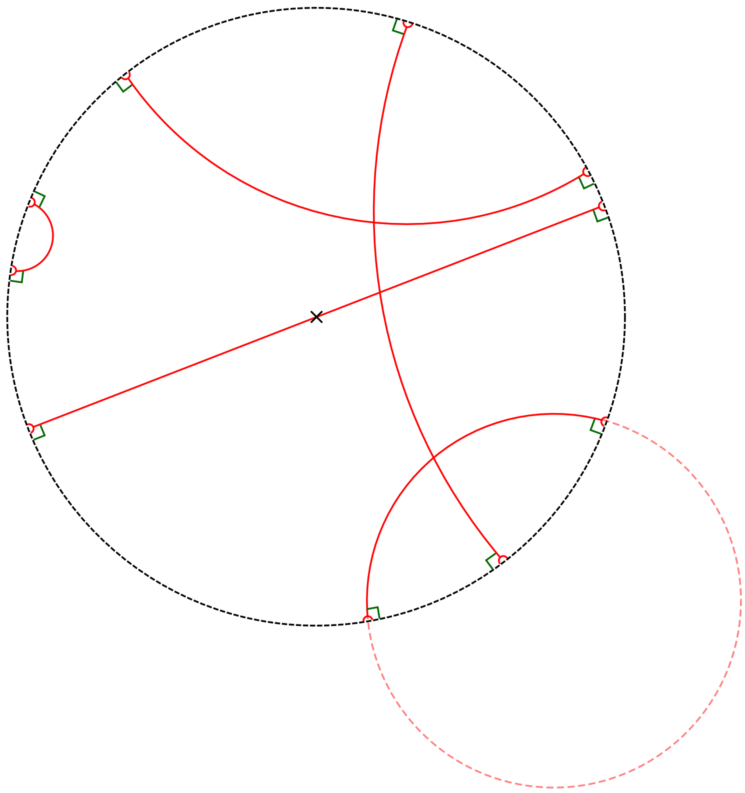


Figure 10.3: Geodesics in the Poincaré disk.

Proof. It follows from our definition of the Poincaré ball that geodesics in B^n are the image of geodesics in $\mathcal{H}^+$ under the stereographic projection s.

First let us show that geodesics through the origin are diameters. Any such geodesic is the image of a geodesic in $\mathcal{H}^+$ through the point $(0, \ldots, 0, 1)$, which is the intersection of $\mathcal{H}^+$

with a vertical 2-plane P. It is easy to see from the analytic expression of s that the image of $P \cap \mathcal{H}^+$ is $P \cap B^n$, which is a diameter.

Now let l be a geodesic of B^n that does not go through the origin. Let $x_0 \in l$. Since hyperbolic isometries act transitively, there exists $f \in \mathrm{Isom}(B^n, g_{B^n})$ such that $f(x_0) = 0$. Therefore $f(l) =: l'$ is a geodesic through the origin, so l' is a diameter. One can write $l' = C \cap B^n$, where C is a circle of $\widehat{\mathbb{R}^n}$ orthogonal to S^{n-1}. Since f^{-1} is a Möbius transformation, it is conformal and sphere-preserving, therefore $f^{-1}(C)$ is circle of $\widehat{\mathbb{R}^n}$ orthogonal to S^{n-1}. We conclude that l is an arc of Euclidean circle orthogonal to S^{n-1}.

Conversely, let us argue that any diameter or arc of Euclidean circle orthogonal to S^{n-1} is a Poincaré geodesic. Consider such an arc l and denote its endpoints $I, J \in S^{n-1}$. Let l_0 be any geodesic through the origin, it is a diameter with endpoints $I_0, J_0 \in S^{n-1}$. There exists a Möbius transformation $f \in \mathrm{Möb}(S^{n-1})$ such that $f(I_0) = I$ and $f(J_0) = J$. Indeed, it is not hard to argue with a little work that $\mathrm{Möb}(S^{n-1})$ acts 2-transitively on S^{n-1} (when $n = 2$, it actually acts 3-transitively by Theorem 7.70). Let $\hat{f}$ be the Poincaré extension of f. Since $\hat{f}$ is a Möbius transformation, it sends l_0 to a circle of arc that intersects S^{n-1} orthogonally at I and J. We leave it as an exercise of Euclidean geometry to show that such an arc is unique, therefore $\hat{f}(l_0) = l$. On the other hand, $\hat{f}(l_0)$ is a geodesic since $\hat{f}$ is an isometry of the Poincaré ball. ∎

10.2 The Poincaré half-space model

10.2.1 Definition via the Cayley transform

We recall that the Cayley transform is a map $c \colon H^n \to B^n$, where $H^n \subseteq \mathbb{R}^n$ is the upper half-space. It is the restriction of an orientation-preserving Möbius transformation of $\widehat{\mathbb{R}^n}$, in particular c is a conformal equivalence between H^n and B^n. See § 9.4.2 for details and the analytic expression of the Cayley transform (also § 9.5.4 for $n = 2$).

Definition 10.15. The Poincaré upper half-plane (H^n, g_{H^n}) is the inverse image of the Poincaré ball (B^n, g_{B^n}) by the Cayley transform.

As before, we immediately obtain that the Poincaré upper half-plane is a model of hyperbolic space:

Theorem 10.16. *The Poincaré half-space (H^n, g_{H^n}) is a complete, simply-connected Riemannian manifold of constant sectional curvature -1.*

10.2.2 Riemannian metric

The Poincaré metric g_{H^n} can be computed as the pullback of g_{B^n} by the Cayley transform c. We leave the computation as an exercise to the reader (Exercise 10.1). One finds:

$$\mathrm{d}s^2 = \frac{\mathrm{d}x_1^2 + \cdots + \mathrm{d}x_n^2}{{x_n}^2}\,.$$

We note once again that g_{H^n} is a conformal metric (i.e. conformally equivalent to the Euclidean metric g_0), with conformal factor $f(x) = \frac{1}{x_n^2}$. This was to be expected: we already know that g_{B^n} is a conformal metric in B^n, and the Cayley transform is a conformal map.

Remark 10.17. As expected, the Riemannian metric blows up when $x_n \longrightarrow 0$, that is when x approaches $\partial H^n = \widehat{\mathbb{R}^{n-1}}$.

10.2.3 Distance

The Poincaré distance on H^n can be computed explicitly as $d_{H^n}(x,y) = d_{B^n}(c(x), x(y))$. Indeed, since the Cayley transform is a Riemannian isometry, it is also a metric isometry. After a few lines of calculations which we leave to the reader, one finds:

$$d(x,y) = \operatorname{arcosh}\left(1 + \frac{\|x-y\|^2}{2x_n y_n}\right) \tag{10.2}$$

Alternatively, one may again express the distance in terms of a cross-ratio:

$$\begin{aligned} d(x,y) &= \ln[x,y,J,I] \\ &= \ln\frac{|Jx||Iy|}{|Jy||Ix|}\,. \end{aligned}$$

Here, $I, J \in \widehat{\mathbb{R}^{n-1}}$ are now the ideal endpoints of the geodesic through x and y, which is a circle arc orthogonal to $\widehat{\mathbb{R}^{n-1}}$ (see § 10.2.5). The proof of this identity is quickly derived from the Poincaré ball case: since the Cayley transform is (the restriction of) a Möbius transformation of $\widehat{\mathbb{R}^n}$, it preserves cross-ratios.

10.2.4 Isometries

Theorem 10.18. *The group of isometries of the Poincaré half-space is exactly the Möbius group of the upper half-space:*

$$\begin{aligned} \operatorname{Isom}(H^n, g_{H^n}) &= \text{Möb}(H^n) \\ \operatorname{Isom}^+(H^n, g_{H^n}) &= \text{Möb}^+(H^n) \end{aligned}$$

Proof. Since (H^n, g_{H^n}) is the inverse image of (B^n, g_{B^n}) by the Cayley transform $c\colon H^n \longrightarrow B^n$, the group of isometries of (H^n, g_{H^n}) is conjugate to that of (B^n, g_{B^n}) by the Cayley transform: $\operatorname{Isom}(H^n, g_{H^n}) = c^{-1}\left(\operatorname{Isom}(B^n, g_{B^n})\right) c$. On the other hand, we know that $\operatorname{Isom}(B^n, g_{B^n}) = \text{Möb}(B^n)$, and the Cayley transform conjugates $\text{Möb}(H^n)$ and $\text{Möb}(B^n)$. ∎

Corollary 10.19. *We have isomorphisms:*

$$\mathrm{Isom}(H^n, g_{H^n}) \approx \mathrm{M\ddot{o}b}(\widehat{\mathbb{R}^{n-1}}) \approx \mathrm{PO}(n,1)$$
$$\mathrm{Isom}^+(H^n, g_{H^n}) \approx \mathrm{M\ddot{o}b}^+(\widehat{\mathbb{R}^{n-1}}) \approx \mathrm{PO}^+(n,1)$$

In dimension 2, the Poincaré half-plane $H^2 = \mathbb{H}$ can be identified as a subset of $\hat{\mathbb{C}}$, and the orientation-preserving Möbius group of $\mathbb{H}$ is identified to $\mathrm{PSL}(2,\mathbb{R})$ acting by fractional linear transformations. This is also the group of complex automorphisms of $\mathbb{H}$. (See § 9.5.4 for details.)

Corollary 10.20. *The group of orientation-preserving isometries of the Poincaré half-plane is:*

$$\mathrm{Isom}^+(H^2, g_{B^2}) \approx \mathrm{Aut}(\mathbb{H}) \approx \mathrm{PSL}(2,\mathbb{R})\,.$$

In dimension 3, the Poincaré half-space H^3 can be identified to $\mathbb{C}\times\mathbb{R}_{>0}$, and any isometry of H^3 is uniquely determined by its extension to the boundary $\partial H^3 = \hat{\mathbb{C}}$, which is a Möbius transformation of $\hat{\mathbb{C}}$. We have seen in § 9.5.2 that the orientation-preserving Möbius group of $\hat{\mathbb{C}}$ is identified to $\mathrm{PGL}(2,\mathbb{C})$ acting by fractional linear transformations, and that this is also the group of complex automorphisms of $\hat{\mathbb{C}}$.

Corollary 10.21. *The group of orientation-preserving isometries of the* 3*-dimensional Poincaré half-space is:*

$$\mathrm{Isom}^+(H^3, g_{H^3}) \approx \mathrm{Aut}(\hat{\mathbb{C}}) \approx \mathrm{PGL}(2,\mathbb{C})\,.$$

10.2.5 Geodesics

Theorem 10.22. *The (unparametrized) geodesics of the Poincaré half-space (H^n, g_{H^n}) are the intersections of H^n with circles in $\widehat{\mathbb{R}}^n$ that are orthogonal to $\partial H^n = \widehat{\mathbb{R}^{n-1}}$.*

Remark 10.23. A circle in $\widehat{\mathbb{R}^n}$ is either a Euclidean circle in $\mathbb{R}^n$, or $l \cup \{\infty\}$ where l is a straight line in $\mathbb{R}^n$. Therefore geodesics of the Poincaré half-space are either arcs of Euclidean circles orthogonal to $\widehat{\mathbb{R}^{n-1}}$, or vertical straight lines. See Figure 10.4 for a few geodesics in the Poincaré half-plane ($n = 2$).

Proof. Geodesics in H^n are the inverse images of geodesics in B^n by the Cayley transform, and conversely. Since the Cayley transform is (the restriction of) a Möbius transformation of $\widehat{\mathbb{R}^n}$, it maps circles orthogonal to ∂H^n to circles orthogonal to ∂B^n, and conversely. ∎

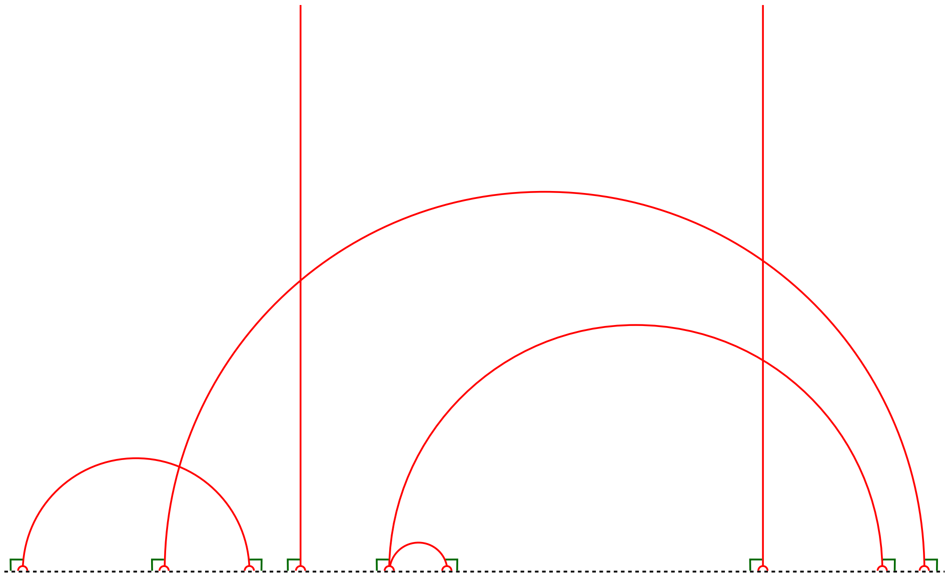


Figure 10.4: Geodesics in the Poincaré half-plane.

10.3 Exercises

Exercise 10.1.

Poincaré metric

Feel free to take $n = 2$ in this exercise. You can always do the general case afterwards.

(1) Recover the expression of the stereographic projection $s\colon \mathcal{H}^+ \longrightarrow B^n$.

(2) Recall the expressions of the Riemannian metrics $g_{\mathcal{H}^+}$ and g_{B^n} and recover the fact that s is a Riemannian isometry.

(3) Recover the expression of the Cayley transform $c\colon H^n \longrightarrow B^n$.

(4) Recall the expression of the metric g_{H^n} and recover that c is a Riemannian isometry.

Exercise 10.2.

Curvature of the Poincaré metric

Let $\Omega \subseteq \mathbb{R}^n$ and let $g = e^{2\varphi} g_0$ be a conformal metric in Ω. Let u, v be an orthonormal pair of vectors in $\mathbb{R}^n$ and denote P the plane spanned by u and v. The following formula (reference:

[Kap]) gives the sectional curvature of the metric g at a point $x \in \Omega$ in the direction of P:

$$K_P = -e^{-2\varphi}\left[\mathrm{D}^2\varphi(u,u) + \mathrm{D}^2\varphi(v,v) + \|\nabla\varphi\|^2 - \langle\nabla\varphi, u\rangle^2 - \langle\nabla\varphi, v\rangle^2\right] .$$

(We have denoted $\nabla\varphi$ the gradient of φ.)

(1) Recover the curvature of the Poincaré metric in B^n by direct computation.

(2) Let $K < 0$. Can you find a metric of constant sectional curvature K in B^n?

(3) Same questions for H^n.

Exercise 10.3.

Poincaré vs Klein ball

(1) Show that the natural identification between the Poincaré ball and the Beltrami–Klein ball is given by the map

$$\begin{aligned} \varphi\colon B^n_{\mathrm{P}} &\longrightarrow B^n_{\mathrm{K}} \\ x &\longmapsto \frac{2x}{1+\|x\|^2} . \end{aligned}$$

(2) Recover that φ is a Riemannian isometry by direct computation. *Feel free to take* $n = 2$.

Exercise 10.4.

Poincaré vs Klein ball: the distance

(1) Let x, x' be two real numbers in $[0, 1)$ such that $x' = \frac{2x}{1+x^2}$. Show that $\frac{1+x'}{1-x'} = \left(\frac{1+x}{1-x}\right)^2$ and derive that $\operatorname{artanh} x' = 2 \operatorname{artanh} x$.

(2) Recover the fact that the map φ of Exercise 10.3 is a metric isometry, i.e. $d(\varphi(x), \varphi(y)) = d(x, y)$, in the case $y = 0$.

Exercise 10.5.

Poincaré vs Klein ball: isometries

$\mathrm{PO}(n, 1)$ acts by isometries on the Klein ball and the Poincare ball. Is this the same action on B^n? Show that the map φ of Exercise 10.3 conjugates the two actions.

Exercise 10.6.

Hemisphere model

Let S^n be the unit sphere in $\mathbb{R}^{n+1}$ and denote S^n_+ the upper hemisphere (with $x_{n+1} > 0$). We also denote $S = (0, \dots, 0, -1)$ the "South pole" of S^n. We recall that the Poincaré ball may be seen as the unit ball in $R^n \subseteq R^{n+1}$.

(1) Consider the stereographic projection $s \colon S^n \longrightarrow \widehat{\mathbb{R}^n}$. Find its analytic expression. Show that s restricts to a diffeomorphism $S^n_+ \longrightarrow B^n$.

(2) By definition, the ***hemisphere model*** $(S^n_+, g_{S^n_+})$ of hyperbolic space is the inverse image of the Poincaré ball (B^n, g_{B^n}) by the stereographic projection s. Prove that $g_{S^n_+}$ can be written:

$$\mathrm{d}s^2 = \frac{\mathrm{d}x_1^2 + \cdots + \mathrm{d}x_{n+1}^2}{x_{n+1}^2}\,.$$

In what sense is the hemisphere model a conformal model?

Exercise 10.7.

Relations between models

(1) Show that the different models of hyperbolic space are related as showed by the diagram in Figure 10.5.

(2) Show that geodesics in the hemisphere model are semi-circles that are orthogonal to the equator. Explain Figure 10.6.

(3) Recover that geodesics in the Poincaré half-space model are semi-circles that are orthogonal to the boundary.

Exercise 10.8.

Matrix model of hyperbolic 3-space

Let H denote the set of 2×2 matrices with complex coefficients that are Hermitian symmetric:

$$H = \{A \in \mathcal{M}_{2\times 2}(\mathbb{C}) \mid A^* = A\}$$

where we denote $A^* = \bar{A}^{\mathsf{T}}$.

(1) Let $q(A) = -\det(A)$. Show that $q(A)$ is a quadratic form on H, with associated symmetric bilinear form $b(A, B) = -\frac{1}{2}\operatorname{tr}(A \operatorname{Comat}(B)^{\mathsf{T}})$.

(2) Show that (H, b) is isomorphic to $\mathbb{R}^{3,1}$ via

$$(x_1, x_2, x_3, x_4) \mapsto \begin{bmatrix} x_1 + x_4 & x_2 + ix_3 \\ x_2 - ix_3 & x_1 - x_4 \end{bmatrix}.$$

(3) Let $H_1 = H \cap \mathrm{SL}(2, \mathbb{C})$. Show that H_1 is a model of hyperbolic 3-space. What is the Riemannian metric?

(4) Show that $\mathrm{SL}(2, \mathbb{C})$ acts on H_1 by isometries via $M \cdot A = M\,A\,M^*$. What is the stabilizer of I_2? Recover that $\mathrm{Isom}^+(\mathbb{H}^3) \approx \mathrm{PSL}(2, \mathbb{C})$ and $\mathbb{H}^3 \approx \mathrm{PSL}(2, \mathbb{C})/\mathrm{PSU}(2)$.

Exercise 10.9.

Hyperbolic subspace

Propose a definition of a hyperbolic subspace of a hyperbolic space $X = \mathbb{H}^n$, and describe the hyperbolic subspaces in all the different models of $\mathbb{H}^n$.

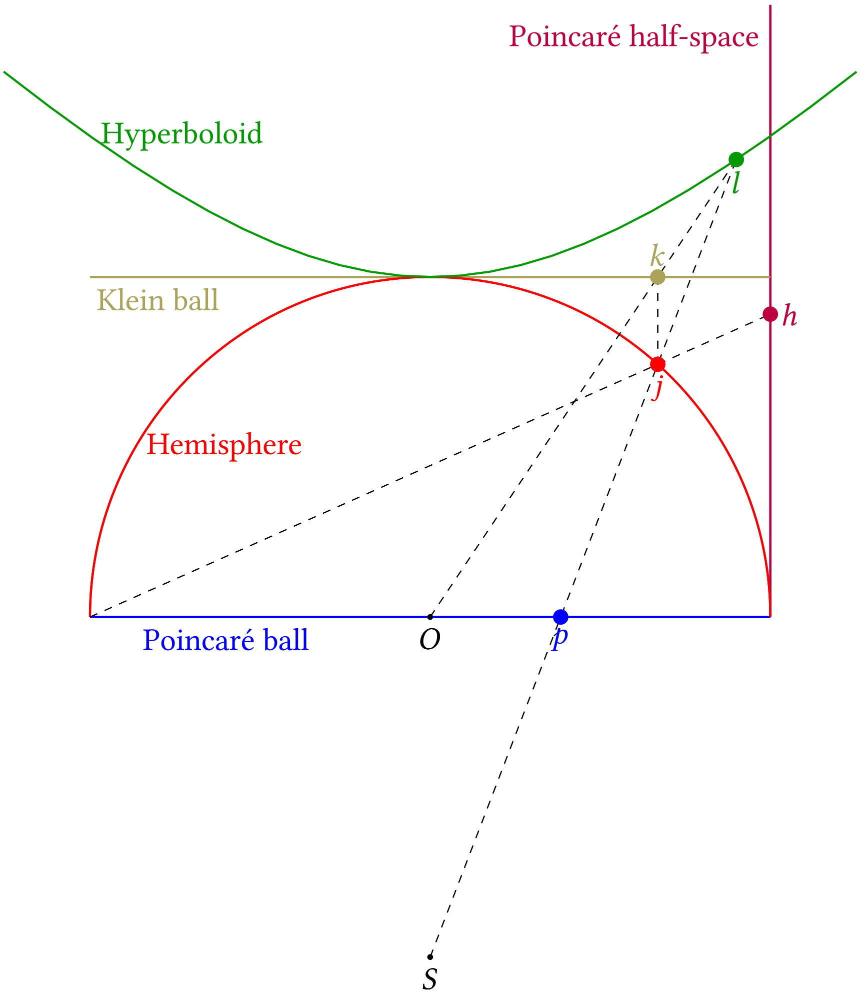


Figure 10.5: Relation between models of hyperbolic space.

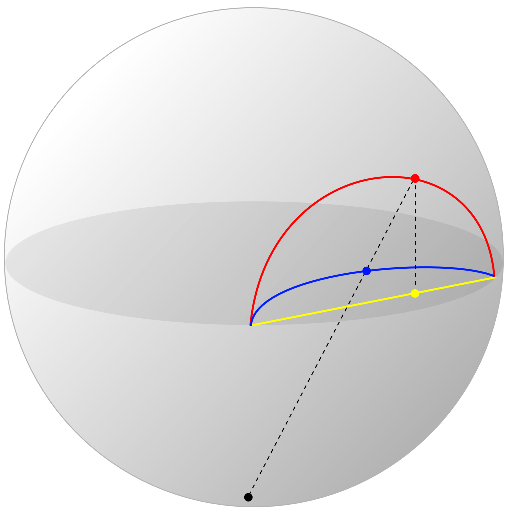


Figure 10.6: Geodesics in Poincaré ball, Klein ball, and hemisphere models.

Part V

Ideal boundary and classification of isometries

"The way I see the picture," said Gromov, "is that … we took two different, but sometimes overlapping, routes: Thurston concentrated on the most beautiful and difficult aspects of the field (hyperbolic 3-manifolds) and myself on the most general ones (hyperbolic groups)."

– Mikhail Gromov[2]

[2]Simons foundation, 2014. www.simonsfoundation.org/2014/12/22/mikhail-gromov/

CHAPTER 11

Ideal boundary of hyperbolic space

Disclaimer: This chapter is a draft.

In this chapter, we introduce the ideal boundary of hyperbolic space and study some of its most important properties. For instance, we will see that any geodesic is uniquely determined by its pair of ideal endpoints. We will also discuss the related notions of Busemann functions and horospheres. In the next chapter, we will make critical use of the ideal boundary in order to classify isometries of hyperbolic space.

The ideal boundary is not strictly speaking part of hyperbolic space: its points are "at infinity". Nevertheless, it can be defined intrinsically from hyperbolic space, and offers a compactification of it that is geometrically meaningful.

Most of the notions of this chapter are naturally defined in a much more general framework, namely metric spaces of nonpositive curvature. Specifically, we shall use properties of hyperbolic space that hold more generally in CAT(0) metric spaces and/or Gromov hyperbolic metric spaces. For the reader interested to learn more about this point of view, I recommend [BH99]. Other excellent references include [BBI01; CDP90; GH90].

11.1 Metric properties of hyperbolic space

11.1.1 Basic properties

Throughout this chapter, let $(X, d) := \mathbb{H}^n$ denote the metric space that is n-dimensional hyperbolic space $\mathbb{H}^n$ with its distance function. (We take $n \geqslant 2$, although $n = 1$ is also acceptable.) We can alternatively use any of the models of hyperbolic space, since they are all isometric.

Let us point out that the notion of geodesic makes sense in a metric space: it is defined as map $\gamma\colon I \to X$, where $I \subseteq \mathbb{R}$ is an interval, such that for any sufficiently close $t_0, t_1 \in I$, $d(\gamma(t_0), \gamma(t_1)) = v|t_1 - t_0|$ for some constant $v > 0$ (the speed of the geodesic). When (X, d) is a manifold with the distance induced from a Riemannian metric, geodesics in (X, d) coincides with Riemannian geodesics (it is a fundamental theorem of Riemannian geometry that geodesics can be characterized as locally length-minimizing curves.)

Proposition 11.1. *Hyperbolic space $(X, d) = \mathbb{H}^n$ enjoys the following properties:*

(i) It is a complete metric space.

(ii) It is a proper metric space: any closed ball is compact.

(iii) For any two distinct points $x, y \in X$, there exists a unique geodesic γ from x to y up to reparametrization, moreover $d(x, y) = L(\gamma)$ (length of γ).

Remark 11.2. Property (iii) is sometimes called ***strong geodesic convexity***. It implies that (X, d) is ***uniquely geodesic***, which is the slightly weaker version: for any two distinct points $x, y \in X$, there exists a unique geodesic γ from x to y up to reparametrization such that $d(x, y) = L(\gamma)$. This implies in turn that X is a ***length space***: the distance between any two points is equal to the infimum of the lengths of rectifiable curves between them. Note that by definition, the Riemannian distance makes any Riemannian manifold a length space.

Proof. For (i), we use the famous Hopf-Rinow theorem of Riemannian geometry: a Riemannian manifold is complete as a metric space if and only if it is geodesically complete, i.e. all geodesics are defined on $\mathbb{R}$. We have seen that hyperbolic geodesics are defined on $\mathbb{R}$ in § 5.4 (see Corollary 5.10).

One way to prove (ii) is the following: let $B = \{x \in X \mid d(x, x_0) \leqslant r\}$ be a closed ball and consider the Riemannian exponential map $\exp_{x_0}\colon \mathrm{T}_{x_0} X \to X$. By geodesic completeness, $\exp_{x_0}$ is globally well-defined on $\mathrm{T}_{x_0} X$. It follows immediately from (iii) and the definition of the Riemannian exponential that $B = \exp(B_{\mathrm{E}})$ where $B_{\mathrm{E}} = \{v \in \mathrm{T}_{x_0} X \mid \|v\| \leqslant r\}$. Of course, B_{E} is compact as a closed bounded set in a Euclidean space, therefore $B = \exp(B_{\mathrm{E}})$ is compact by continuity of $\exp_{x_0}$. Let us mention that a more intrinsic proof consists in arguing that any complete and locally compact length space is proper: see [BH99, Cor. 3.8 in Chap. I.3].

We have already proved (iii) in the hyperboloid model: see Corollary 5.11. ∎

11.1.2 Convexity of the distance function

Consider the distance function on X: it is a map

$$d\colon X \times X \to [0, +\infty)\,.$$

It is a general feature of CAT(0) metric spaces that the distance function is convex on $X \times X$. We shall not discuss CAT(0) metric spaces in general, because we are essentially interested in this particular property. Let us only mention that by definition, a CAT(0) metric space [1] is a space where geodesic triangles are thinner than Euclidean triangles with the same side lengths: see Figure 11.1. Any Hadamard manifold (complete, simply connected, with nonpositive sectional curvature) is a CAT(0) metric space. For a precise definition and a systematic treatment of CAT(k) spaces, we refer to [BH99].

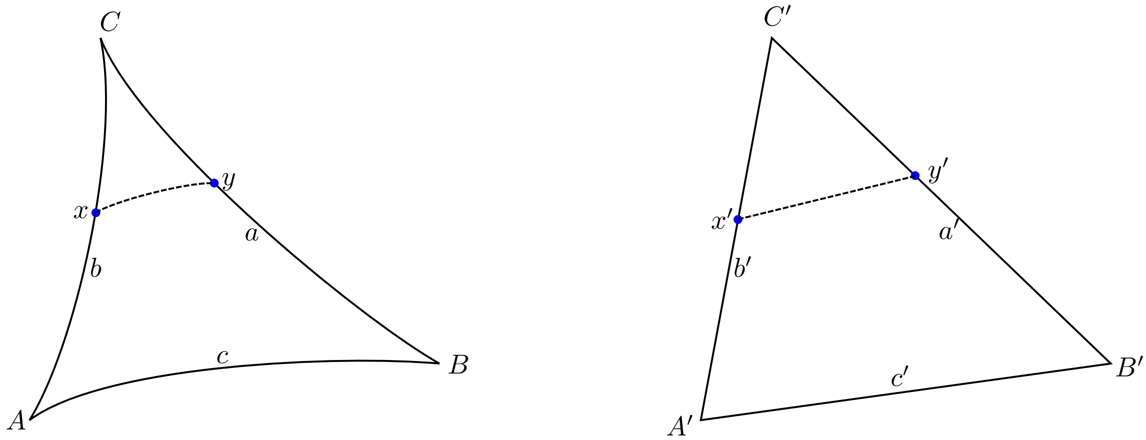


Figure 11.1: In a CAT(0) metric space, geodesic triangles are "slimmer" than Euclidean triangles with the same side lengths: in this schematic picture, we have $d_X(x, y) \leqslant d_{\mathbb{R}^2}(x', y')$.

The fact that the distance function is convex translates concretely as follows:

Theorem 11.3. *Let γ_1 and γ_2 be any two geodesics in $X = \mathbb{H}^n$, not necessarily with same speed. The function $t \mapsto d(\gamma_1(t), \gamma_2(t))$ is convex on $\mathbb{R}$.*

We give a direct proof below of Theorem 11.3, using the explicit expression of the distance function in the hyperboloid model. (Another direct proof can be found in [Thu97, Theorem 2.5.8].) Let us nevertheless give a sketch of what a more intrinsic proof would look like. First of all, it is very straightforward to show that the distance function is convex in any CAT(0) metric space: see [BH99, Prop. 2.2 in Chap II.2][2]. Secondly, one can show [BH99, Ex 1.9d in Chap. II.1] that the CAT(0) condition is equivalent to the property that, for any geodesic triangle with side lengths a, b, c and opposite angles α, β, γ, we have:

$$c^2 \geqslant a^2 + b^2 - 2ab \cos \gamma \,.$$

(Note that the equality case is the law of cosines in Euclidean geometry.) When $X = \mathbb{H}^n$ is hyperbolic space, this inequality can be derived (see e.g. [Duc18, Prop. 3.4]) from the

[1]Quoting [BH99]: The terminology "CAT(k)" was coined by M. Gromov [Gro87, p. 119]. The initials are in honour of E. Cartan, A.D. Alexandrov and V.A. Toponogov, each of whom considered similar conditions in varying degrees of generality.

[2]It is incorrectly assumed in [BH99] that the two geodesics have same (unit) speed, but the proof works without any changes for arbitrary geodesics.

hyperbolic law of cosines (Theorem 14.8). Of course, writing the details of this proof involves significantly more work than our direct proof below, but this proof can be extended to show the much more general fact that any Riemannian manifold of sectional curvature $\leqslant k$ is locally CAT(k). This was originally proved by Cartan in 1928 [Car88] for $k = 0$ and Alexandrov [Ale51] in the general case. We refer to [BH99, Chap. II.1 Appendix] for details.

Proof of Theorem 11.3. We work in the hyperboloid model. We know (Theorem 5.8) that the geodesics γ_i ($i \in \{1,2\}$) are of the form:

$$\gamma_i(t) = \cosh(\|v_i\|t)p_i + \sinh(\|v_i\|t)\frac{v_i}{\|v_i\|}$$

where p_i is a point on the hyperboloid, i.e. $p_i \in \mathbb{R}^{n,1}$ with $\langle p_i, p_i\rangle = -1$, and v_i is a tangent vector to the hyperboloid at p_i, i.e. $p_i \in \mathbb{R}^{n,1}$ with $\langle v_i, p_i\rangle = 0$.

The distance between $\gamma_1(t)$ and $\gamma_2(t)$ is given by:

$$d(t) = \operatorname{arcosh}\left(-\langle\gamma_1(t), \gamma_2(t)\rangle\right) .$$

It is straightforward to compute $d(t)$ and see that it is a $\mathcal{C}^\infty$ function of t, except possibly at $t = 0$ if $p_1 = p_2$. If $p_1 \neq p_2$, we can show that d is convex by proving that $d''(t) \geqslant 0$ for all t. It is sufficient to show that $d''(0) \geqslant 0$, since the other cases are obtained by reparametrizing the geodesics. As for the case $p_1 = p_2$, one can easily argue convexity by passing to the limit in the convexity inequality when $p_2 \longrightarrow p_1$. In summary, we can assume $p_1 \neq p_2$ and we want to show that $d''(0) \geqslant 0$.

By direct computation, one finds:

$$d''(0) = \sqrt{1+c^2}\,\frac{A-B}{c}$$

where

$$\begin{aligned} A &= \|v_1\|^2 + \|v_2\|^2 - 2\frac{\langle v_1, v_2\rangle}{\sqrt{1+c^2}} \\ B &= \frac{(\langle p_1, v_2\rangle + \langle v_1, p_2\rangle)^2}{c^2} \end{aligned}$$

and we have denoted $c^2 = \langle p_1, p_2\rangle^2 - 1$. (Note: these computations are guided by the fact that when $c \longrightarrow 0$, we approach the Euclidean scenario.) Thus it remains to show that $A \geqslant B$. Let us introduce the vectors:

$$\begin{aligned} u &= \frac{1}{c}\left(\langle p_1, p_2\rangle p_1 + p_2\right) \\ w_2 &= \frac{\langle p_1, v_2\rangle}{-1 - \langle p_1, p_2\rangle}(p_1 - p_2) - v_2 \,. \end{aligned}$$

It is immediate to check that $\langle u, p_1\rangle = \langle w_2, p_1\rangle = 0$, therefore these are two tangent vectors to the hyperboloid at p_1. Moreover, $\|u\| = 1$ and $\|w_2\| = \|v_2\|$. (Note: u is the initial velocity of the unit geodesic from p_1 to p_2, and w_2 is the inverse parallel transport of v_2 along that geodesic). It is straightforward to check that $B = \langle u, v_1 - w_2 \rangle^2$, and we leave it as an exercise to show that $A \geqslant \|v_1 - w_2\|^2$ (with equality if and only if the vectors v_1, w_2, and u are collinear). We conclude that $A \geqslant B$ by the Cauchy-Schwarz inequality (in the tangent space $\mathrm{T}_{p_1}\mathcal{H}^+$, which is positive definite). ∎

Tracing the equality case in the proof above, we can improve the previous theorem:

Theorem 11.4. *Given any two geodesics γ_1 and γ_2 in $X = \mathbb{H}^n$ (not necessarily with same speed), the function $t \mapsto d(\gamma_1(t), \gamma_2(t))$ is convex on $\mathbb{R}$. Moreover, it is strictly convex unless γ_1 and γ_2 are the same unoriented geodesic up to reparametrization.*

Proof. In the proof of Theorem 11.3, we see that $d''(0) > 0$ unless $A = \|v_1 - w_2\|^2$, which occurs if and only if the vectors v_1, w_2, and u are collinear. Since u is the initial tangent vector of the unit geodesic γ from p_1 to p_2, the fact that v_1 is parallel to u means that $\gamma_1 = \gamma$ up to reparametrization. On the other hand, since w_2 is the inverse transport of v_2 along γ, and u is the inverse parallel transport of the tangent vector u_2 to γ at p_2, the fact that w_2 is parallel to u implies that v_2 is parallel to u_2. This means that $\gamma_2 = \gamma$ up to reparametrization. We conclude that $\gamma_1 = \gamma = \gamma_2$ up to reparametrization. ∎

Note that if γ_1 and γ_2 are two parametrizations of the same unoriented geodesics, one can write $\gamma_1(t) = \gamma_2(at+b)$, with $a \in \mathbb{R} - \{0\}$ and $b \in \mathbb{R}$. We then have $d(\gamma_1(t), \gamma_2(t)) = |(at+b) - t|$. This is a (piecewise) linear function of t, and it is constant if and only if $a = 1$. The case $a = 1$ means that γ_1 and γ_2 have same orientation and same speed, equivalently $\gamma_1(t) = \gamma_2(t - t_0)$ for some $t_0 \in \mathbb{R}$.

Corollary 11.5. *Let γ_1 and γ_2 be two complete geodesics in $X = \mathbb{H}^n$ such that $d(\gamma_1(t), \gamma_2(t))$ is bounded. Then $\gamma_1 = \gamma_2$ up to a reparametrization $t \mapsto t - t_0$.*

Proof. By Theorem 11.4, the function $t \mapsto d(\gamma_1(t), \gamma_2(t))$ is strictly convex unless γ_1 and γ_2 are the same unoriented geodesic up to reparametrization. Since a strictly convex function on $\mathbb{R}$ cannot be bounded, we conclude that γ_1 and γ_2 are the same unoriented geodesic up to reparametrization. Moreover, the discussion above shows that we must be in the case $a = 1$, otherwise $d(\gamma_1(t), \gamma_2(t))$ is again not bounded. ∎

Remark 11.6. Theorem 11.4 and Corollary 11.5 reflect the fact that $X = \mathbb{H}^n$ has negative curvature (bounded away from zero), in contrast to any CAT(0) space: for instance, two distinct parallel lines in the Euclidean plane furnish a counter-example to Corollary 11.5.

It is not hard to extend Theorem 11.3 to the case where one of the geodesics has zero speed, i.e. is a constant curve, although technically this is not called a geodesic.

Corollary 11.7. *For any fixed $y \in X = \mathbb{H}^n$, the function $x \mapsto d(x, y)$ is convex on X. In other words, the function $t \mapsto d(\gamma(t), y)$ is convex on $\mathbb{R}$ for any geodesic γ. Moreover, it is strictly convex unless γ goes through y.*

Proof. Let v be any tangent vector at y. For any $\varepsilon > 0$, the function $t \mapsto d(\gamma(t), \gamma_{\varepsilon v}(t))$ is convex by Theorem 11.3. By passing to the limit in the convexity inequality when $\varepsilon \longrightarrow 0$, we obtain that $t \mapsto d(\gamma(t), y)$ is also convex.

Alternatively, we could write a direct proof from scratch using the explicit expression of $d(\gamma(t), y)$ in the hyperboloid model. The proof is then a simpler version of the proof of Theorem 11.3. It is also the best way to argue strict convexity. We leave out the details as an exercise. ■

11.1.3 Gromov hyperbolicity

Let (X, d) be a geodesic metric space (there exists a length-minimizing geodesic between any two points). Consider a geodesic triangle, which consists of three vertices and three sides, i.e. length-minimizing geodesics between the vertices. Such a triangle is called ***δ-slim*** (where $\delta \geqslant 0$) if any side is contained in the δ-neighborhood of the union of the other two sides. See Figure 11.2.

Definition 11.8. A geodesic metric space (X, d) is called ***Gromov hyperbolic*** if there exists $\delta \geqslant 0$ such that any geodesic triangle is δ-slim.

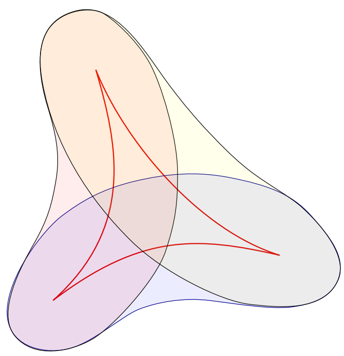


Figure 11.2: A δ-slim triangle: any of its side is contained in the δ-neighborhood of the union of the other two sides.

Example 11.9. Any geodesic metric space of bounded diameter is Gromov hyperbolic. The Euclidean plane is not Gromov hyperbolic: bigger and bigger triangles of the same aspect ratio require larger and larger δ's.

Remark 11.10. Contrary to the CAT(0) property or the convexity of the distance function, Gromov hyperbolicity only reflects negative curvature on a large scale, as opposed to an infinitesimal or local scale. One says that Gromov hyperbolicity is a ***coarse*** property.

Theorem 11.11. *The hyperbolic space $X = \mathbb{H}^n$ is Gromov hyperbolic.*

Proof. See Exercise 14.6. ■

11.2 The ideal boundary

11.2.1 Visual boundary and ideal boundary

Let us call ***geodesic ray*** in $X = \mathbb{H}^n$ a unit geodesic defined on an interval of the form $[t_0, +\infty)$.

Definition 11.12. Two geodesic rays r_1 and r_2 are called ***asymptotic*** when the distance $d(r_1(t), r_2(t))$ is bounded when $t \to +\infty$.

Remark 11.13. The ***Hausdorff distance*** between two subsets $A, B \subseteq X$ is the infimum of all $\delta > 0$ such that A is contained in the δ-neighborhood of B and conversely. (This is not a proper distance, because it can be infinite and it is equal to zero whenever A and B have same closure.) It is easy to show that two geodesic rays are asymptotic if and only if they have finite Hausdorff distance.

Being asymptotic defines an equivalence relation $\sim$ on the set of all geodesic rays. Let us denote $r(+\infty)$ the equivalence class of a geodesic ray r. If γ is a complete geodesic, we also let $\gamma(+\infty)$ denote the equivalence class of the ray $t \in [0, +\infty) \mapsto \gamma(t)$, and $\gamma(-\infty)$ the equivalence class of the ray $t \in [0, +\infty) \mapsto \gamma(-t)$.

Definition 11.14. The ***ideal boundary*** (or ***Gromov boundary***, or ***boundary at infinity***) of $X = \mathbb{H}^n$ is the set of all equivalence classes of geodesic rays, denoted $\partial_\infty X$.

The Gromov boundary can be defined for any metric space, and enjoys some good properties when X is Gromov hyperbolic. On the other hand, we have the notion of visual boundary ("boundary at infinity in the vision of an observer"), which is best suited to CAT(0) spaces:

Definition 11.15. Let $x_0 \in X = \mathbb{H}^n$. The ***visual boundary*** $\partial_\infty^{x_0} X$ is the set of all equivalence classes of geodesic rays starting from x_0.

Given our definition of the ideal boundary and the visual boundary, it is clear that $\partial_\infty^{x_0} X$ is a subset of $\partial_\infty X$. In a general metric space, the two can be different, but in our case of interest $X = \mathbb{H}^n$ they are the same.

Lemma 11.16. *Let $x_0 \in X = \mathbb{H}^n$. Any geodesic ray r in X is asymptotic to a unique geodesic ray $\hat{r}$ starting from x_0.*

Proof. Let us first show uniqueness: assume that r_1 and r_2 are two geodesic rays defined on $[0, +\infty)$ with $r_1(0) = r_2(0) = x_0$, and are asymptotic. By Theorem 11.3, the function $f\colon t \in [0, +\infty) \mapsto d(r_1(t), r_2(t))$ is convex. Moreover, f is nonnegative, and by assumption f is bounded and $f(0) = 0$. Such a function must be constant equal to zero. This shows that $r_1 = r_2$.

Let us now show existence. Let r_n be the geodesic ray starting from x_0 that goes through $r(n)$. Since X is proper, any closed ball $B(x_0, R)$ is compact, and we can apply the Arzèla-Ascoli theorem to find a subsequence of r_n that converges uniformly on such balls to some limit $\hat{r}$. It is easy to argue that $\hat{r}$ is also a geodesic ray. It remains to show that $\hat{r}$ is asymptotic to r. Consider the geodesic triangle with vertices x_0, $r(0)$, and $r(n)$. The fact that it is δ-slim implies that the side $[x_0, r(n)]$ is contained in the δ'-neighborhood of the side $[r(0), r(n)]$ where $\delta' = \delta + d(x_0, r(0))$, and conversely. In other words, the segments $[x_0, r(n)]$ and $[r(0), r(n)]$ are within Hausdorff distance $\leqslant \delta'$. Passing to the limit when $n \to +\infty$, we obtain that the geodesic rays r and $\hat{r}$ are within Hausdorff distance $\leqslant \delta'$, therefore they are asymptotic. ■

Remark 11.17. The uniqueness part of the proof works in any $\mathrm{CAT}(0)$ metric space. The existence part works in any proper Gromov hyperbolic space, but a different argument exists for complete $\mathrm{CAT}(0)$ metric spaces: see [BH99, Prop. 8.2].

Theorem 11.18. *For any $x_0 \in \mathbb{H}^n$, we have an identification $\partial_\infty X \approx \partial_\infty^{x_0} X$. Moreover, $\partial_\infty^{x_0} X$ can be identified to the unit tangent space $\mathrm{T}^1_{x_0} X := \{u \in \mathrm{T}_{x_0} X \mid \|u\| = 1\}$.*

Proof. It is clear that $\partial_\infty^{x_0} X$ is a subset of $\partial_\infty X$. In order to show that they are the same, we need to show that the map $\partial_\infty^{x_0} X \to \partial_\infty X$ is surjective, which is to say that any geodesic ray r starting from some point $x \in X$ is asymptotic to some ray $\hat{r}$ starting from x_0. This follows from the existence part of Lemma 11.16.

For the second assertion, first observe that the uniqueness part of Lemma 11.16 says that the equivalence relation on geodesic rays starting from x_0 is trivial, in other words there is a unique geodesic ray representing each element of $\partial_\infty^{x_0} X$. Such a geodesic ray is uniquely determined by its initial tangent vector $u \in \mathrm{T}^1_{x_0} X$. ■

11.2.2 Topology

Let $X = \mathbb{H}^n$ and let us denote $\bar{X}^\infty := X \sqcup \partial_\infty X$. There is a natural topology on $\bar{X}^\infty$ such that, for any geodesic ray r in X, $r(t) \to r(+\infty)$ when $t \to +\infty$. There are various ways to define this topology, here is one of them. Fix x_0 in X. For any $x \in \bar{X}^\infty$, there is a unique geodesic segment (when $x \in X$) or ray (when $x \in \partial_\infty X$), which we denote r_x, from x_0 to x. By definition, we say that $x_n \to x$ in $\bar{X}^\infty$ when $r_{x_n} \to r_x$ locally uniformly. We leave as an exercise to the reader to show that this is a well-defined topology on $\bar{X}^\infty$ and that it does not depend on the choice of x_0.

Theorem 11.19. *Let $X = \mathbb{H}^n$ and consider $\bar{X}^\infty = X \sqcup \partial_\infty X$ with the topology defined above.*

(i) The identifications $\partial_\infty X \approx \partial_\infty^{x_0} X \approx \mathrm{T}^1_{x_0} X$ are homeomorphisms. In particular, $\partial_\infty X$ is a topological $(n-1)$-sphere.

(ii) The inclusion $X \to \bar{X}^\infty$ is a compactification of X: it is a homeomorphism to its image, which is dense, and $\bar{X}^\infty$ is compact. Topologically, $\bar{X}^\infty$ is a closed n-ball.

We leave the proof of Theorem 11.19 as an exercise for the most diligent readers.

Remark 11.20. There are various ways to compactify a topological space, the simplest being the one-point compactification. However, depending on the context, one may seek compactifications where the points at infinity retain some interesting information, so that the compactified space is insightful. The compactification of hyperbolic space (or more generally, a CAT(0) or a Gromov hyperbolic metric space) is an example of compactification that is geometrically meaningful. Other important examples include: the end compactification of a topological space, the Stone-Čech compactification of a topological space, and the projective compactification of a vector space. We have seen the latter in Chapter 7: embedding $\mathbb{K}^n$ as an affine hyperplane in $\mathbb{K}P^n$ is indeed a compactification (where $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$).

11.2.3 Essential properties

Let us put on the record a couple of essential properties of the ideal boundary, in addition to Theorem 11.18 and Theorem 11.19.

Theorem 11.21. *Let $X = \mathbb{H}^n$. For any two distinct points $x, y \in \bar{X}^\infty$, there exists a unique geodesic from x to y.*

Proof. When x and y are both in X, we already know that there exists a unique geodesic from x to y (see Proposition 11.1 (iii)). When $x \in X$ and $y \in \partial_\infty X$, the existence and uniqueness of a geodesic ray r starting from x such that $r(+\infty) = y$ is the content of Lemma 11.16. Finally, when x and y are both ideal points, the proof of the existence and uniqueness of a geodesic such that $\gamma(-\infty) = x$ and $\gamma(+\infty) = y$ can be conducted similarly to the proof of Lemma 11.16; we leave out the details. ∎

Considering the case where x and y are both ideal points, we immediately get:

Corollary 11.22. *Any complete geodesic $\gamma\colon \mathbb{R} \to X$ is uniquely determined by its pair of ideal points $\{\gamma(-\infty), \gamma(+\infty)\}$.*

The next theorem will be important in the next chapter:

Theorem 11.23. *Let $X = \mathbb{H}^n$. Any isometry $f\colon X \to X$ uniquely extends to a continuous map $\hat{f}\colon \bar{X}^\infty \to \bar{X}^\infty$, and the restriction of $\hat{f}$ to $\partial_\infty X$ is a homeomorphism $\partial_\infty X \to \partial_\infty X$.*

Proof. It is a straightforward exercise to check that the map $\hat{f}$ defined on $\partial_\infty X$ by $\hat{f}(r(+\infty)) := (f \circ r)(+\infty)$ is well-defined and extends f continuously. Moreover, $\widehat{f^{-1}} = \hat{f}^{-1}$, therefore $\hat{f}$ is a homeomorphism of $\bar{X}^\infty$, and it restricts to a homeomorphism of $\partial_\infty X$. ∎

Remark 11.24. As we shall see below, in the Poincaré ball model $X = B^n$, the ideal boundary is $\partial_\infty X = \partial B^n = S^{n-1}$. We already know from Theorem 10.9 that any isometry $f\colon X \to X$ uniquely extends to $\partial B^n = S^{n-1}$. This provides an alternative proof of Theorem 11.23. This proof is much more specific to $X = \mathbb{H}^n$ (as opposed to X being any Gromov hyperbolic metric space), but it also gives more information: Theorem 10.9 additionally tells us that the boundary map is a Möbius transformation of S^{n-1}, and uniquely determines f.

11.3 The ideal boundary in each model

One way to describe the ideal boundary of hyperbolic space in each of the different models is to choose our favorite base point in the model, and associate a natural "point at infinity" to each geodesic ray from that point, thus providing an identification of the visual boundary.

11.3.1 Ideal boundary of the hyperboloid model

Choose the base point $p_0 = (0, \ldots, 0, 1) \in \mathcal{H}^+$. Any geodesic ray starting from p_0 is of the form $r(t) = \cosh(t)p_0 + \sinh(t)v$, where v is a unit tangent vector at p_0. When $t \to +\infty$, $r(t) \sim e^t u$ where $u = p_0 + v$ is a lightlike vector. Thus the geodesic ray $r(t)$ is asymptotic to the lightlike line $l = \mathbb{R}u$. Conversely, any lightlike line l can be written $l = \mathbb{R}u$ where u is a lightlike vector of the form $u = (v_0, 1)$. Letting $v = (v_0, 0)$, we have that the geodesic ray $r(t)$ as above is asymptotic to l. In conclusion:

Theorem 11.25. *The ideal boundary of the hyperboloid model $\mathcal{H}^+ \subseteq \mathbb{R}^{n,1}$ may be identified to the set of lightlike lines in $\mathbb{R}^{n,1}$.*

Note that the set of lightlike lines in $\mathbb{R}^{n,1}$ is called the projectivized light cone, which we have encountered several times in this course. As a projective quadric, it is called an ellipsoid, and it is a topological sphere as expected.

11.3.2 Ideal boundary of the Klein model

We recall that there are two variations of the Klein model: the Cayley-Klein model, which is a projective model, and the Beltrami-Klein model, which is the Cayley-Klein model projected in an affine chart.

The Cayley-Klein model is the interior Ω^- of an ellipsoid $\mathcal{Q}$ in projective space $\mathcal{P} = \mathbb{R}P^n$, and geodesics are projective lines (or rather chords, i.e. projective lines restricted to Ω^-). It is clear that given any base point $x_0 \in \Omega^-$, each geodesic ray starting from x is uniquely determined by its intersection with $\mathcal{Q}$. In conclusion:

Theorem 11.26. *The ideal boundary of the Cayley-Klein model $\Omega^- \subseteq \mathcal{P}$ is the ellipsoid $\mathcal{Q}$.*

Remark 11.27. The ellipsoid $\mathcal{Q}$ is none other than the projectivized light cone of $\mathbb{R}^{n,1}$. The hyperboloid $\mathcal{H}^+ \subseteq \mathbb{R}^{n,1}$ and the Cayley-Klein model $\Omega^- \subseteq \mathbf{P}(\mathbb{R}^{n,1})$ thus have the same ideal boundary. Can you explain this "coincidence"? See Exercise 11.2.

Let us now turn to the Beltrami-Klein model. This is the unit ball $B^n \subseteq \mathbb{R}^n$ equipped with a Riemannian metric such that the geodesics in B^n are the chords (intersection of B^n with Euclidean straight lines in $\mathbb{R}^n$). Taking $x_0 = 0$ (the Euclidean center of B^n), a geodesic ray starting from x_0 is a Euclidean radius of B^n. Clearly, each such ray is uniquely determined by its intersection with $\partial B^n = S^{n-1}$. See Figure 11.3. In conclusion:

Theorem 11.28. *The ideal boundary of the Beltrami-Klein ball $B^n \subseteq \mathbb{R}^n$ is the sphere $\partial B^n = S^{n-1}$.*

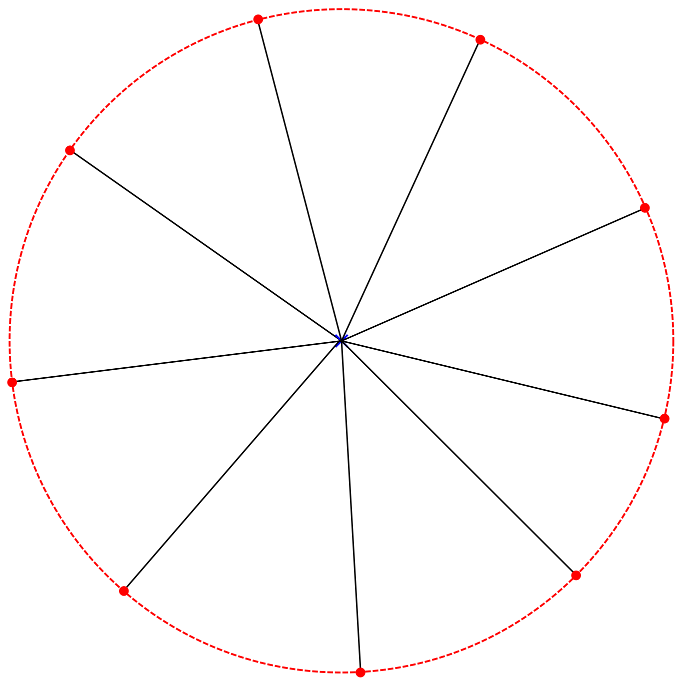


Figure 11.3: Visual boundary of the Beltrami-Klein disk (or the Poincaré disk) seen from the origin.

11.3.3 Ideal boundary of the Poincaré models

The Poincaré ball is the unit ball $B^n \subseteq \mathbb{R}^n$ equipped with a Riemannian metric such that the geodesics in B^n are arcs of Euclidean circles orthogonal to the boundary $\partial B^n = S^{n-1}$, and Euclidean diameters of B^n. Taking $x_0 = 0$ (the Euclidean center of B^n), a geodesic ray starting from x_0 is a Euclidean radius of B^n, just like in the Beltrami-Klein model (although the parametrization is different). Clearly, each such ray is uniquely determined by its intersection with $\partial B^n = S^{n-1}$ (again, see Figure 11.3). In conclusion:

Theorem 11.29. *The ideal boundary of the Poincaré ball $B^n \subseteq \mathbb{R}^n$ is $\partial B^n = S^{n-1}$.*

As for the Poincaré half-space $H^n \subseteq \mathbb{R}^n$, geodesics are Euclidean half-circles orthogonal to the boundary $\partial H^n = \widehat{\mathbb{R}^{n-1}}$. One can show again that each geodesic ray starting from some point $x_0 \in H^n$ is uniquely determined by its intersection with ∂H^n. One could either prove this directly,or derive it from the Poincaré ball case using the Cayley transform. In conclusion:

Theorem 11.30. *The ideal boundary of the Poincaré half-space $H^n \subseteq \mathbb{R}^n$ is $\partial H^n = \widehat{\mathbb{R}^{n-1}}$.*

11.4 Busemann functions and horospheres

11.4.1 Busemann functions

Let $X = \mathbb{H}^n$. For any geodesic ray $r\colon [0,+\infty) \longrightarrow X$, let us define the ***Busemann function*** (relative to r) as:

$$\begin{aligned} B_r\colon X &\longrightarrow \mathbb{R} \\ x &\longmapsto \lim_{t\to+\infty} \left(d(x,r(t)) - t\right) . \end{aligned}$$

Proposition 11.31. *For any geodesic ray r, the Busemann function $B_r\colon X \longrightarrow \mathbb{R}$ is well-defined, Lipschitz continuous with constant 1, and convex on X. Moreover B_{r_1} and B_{r_2} differ by an additive constant if and only if $r_1(+\infty) = r_2(+\infty)$.*

Proof. For any $x \in X$, the function $g\colon t \mapsto d(x,r(t)) - t$ is nonincreasing. Indeed, for $s \leqslant t$ we have $g(t) - g(s) = d(x,r(t)) - d(x,r(s)) - (t-s)$; by the triangle inequality $d(x,r(t)) - d(x,r(s)) \leqslant d(r(t),r(s)) = t-s$ so we obtain $g(t) - g(s) \leqslant 0$. Moreover, $g(t)$ is bounded below by $-d(x,r(0))$, since $t = d(r(0),r(t)) \leqslant d(r(0),x) + d(x,r(t))$. It follows that $g(t)$ converges when $t + \infty$ to some limit $B_r(x)$. By Dini's theorem, the convergence is locally uniform.

It follows from the triangle inequality that $|B_r(x) - B_r(y)| \leqslant d(x,y)$, i.e. B_r is Lipschitz continuous with constant 1. The convexity of B_r is immediately derived from the convexity of the distance function on $X = \mathbb{H}^n$ (Theorem 11.3).

If B_{r_1} and B_{r_2} differ by an additive constant, we may assume that $B_{r_1} = B_{r_2} =: B$ after reparametrizing r_1 or r_2. Let $t_0 \in [0,+\infty)$ and consider the closed convex set $C := \{B \leqslant -t_0\} \subseteq X$. Note that $B(r_1(t_0)) = -t_0$, therefore $r_1(t_0) \in C$. In fact, for $t \geqslant t_0$, $r_1(t_0)$ is the projection of $r_1(t)$ on C. Let us admit the previous point (see [BH99, Prop. 8.22 in Chap. II.8]) or leave it as an exercise. Similarly, $r_2(t_0)$ is the projection of $r_2(t)$ on $C$for $t \geqslant t_0$. It follows that $d(r_1(t),r_2(t)) \leqslant d(r_1(t_0),r_2(t_0))$ is bounded for $t \geqslant t_0$, hence r_1 and r_2 are asymptotic. Conversely, assume that r_1 and r_2 are asymptotic, and let us show that $B_{r_1} - B_{r_2}$ is constant. The function $t \mapsto d(r_1(t),r_2(t))$ is convex and bounded, therefore it has a finite limit when $t \longrightarrow +\infty$. After reparametrizing of r_1 or r_2, we can assume that $\lim_{t\to+\infty} d(r_1(t),r_2(t)) = 0$. By the triangle inequality, $|B_{r_1}(x) - B_{r_2}(x)| \leqslant \lim_{t\to+\infty} d(r_1(t),r_2(t))$, so we conclude that $B_{r_1} = B_{r_2}$. ■

Let now $\xi \in \partial_\infty X$ and let us define the ***Busemann function*** (relative to ξ) as:

$$\begin{aligned} B_\xi\colon X \times X &\longrightarrow \mathbb{R} \\ (x,y) &\longmapsto \lim_{t\to+\infty} \left(d(x,r(t)) - d(y,r(t))\right) \end{aligned}$$

where r is any geodesic ray with $r(+\infty) = \xi$. In other words, $B_\xi(x,y) = B_r(x) - B_r(y)$.

Proposition 11.32. *For any $\xi \in \partial_\infty X$, the Busemann function $B_\xi \colon X \times X \to \mathbb{R}$ is well-defined and continuous.*

Proof. As pointed out above, $B_\xi(x, y) = B_r(x) - B_r(y)$. It follows from the previous proposition that $B_\xi(x, y)$ is independent of the choice of geodesic ray r such that $r(+\infty) = \xi$. Moreover, B_ξ is clearly continuous since B_r is continuous. ∎

As an example, let us compute a Busemann function in the Poincaré half-space $X = H^n$.

Proposition 11.33. *In the Poincaré half-space $H^n \subseteq \mathbb{R}^n$, the Busemann function relative to the ideal point $\xi = \infty \in \partial H^n$ is:*

$$B_\xi(x, y) = \ln(y_n) - \ln(x_n) \ .$$

Proof. Let us choose a geodesic ray r in H^n such that $r(+\infty) = \xi$. Recall that geodesics having ∞ as an endpoint in the Poincaré half-space model are Euclidean vertical straight lines. We can take $r(t) = (0, \ldots, 0, e^t)$. Indeed, $r(t)$ parametrizes the vertical straight line from 0 to ∞, and it is immediate to check that $r'(t) = (0, \ldots, 0, e^t)$ has unit norm with respect to the Poincaré metric $\frac{\mathrm{d}x_1^2 + \ldots \mathrm{d}x_n^2}{x_n^2}$, hence $r(t)$ is a geodesic ray.

The distance from a point $x = (x_1, \ldots x_n)$ is given by (see (10.2)) $d(x, r(t)) = \operatorname{arcosh} A(t)$ where

$$\begin{aligned} A(t) &= 1 + \frac{x_1^2 + \cdots + x_{n-1}^2 + (x_n - e^t)^2}{2x_n e^t} \\ &= \frac{x_1^2 + \cdots + x_n^2 + e^{2t}}{2x_n e^t} = e^t a(t) \end{aligned}$$

with $a(t) = \frac{1}{2x_n}\left(1 + e^{-2t}\left(x_1^2 + \cdots + x_n^2\right)\right)$. Since $\operatorname{arcosh}(A(t)) = \ln\left(A(t) + \sqrt{A(t)^2 - 1}\right)$, when $t \to +\infty$ we have

$$\begin{aligned} d(x, r(t)) &\approx \ln\left(2A(t)\right) = t + \ln(2a(t)) \\ &\approx t + \ln\left(\frac{1}{x_n}\right). \end{aligned}$$

We conclude that $B_r(x) = -\ln x_n$, and $B_\xi(x, y) = B_r(x) - B_r(y) = \ln(y_n) - \ln(x_n)$. ∎

11.4.2 Horospheres

Definition 11.34. A ***horosphere*** in $X = \mathbb{H}^n$ is a level set of a Busemann function B_r for some geodesic ray r. When $n = 2$, a horosphere is also called ***horocycle***.

One says that a horosphere given by a level set of B_r is *centered at* $\xi := r(+\infty)$. The next proposition follows immediately from the discussion of the previous subsection:

Proposition 11.35. *For any $\xi \in \partial_\infty X$ and any $x_0 \in X$, there exists a unique horosphere centered at ξ going through x_0; it is the set $\{x \in X \mid B_\xi(x, x_0) = 0\}$.*

The next proposition is also an immediate consequence of the discussion of the previous subsection:

Proposition 11.36. *Let $\xi \in \partial_\infty X$. Any geodesic γ with $\gamma(+\infty) = \xi$ intersects each horosphere centered at ξ exactly once.*

Proof. Let $x_0 = \gamma(t_0)$. We know that there exists a horosphere S centered at ξ going through x_0. Let us show that x_0 is the only intersection of S and γ. Let $r(t) = \gamma(t)$ for $t \in [t_0, +\infty)$. Since r is a geodesic ray with endpoint ξ, horospheres centered at ξ are level sets of the Busemann function B_r. Note that for any $x = \gamma(t_1)$ on the geodesic, $d(x, r(t)) = |t - t_1| - t$, we easily derive that $B_r(x) = -t_1$. In particular, S is the $-t_0$ level set of B_r, and it does not go through $x = r(t_1)$ unless $t_1 = t_0$. ∎

Proposition 11.37. *Let f be an isometry of X, and still denote f its extension to $\partial_\infty X$. For any $\xi \in \partial_\infty X$, f maps bijectively horospheres centered at ξ to horospheres centered at $f(\xi)$.*

Proof. Let r be a geodesic ray with $r(+\infty) = \xi$, then $f \circ r$ is a geodesic ray with $f \circ r(+\infty) = f(\xi)$. The fact that f is an isometry implies that $B_{f \circ r} = B \circ f^{-}1$. It follows that $S \subseteq X$ is a level set of B_r if and only if $f(S)$ is a level set of $B_{f \circ r}$. In other words, S is a horosphere centered at ξ if and only if $f(S)$ is a horosphere centered at $f(\xi)$. ∎

Now let us describe horospheres in the Poincaré models.

Theorem 11.38. *In the Poincaré ball $X = (B^n, g_{B^n})$ or in the Poincaré half-space ball $X = (H^n, g_{H^n})$, the horospheres centered at any $\xi \in \partial_\infty X$ are the Euclidean hyperspheres of $\mathbb{R}^n$ contained in X that are tangent to $\partial_\infty X$ at ξ.*

Figure 11.4 and Figure 11.5 feature a few horocycles in the Poincaré disk and in the Poincaré half-plane respectively.

Remark 11.39. In the Poincaré half-space $X = H^n$, recall that $\partial H^n = \widehat{\mathbb{R}^{n-1}}$. In the case where $\xi = \infty$, Theorem 11.38 must be understood as: Horospheres centered at ξ are horizontal hyperplanes, i.e. subsets $\{x_n = c\}$ with $c > 0$. See Figure 11.6.

Proof of Theorem 11.38. First we argue that it is enough to show the theorem for one particular ideal point $\xi_0 \in \partial_\infty X$. Recall that if X is the Poincaré ball or the Poincaré half-space, then any isometry $f \in \mathrm{Isom}(X)$ is uniquely determined by its extension to $\partial_\infty X$, which we abusively still denote f, and which is a Möbius transformation of $\partial_\infty X$. Since the Möbius group acts transitively on $\partial_\infty X$, if $\xi \in \partial_\infty X$ is any other ideal point, we can find an isometry $f \in \mathrm{Isom}(X)$ such that $f(\xi_0) = \xi$. By Proposition 11.37, f maps horospheres centered at ξ_0 to horospheres centered at ξ. On the other hand, f is a Möbius transformation of X, therefore it is sphere-preserving (see Theorem 9.18), and it also preserves tangency to $\partial_\infty X$. In conclusion, it is

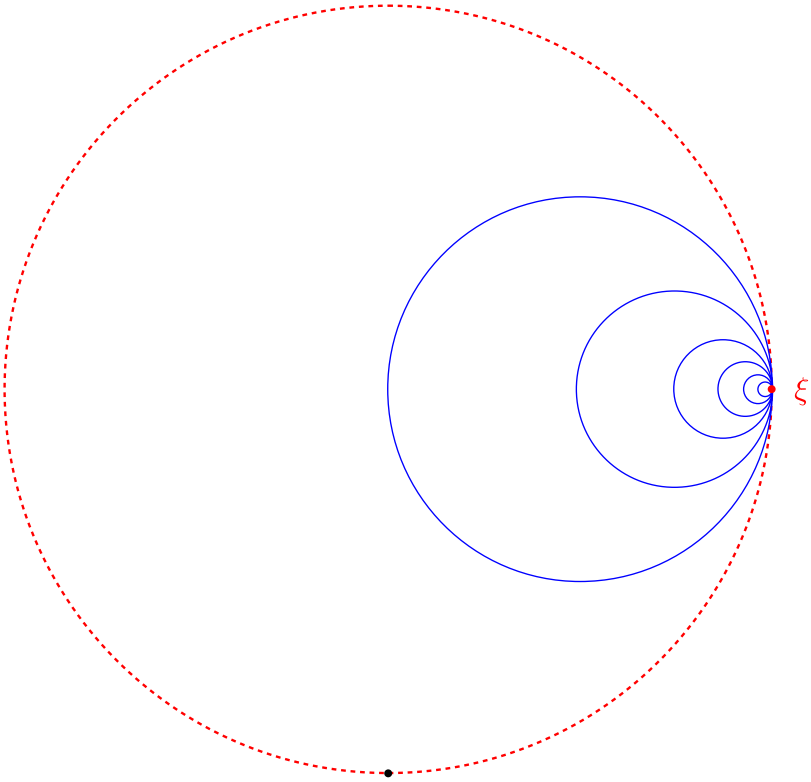


Figure 11.4: Horocycles in the Poincaré disk.

enough to show the theorem at ξ_0. Moreover it is enough to do the case $X = H^n$, because the case $X = B^n$ can then be derived using the Cayley transform.

Thus we take $X = H^n \subseteq \mathbb{R}^n$ and let us pick $\xi_0 = \infty \in \partial_\infty X$. In this case, the (generalized) Euclidean hyperspheres tangent to ξ_0 are the horizontal Euclidean hyperplanes in H^n. We want to show that such are the horospheres centered at ξ_0. For any $x \in H^n$, the horosphere through x is $S = \{y \in H^n \mid B_{\xi_0}(x, y) = 0\}$. By Proposition 11.33, we immediately find $S = \{y \in H^n \mid y_n = x_n\}$. In other words, S is the horizontal hyperplane through x. ■

We leave as an exercise (rather, several exercises) to the curious reader to describe horospheres in the other models of hyperbolic space. In Chapter 5, there was an exercise that claims to describe horocycles on the hyperboloid when $n = 2$: see Exercise 5.4. Exercise 11.6 proposes to prove an analogous result in any dimension. As for the Klein models, a characterization is suggested in Exercise 11.7.

Let us conclude this chapter with the following important property of horospheres:

Theorem 11.40. *Any horosphere $S \subseteq \mathbb{H}^n$ is a Euclidean space. In other words, any horosphere $S \subseteq \mathbb{H}^n$ is a complete simply-connected hypersurface with vanishing curvature. Equivalently, there exists an isometry $S \xrightarrow{\sim} \mathbb{R}^n$.*

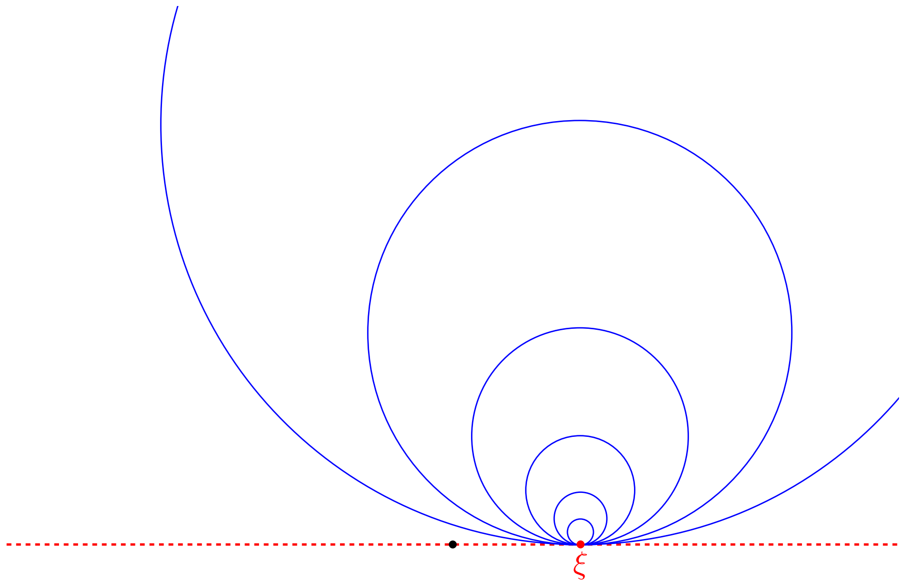


Figure 11.5: Horocycles in the Poincaré half-plane.

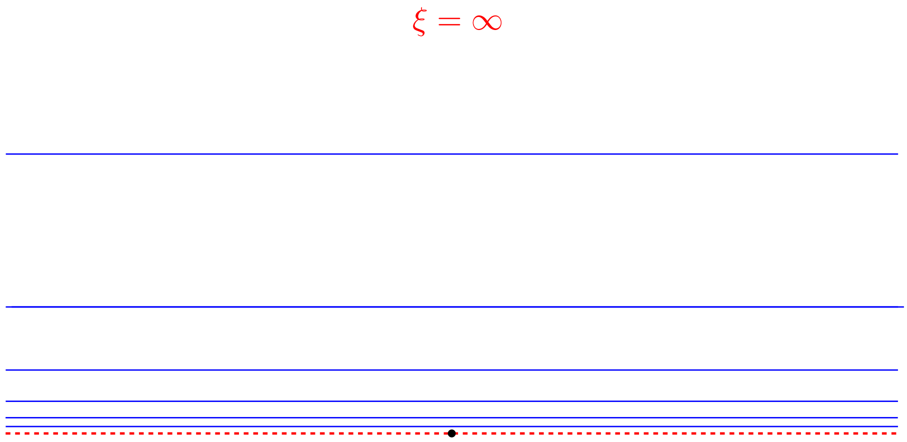


Figure 11.6: Horocycles centered at $\xi = \infty$ in the Poincaré half-plane.

Proof. Since all models of $\mathbb{H}^n$ are isometric, it is enough to do the proof in the Poincaré half-space model. Moreover, since horospheres at some ideal point ξ_0 are mapped isometrically to horospheres at all other ideal points (see proof of Theorem 11.38), it is enough to consider horospheres at ξ_0.

Let us $\xi_0 = \infty$. We have seen that horospheres at ξ_0 are horizontal hyperplanes contained in H^n. Consider such a horosphere $S = \{x \in \mathbb{H}^n \mid x_n = c\}$ (where $c > 0$ is a constant). Recall that the hyperbolic metric in H^n is:

$$g_{H^n} = \frac{dx_1^2 + \cdots + dx_{n-1}^2 + dx_n^2}{x_n^2} .$$

Clearly $(x_1, \ldots, x_{n-1})$ offer a global system of coordinates on S, and the induced metric on S is simply:

$$g_S = \frac{dx_1^2 + \cdots + dx_{n-1}^2}{c^2} .$$

Up to the constant scaling factor $\frac{1}{c^2}$, this is the standard Euclidean metric g_0 on $\mathbb{R}^{n-1}$. Regardless, this is a complete Euclidean metric (in fact, g_S is isometric to g_0 via $x \mapsto x/c$). ■

11.5 Exercises

Exercise 11.1.

(*) Quasi-isometric spaces

Let (X, d_X) and (Y, d_Y) be metric spaces. A map $f: X \to Y$ is called a ***quasi-isometry*** if:

(i) f is coarsely Lipschitz: there exists $A \geqslant 1, B \geqslant 0$ such that for all $x_1, x_2 \in X$:

$$\frac{1}{A} d_X(x_1, x_2) - B \leqslant d_Y(f(x_1), f(x_2)) \leqslant A d_X(x_1, x_2) + B \, .$$

(ii) f is coarsely surjective: there exists $C \geqslant 0$ such that for all $y \in Y$, there exists $x \in X$ such that $d(f(x), y) \leqslant C$.

When there exists a quasi-isometry $f: X \to Y$, one says that the metric spaces X and Y are ***quasi-isometric***.

(1) Show that any metric space of finite diameter is quasi-isometric to a point.

(2) Show that $\mathbb{R}^2$ and $\mathbb{H}^2$ are not quasi-isometric.

(3) Show that any quasi-isometry $f: \mathbb{H}^m \to \mathbb{H}^n$ extends to a homeomorphism $\partial_\infty \mathbb{H}^m \to \partial_\infty \mathbb{H}^n$. Conclude that $\mathbb{H}^m$ is quasi-isometric to $\mathbb{H}^n$ if and only if $m = n$.

Exercise 11.2.

Ideal boundary of the hyperboloid and the Cayley–Klein models

We identified both the ideal boundary of the hyperboloid model $\mathcal{H}^+ \subseteq \mathbb{R}^{n,1}$ and the ideal boundary of the Cayley–Klein model $\Omega^- \subseteq \mathbf{P}(\mathbb{R}^{n,1})$ as the projectivized light cone of $\mathbb{R}^{n,1}$. Can you explain this "coincidence"?

Exercise 11.3.

Busemann function in the Poincaré disk

Let $X = (B^2, g_{B^2})$ be the Poincaré disk. We use the complex coordinate z on the unit disk $\mathbb{D} \approx B^2$.

(1) For any $\xi \in \partial_\infty X = \{z \in \mathbb{C} \mid |z| = 1\}$, check that the geodesic ray $r_\xi: [0, +\infty) \to X$ such that $r(0) = 0$ and $r(+\infty) = \xi$ has the expression: $r(t) = \tanh(t/2)\, \xi$.

(2) Show that the Busemann function B_r is given by

$$B_r(z) = -\ln\left(\frac{1-|z|^2}{|z-\xi|^2}\right) .$$

(3) Recover the fact that horocycles centered at ξ are Euclidean circles tangent to $\partial_\infty X$ at ξ.

Exercise 11.4.

Horospheres as limit of spheres

Let $x_0 \in \mathbb{H}^n$ and let $P \subseteq \mathrm{T}_{x_0}\mathbb{H}^n$ be a hyperplane.

(1) Show that for all $r > 0$, there exists exactly two hyperspheres $S_1(r)$ and $S_2(r)$ in $\mathbb{H}^n$ that go through x_0 and are tangent to P.

(2) Show that there exists exactly two horospheres S_1 and S_2 in $\mathbb{H}^n$ that go through x_0 and are tangent to P.

(3) Show that $\{\lim_{r\to+\infty} S_1(r), \lim_{r\to+\infty} S_2(r)\} = \{S_1, S_2\}$.

Exercise 11.5.

Horospheres as hypersurfaces with asymptotic normal geodesics

(1) Let S be a horosphere centered at $\xi \in \partial_\infty\mathbb{H}^n$. Show that for any $x_0 \in S$, the geodesic going through x and with ideal endpoint ξ intersects S orthogonally. Show that it is also orthogonally transverse to any other horosphere centered at ξ.

(2) Show that a complete hypersurface $S \subseteq \mathbb{H}^n$ is a horosphere if and only if all geodesics that intersect S orthogonally share an ideal endpoint.

Exercise 11.6.

Horospheres in the hyperboloid model

Show that in the hyperboloid model $\mathcal{H}^+ \subseteq \mathbb{R}^{n,1}$, horospheres are given by the intersection of $\mathcal{H}^+$ with hyperplanes of $\mathbb{R}^{n,1}$ whose normal lies in the light cone. Show that when $n = 2$, these are parabolas (also see Exercise 5.4).

Exercise 11.7.

Horospheres in the Klein model

Show that in the Beltrami–Klein disk $B^2 \subseteq \mathbb{R}^2$, the horocycles centered at $\xi \in S^1$ are the Euclidean ellipses contained in B^2 that have a contact of order 4 with S^1 at ξ. Suggest and prove an analogous characterization in higher dimensions. Argue that this characterization also makes sense in the Cayley–Klein model.

Exercise 11.8.

Isometries fixing an ideal point

Let $X = \mathbb{H}^n$ and $\xi \in \partial_\infty X$.

(1) Show that if $f \in \mathrm{Isom}(X)$ fixes ξ, then f maps any horosphere S centered at ξ to some other such horosphere S'. *Optional: in what case do we have $S' = S$?*

(2) Recall that any horosphere S is isometric to $\mathbb{R}^{n-1}$. Recall explicitly the isometric identification $S \approx \mathbb{R}^{n-1}$ when S is a horosphere centered at $\xi = \infty$ in the Poincaré half-space model. Show that f induces an affine similarity of $\mathbb{R}^{n-1}$.

(3) Recover the fact that the subgroup of the Möbius group of S^{n-1} fixing a point is isomorphic to the group of affine similarities of $\mathbb{R}^{n-1}$ (see Exercise 9.6).

CHAPTER 12

Isometries of hyperbolic space

Disclaimer: This chapter is a draft.

In this chapter, we study the isometries of hyperbolic space and establish a classification thereof. We have already described the group of isometries in the different models (hyperboloid, Klein models, Poincaré models), and how it acts on each model; however we have yet to analyze the geometric behavior of isometries.

As an analogy, consider the group $\mathrm{E}^+(3) = \mathrm{Isom}^+(\mathbb{R}^3)$ of motions (orientation-preserving isometries) of Euclidean space. As a group, this is $\mathrm{E}^+(3) \approx \mathrm{SO}(3) \ltimes \mathbb{R}^3$, acting on $\mathbb{R}^3$ by affine transformations. But what do Euclidean isometries actually look like? As is well-known, they fall into distinct types: translations, rotations, and screw rotations (to include orientation-reversing isometries, there is also reflections, glide reflections, and rotation-reflections). This classification is easily generalized in any dimension.

The goal of this chapter is to present a similar classification of isometries of hyperbolic space $\mathbb{H}^n$. In order to do so, we will make a crucial use of the ideal boundary of hyperbolic space introduced in the previous chapter. Essentially, isometries can be classified according to their dynamics, which can be read off their extended action on the ideal boundary of hyperbolic space. Just like the notion of ideal boundary, this paradigm to classify isometries holds in a broad class of metric spaces. We attempt a presentation that is suggestive of this generality[1], but also discuss the specific features of the case of hyperbolic space.

After studying the isometries of hyperbolic space in arbitrary dimensions, we specialize to the 2- and 3-dimensional cases. We shall see that in the Poincaré half-space model, orientation-preserving isometries can be concretely described and characterized using matrices in $\mathrm{SL}(2, \mathbb{R})$

[1] To learn more about the classification of isometries in metric spaces of nonpositive curvature, I recommend [BH99] (for CAT(0) spaces) and [GH90] (for Gromov hyperbolic spaces).

(in the 2-dimensional case) or $\mathrm{SL}(2,\mathbb{C})$ (in the 3-dimensional case).

12.1 Classification

Let (X,d) be a metric space and let $f\colon X\to X$ be an isometry. By definition, the ***displacement function of*** f is $d_f(x) := d(x,f(x))$, and the ***translation length*** of f is $l_f := \inf_{x\in X} d_f(x)$.

Definition 12.1. An isometry $f\colon X\to X$ is called:

- ***elliptic*** if $l_f = 0$ is attained, i.e. f has a fixed point.
- ***hyperbolic*** (or ***loxodromic***) if $l_f > 0$ and is attained.
- ***parabolic*** if l_f is not attained.

Remark 12.2. A quick note about the terminology: for isometries of the second type, we will favor the term *hyperbolic* when X is a generic metric space, and *loxodromic* when $X=\mathbb{H}^n$ is hyperbolic space. There are two reasons to avoid using “hyperbolic” when $X=\mathbb{H}^n$: 1. Any isometry of $\mathbb{H}^n$ could reasonably be called a “hyperbolic isometry”, just like any isometry of $\mathbb{R}^n$ is called a Euclidean isometry, and 2. It is common in the math literature to call “hyperbolic isometry” a subclass of loxodromic isometries, although we find this a poor choice of terminology (we will use instead the term “translation”, see Definition 12.15).

Example 12.3. In Euclidean space $\mathbb{R}^n$, every isometry is either hyperbolic (translations, screw rotations, glide reflections) or elliptic (rotations, reflections, rotation-reflections). An isometry that is either elliptic or hyperbolic is called ***semisimple***, hence every Euclidean isometry is semisimple (i.e. there are no parabolics).

The main goal of this section is to present a characterization of elliptic, hyperbolic, and parabolic isometries of hyperbolic space, which we condense in the following three theorems. We postpone the definition of all the new terms appearing in these theorems (orbit, limit set, attracting/repelling/neutral fixed points, translation axis) until after their statement.

Theorem 12.4. *Let $X=\mathbb{H}^n$ and let $f\colon X\to X$ be an isometry. The following are equivalent:*

(i) f is elliptic.

(ii) Some/every orbit of f is bounded.

(iii) f has 0, 2, or infinitely many fixed points on $\partial_\infty X$, all of which are neutral fixed points.

Theorem 12.5. *Let $X=\mathbb{H}^n$ and let $f\colon X\to X$ be an isometry. The following are equivalent:*

(i) f is hyperbolic.

(ii) f has a translation axis.

(iii) *f has exactly two fixed points on $\xi^-, \xi^+ \in \partial_\infty X$, one attracting ($\xi^+$) and one repelling ($\xi^-$).*

Theorem 12.6. *Let $X = \mathbb{H}^n$ and let $f : X \to X$ be an isometry. The following are equivalent:*

(i) *f is parabolic.*

(ii) *f preserves some/any horosphere S centered at some point $\xi \in \partial_\infty X$, and has no fixed points in S.*

(iii) *f has exactly one fixed point $\xi \in \partial_\infty X$.*

Before proving these theorems let us define all the terms involved.

Orbits of f. By definition, an orbit of f is a subset of X of the form $\{f^n(x_0), n \in \mathbb{Z}\}$ for some $x_0 \in X$. Here we denote f^n the n-th iterate of f under composition, and f^{-n} is the inverse of f^n.

Fixed points of f at infinity. We have seen (Theorem 11.23) that any isometry $f : X \to X$ extends to the ideal boundary $\partial_\infty X$, and we still denote f the extension to the boundary. Therefore it makes sense to talk about fixed points of f on $\partial_\infty X$.

Attracting and repelling fixed points. A fixed point $\xi \in \partial_\infty f$ is called ***attracting*** if there exists a neighborhood U of ξ in $\partial_\infty X$ such that, for any neighborhood V of ξ, we have $f^n(U) \subseteq V$ for n sufficiently large. The fixed point ξ is called ***repelling*** if ξ is an attracting fixed points of f^{-1}. The fixed point ξ is called ***neutral*** if it is neither attracting nor repelling [2].

Remark 12.7. Assume that f has two fixed points ξ^+, ξ^-, with ξ^+ attracting and ξ^- repelling. If in the definition of attracting [resp. repelling] fixed point one may take for U any neighborhood of ξ^+ that avoids a neighborhood of ξ^- (resp. any neighborhood of ξ^- that avoids a neighborhood of ξ^+), one says that f ***has North-South dynamics*** on $\partial_\infty X$. We will see that any hyperbolic isometry of $X = \mathbb{H}^n$ has North-South dynamics on $\partial_\infty X$.

Translation axis. A geodesic in X is called a ***translation axis*** for an isometry X if f preserves the geodesic but does not fix it pointwise. Concretely, if $\gamma : \mathbb{R} \to X$ is such a geodesic, then there exists a real number $l \neq 0$ such that $f(\gamma(t)) = \gamma(t + l)$ for all $t \in \mathbb{R}$. We shall see that $|l| = l_f$ must be the translation length of f, and that f admits a translation axis if and only if it is a hyperbolic isometry. Moreover, the two fixed points $\xi^-, \xi^+ \in \partial_\infty X$ are the endpoints of its translation axis; in particular, the translation axis is unique by Theorem 11.21.

In order to prove Theorem 12.4, we shall use the notion of minimal bounding ball:

Definition 12.8. Let $A \subseteq X$ be a bounded set. A ***bounding ball*** for A is a closed ball $B \subseteq X$ containing A, and a ***minimal bounding ball*** is a bounding ball of minimal radius.

[2]There is a better definition of attracting, repelling, and neutral fixed points of f: these are respectively fixed points ξ where $f'(\xi)$ is < 1, > 1, or $= 1$. However defining the metric derivative $f'(\xi)$ requires more work, especially since we have not defined any metric on $\partial_\infty X$. To learn more on this, we refer to [GH90] or [DSU17]. Our definition of attracting and repelling fixed points is weaker in general, but equivalent in the case $X = \mathbb{H}^n$.

Lemma 12.9. *For any nonempty bounded subset $A \subseteq X = \mathbb{H}^n$, there exists a unique minimal bounding ball.*

Proof. Consider the function $R\colon X \to [0, +\infty)$ defined by $R(x) := \sup_{y \in A} d(x, y)^2$. Clearly, a minimum bounding ball is a closed ball whose center minimizes R. It is easy to see that R is a proper function on X, therefore it admits minimizers: this proves the existence of a minimum bounding ball.

Let us now prove uniqueness by arguing that R is a strictly convex function on X. Clearly, $R(x) = \sup_{y \in \bar{A}} d(x, y)^2$ is an equivalent definition of R, where $\bar{A}$ indicates the closure of A. By compactness of $\bar{A}$ (because X is a proper metric space: see Proposition 11.1), the supremum is attained in the definition of R. For any fixed $y \in X$, the function $x \mapsto d(x, y)^2$ is strictly convex on X: this can be proved by direct computation in the hyperboloid model; it is an easier version of Corollary 11.7. Therefore R is strictly convex on X as a maximum of strictly convex functions. Conclude by uniqueness of the minimizer of any strictly convex function. ∎

And another useful couple of lemmas, regarding fixed points and ideal fixed points of isometries of $X = \mathbb{H}^n$:

Lemma 12.10. *Let $X = \mathbb{H}^n$ and let $f\colon X \to X$ be an isometry. Then f has at least one fixed point or ideal fixed point.*

Proof. Recall that $\bar{X}^\infty = X \cup \partial_\infty X$ is a topological closed n-ball, as is illustrated by the Poincaré ball model for instance. The celebrated Brouwer fixed point theorem precisely says that any continuous map from a closed n-ball to itself has at least one fixed point. ∎

Lemma 12.11. *Let $X = \mathbb{H}^n$ and let $f\colon X \to X$ be an isometry. The fixed point set $F \subseteq X$ of f is either empty, or reduced to a point, or is a hyperbolic subspace of X. In other words, F is a subset of X that is stable under taking the complete geodesic through any two of its points.*

Proof. Assume F has at least two points, otherwise the lemma is vacuously true. Let x, y be two distinct points in F, and let $\gamma\colon \mathbb{R} \to X$ be the geodesic such that $\gamma(0) = x$ and $\gamma(1) = y$. Since f is an isometry, the curve $f \circ \gamma$ is also a geodesic in X. Moreover, $f \circ \gamma(0) = f(x) = x$ and $f \circ \gamma(1) = f(y) = y$. By uniqueness of the geodesic through x and y, we must have $f \circ \gamma = \gamma$. In other words, $\gamma(t) \in F$ for all $t \in \mathbb{R}$. ∎

Lemma 12.12. *Let $X = \mathbb{H}^n$ and let $f\colon X \to X$ be an isometry.*

(i) If $\xi_1, \xi_2 \in \partial_\infty X$ are two distinct ideal fixed points, then the geodesic with endpoints ξ_1 and ξ_2 is preserved by f.

(ii) If γ is any unit geodesic preserved by f, then $f(\gamma(t)) = \gamma(t + l)$ for all $t \in \mathbb{R}$, where $|l|$ is the translation length of f.

Proof. Since f is an isometry, $f \circ \gamma$ is a geodesic, moreover it has the same endpoints as γ therefore we have $f \circ \gamma$ parametrizes the same geodesic as γ by Theorem 11.21. Since $f \circ \gamma$ is a reparametrization of γ with same speed and orientation, we have $f(\gamma(t)) = \gamma(t+l)$ for some $l \in \mathbb{R}$. We must now show that $|l|$ is the translation length of f. Define the projection P to the geodesic parametrized by γ by putting that for any $x \in X$, $P(x)$ is the unique minimizer of the function $t \mapsto d(x, \gamma(t))^2$. Since this function is strictly convex, the map P is well-defined. Moreover, P is distance nonincreasing, we leave the proof of this claim as an exercise (there are several possible approaches, one may for instance compute the second derivative of P along any geodesic). It is straightforward to argue that $\pi(f(x)) = f(\pi(x))$, therefore we obtain $d(x, f(x)) \leqslant d(\pi(x), \pi(f(x)) = d(\pi(x), f(\pi(x))) = |l|$. This proves that l is the translation distance of f. ∎

We are now ready to prove the characterizations of elliptic isometries, hyperbolic, and parabolic isometries.

Proof of Theorem 12.4. It is obvious that f is an elliptic isometry if and only if f has a fixed point in X. In particular, f has a bounded orbit, since any fixed point is an orbit. More generally, if x_0 is a fixed point, then $d(f^n(x), x_0) = d(x, x_0)$ for any $n \in \mathbb{Z}$ by immediate induction, therefore the orbit of any point $x \in X$ is bounded. Conversely, assume that the orbit S of some point $x \in X$ is bounded. One can consider the unique minimal bounded ball B for S (see Lemma 12.9). Since $f(S) = S$, we have that $f(B) = B$, which means that the center of B must be fixed by f[3].

For the second characterization, first assume that f is elliptic. By Lemma 12.11, the fixed set F is either empty, or reduced to a point, or is a hyperbolic subspace. It follows that the intersection of $\partial_\infty X$ with the closure of F in $X \cup \partial_\infty X$ is either empty (when F is empty or reduced to a point), or consists of two points (when F is a geodesic), or infinitely many points (when F is a hyperbolic subspace of dimension $\geqslant 2$). Moreover, it is straightforward to prove that such points are neutral fixed points of f. To conclude that (ii) implies (iii), we show that f has no other fixed points in $\partial_\infty X$. Let $\xi \in \partial_\infty X$ be a fixed point. Since f is elliptic, it has a fixed point $x \in X$. The geodesic ray from x to ξ must be fixed by f, therefore ξ is indeed in the closure of F.

Let us finally show that (iii) implies that f is elliptic. First note that by Lemma 12.10, if f has no ideal fixed points, then f must have a fixed point in X. Now assume that f has two or more neutral ideal fixed points. For any two such fixed points ξ_1 and ξ_2, the geodesic with endpoints ξ_1 and ξ_2 is preserved by f by Lemma 12.12. Using the notations of Lemma 12.12, if $l \neq 0$ then the geodesic is a translation axis of f (by definition). However we shall see in the proof of Theorem 12.5 that an isometry that has a translation axis is hyperbolic, and has

[3]The idea of this proof goes back to Cartan [Car88], who used the center of mass in place of the center of the minimal bounding sphere to show the existence of a fixed point for the action of any compact group of isometries of a complete and simply connected manifold of nonpositive sectional curvature. This is known as the Cartan fixed point theorem.

no neutral ideal fixed points. In conclusion we must have $l = 0$, in other words the geodesic is fixed pointwise, therefore f is elliptic. ∎

Proof of Theorem 12.5. Assume that f is hyperbolic. Let x_0 be a point where $l = \min d_f$ is attained. We claim that the geodesic through x_0 and $f(x_0)$ is a translation axis for f. Indeed, let γ be the unit parametrization of this geodesic so that $\gamma(0) = x_0$ and $\gamma(l) = f(x_0)$. Consider the geodesic $f \circ \gamma$. By definition of the translation length, we have $d(\gamma(t), f(\gamma(t))) \leqslant l$ for all $t \in \mathbb{R}$. On the other hand, $d(\gamma(t), f(\gamma(t))) = l$ when $t = 0$ and $t = l$. Since the function $t \mapsto d(\gamma(t), f(\gamma(t)))$ is convex and has two distinct minimizers, it must be constant. This implies that $f \circ \gamma$ and γ are the same geodesic up to parametrization (see Corollary 11.5), in fact we must have $f(\gamma(t)) = \gamma(t + l)$ for all $t \in \mathbb{R}$. This proves that γ is a translation axis for f. Note that such a translation axis is unique: if γ_2 is another translation axis, parametrized by unit speed so that $f(\gamma_2(t)) = \gamma_2(t + l)$ for all $t \in \mathbb{R}$ (same translation parameter by Lemma 12.12), then one argues similarly as before that $d(\gamma(t), \gamma_2(t))$ is bounded, and that γ and γ_2 parametrize the same geodesic. Conversely, if f admits a translation axis, then f is a hyperbolic isometry by Lemma 12.12.

Now let us prove that when f is hyperbolic, the induced map (still denoted f) on the ideal boundary $\partial_\infty X$ has North-South dynamics. We know that $\partial_\infty X$ is a compact Hausdorff topological space, and that f has exactly two fixed points $\xi^-, \xi^+ \in \partial_\infty X$ (the endpoints of its axis). In such a situation, it is enough to show that for any $\xi \in \partial_\infty X - \{\xi^-\}$, the point ξ^- is not an accumulation point of the sequence $(f^n(\xi))_{n\in\mathbb{N}}$: this is an exercise of general topology that we leave to the diligent reader. By contradiction, assume that there exists $\xi \neq \xi-$ and a sequence of integers $n_k \longrightarrow +\infty$ such that $\lim_{k\to+\infty} f^{n_k}(\xi) = \xi^-$. Let r be the geodesic ray from x_0 to ξ, where x_0 is some point on the axis of f, and denote $y_0 = r(l_f)$. Clearly, $f^n(\xi)$ is the endpoint of the geodesic ray from $f^n(x_0)$ that goes through $f^n(y_0)$ at time $t = l_f$. On the other hand, ξ^- is the endpoint of the geodesic ray from $f^n(x_0)$ that goes through $f^{n-1}(x_0)$ at time $t = l_f$. The topology on $\partial_\infty X$ implies that if $\lim_{k\to+\infty} f^{n_k}(\xi) = \xi^-$, then $d(f^{n_k}(y_0), f^{n_k-1}(x_0)) \longrightarrow 0$. However this distance is constant equal to $d(y_0, f^{-1}(x_0))$, hence the contradiction.

Finally, let us show that if f has two ideal fixed points ξ^-, ξ^+ on the boundary, which are not neutral, then f has a translation axis. By Lemma 12.12, the geodesic with endpoints ξ^- and ξ^+ is preserved by f, and either entirely consists of fixed points, in which case f is elliptic, or is a translation axis for f. In the first case, we have seen that ξ^- and ξ^+ are neutral fixed points, so it is excluded. ∎

Proof of Theorem 12.6. Let f be a parabolic isometry. Since f has no fixed points in X, f must have at least one ideal fixed point $\xi \in \partial_\infty X$. There can be no other ideal fixed point, for otherwise f would be elliptic or hyperbolic by Lemma 12.12. Conversely, if f has a unique ideal fixed point, then f must be parabolic because Theorem 12.4 and Theorem 12.5 rule out f being elliptic or hyperbolic.

By Proposition 11.37, f must send any horosphere S centered at ξ to another such horosphere S'; let us show that if $S' \neq S$ then f cannot be parabolic[4]. Let $x_0 \in S$ be a point that minimizes $d(x, f(x))$ for all $x \in S$. Such a minimizer exists: indeed, consider a minimizing sequence $(x_n)_{n \in \mathbb{N}}$. By compactness of $S \cup \{\xi\}$, one can extract a converging subsequence in $S \cup \{\xi\}$. The limit cannot be ξ, since $d(x_n, f(x_n)) \to +\infty$ whenever $x_n \in S \to \xi$, we leave this claim as an exercise. Let γ be the geodesic through x_0 with endpoint ξ. Call S_t is the horosphere centered at ξ going through $\gamma(t)$, so that $S_0 = S$ and $S_d = S'$ where $d = d(x_, f(x_0))$. Repeat the same procedure as before to find a minimizer $x_t \in S_t$ of $d(x, f(x))$ for all $x \in S_t$. Since $f^n(S_0) = S_{nd}$ for all $n \in \mathbb{Z}$, we may find a global minimum of $t \in \mathbb{R} \mapsto d(x_t, f(x_t))$ in the interval $[0, d]$. It is straightforward to conclude that this is a minimum of $d(x, f(x))$ over all $x \in X$. This proves that the translation distance of f is attained, so that f cannot be parabolic.

It remains to show that if an isometry f preserves some horosphere S and has no fixed points in S, then f is parabolic. First of all, it is clear that f fixes the center $\xi \in \partial_\infty X$ of S, since ξ is the only ideal point in the closure of S. Secondly, it is immediate from the definition of a horosphere that f must actually preserve any horosphere centered at ξ. If f was elliptic, it would have a fixed point $x_0 \in X$. The whole geodesic through x_0 and with endpoint ξ would then have to be pointwise fixed. This geodesic intersects each horosphere centered at ξ (once), therefore we would find a fixed point of f in S, contrary to the assumption. If f was hyperbolic, by Theorem 12.5 it would have another endpoint $\xi' \in \partial_\infty X$, and the geodesic with endpoints ξ and ξ' would be its axis. Let γ be a unit parametrization of this geodesic, so that $f(\gamma(t)) = \gamma(t + l)$ for all $t \in \mathbb{R}$, where $|l|$ is the translation length of f. Such a geodesic intersects S at a unique point $x_0 = \gamma(t_0)$, therefore $f(x_0) = \gamma(t_0 + l)$ cannot belong to S, contrary to the assumption that f preserves S. ■

12.2 Description

The classification established in the previous section is fundamental, but let us characterize in more detail the elliptic, loxodromic, and parabolic isometries of hyperbolic space $\mathbb{H}^n$. In the next section, we shall give even more explicit descriptions when $n = 2$ and $n = 3$.

For many purposes, it is good enough to understand isometries up to conjugation, in other words to classify conjugacy classes of isometries. Indeed, one can easily derive the properties of an isometry from that of a conjugate: for instance, if f is a loxodromic isometry with axis L, then gfg^{-1} is a loxodromic with axis $g(L)$ and same translation length, etc.

12.2.1 Elliptic isometries

Let f be an elliptic isometry of $\mathbb{H}^n$. We have seen that the set of fixed points F of f is a hyperbolic subspace of $\mathbb{H}^n$, in other words F is a copy (totally geodesic embedding) of $\mathbb{H}^k$

[4] Refer to [BH99, Chap. II.8, Prop. 8.25] for an alternative proof that uses the convexity of the displacement function.

inside $\mathbb{H}^n$. Note that we allow $k = 0$ (F is reduced to a point) and $k = 1$ (F is a geodesic).

Let x_0 be any point in F. The fact that $\mathbb{H}^n$ is uniquely geodesic implies that f is completely determined by its derivative $\mathrm{d}f_{x_0}$. Indeed, for any $x \in \mathbb{H}^n$, we have $f(x) = \gamma_{\mathrm{d}f_{x_0}(u)}(1)$, where $x = \gamma_u(1)$ (in other words f is conjugate to $\mathrm{d}f_{x_0}$ by the Riemannian exponential map $\exp_{x_0}$). The linear map $\mathrm{d}f_{x_0}$ is a linear isometry of the Euclidean vector space $\mathrm{T}_{x_0} M$, and its +1-eigenspace (a.k.a fixed point set) is the tangent subspace to F. Thus the "interesting" part of the action of f resides in the behavior of $\mathrm{d}f$ in the orthogonal complement. Let us record these simple observations:

Theorem 12.13. *Any elliptic isometry of $\mathbb{H}^n$ is uniquely determined by:*

(1) Its set of fixed points $F \subseteq \mathbb{H}^n$, which is a hyperbolic subspace.

(2) For some $x_0 \in F$, a Euclidean isometry of $\mathrm{T}_{x_0}\mathbb{H}^n$, whose +1-eigenspace is $\mathrm{T}_{x_0} F$.

In this description, the splitting of $\mathrm{d}_{x_0}\mathbb{H}^n$ as $V \oplus V^\perp$ corresponds to two orthogonally transverse hyperbolic subspaces of $\mathbb{H}^n$ through x_0, the first (F) being fixed by f, and the second being preserved by f with x_0 as the unique fixed point.

Alternatively to this infinitesimal approach, one can realize that f is adequately described by its set of fixed points F and by a Euclidean isometry by looking at its action on horospheres orthogonally transverse to F. Indeed, it is immediate that any such horosphere S must be preserved by f (why?). Moreover, we recall the important fact that the hyperbolic metric restricts to a Euclidean metric on any horosphere (Theorem 11.40). Therefore f acts by Euclidean isometries on S.

Example 12.14. Consider an elliptic isometry $f \in \mathrm{Isom}^+(\mathbb{H}^3)$ in the Poincaré upper half-space model H^3, whose set of fixed points is the geodesic F with endpoints 0 and ∞. Then f preserves each horosphere centered at ∞, which is a horizontal plane in H^3, and is orthogonal to F. Per the above discussion, f acts in such a plane by Euclidean rotations. In the coordinates $(z = x_1 + ix_2, x_3) \in H^3$, the map f is written $f(z, x_3) = (e^{i\theta}z, x_3)$ for some real number θ. Note that horospheres centered at 0 are also orthogonal to F and preserved by f, as expected. See Figure 12.1 and Figure 12.2.

12.2.2 Loxodromic isometries

Let us now turn to loxodromic isometries of hyperbolic space. We called such isometries *hyperbolic* in a general metric space X (see Definition 12.1), but the term *loxodromic* should be preferred when $X = \mathbb{H}^n$.

Translations

Translations are the "nicest" loxodromic isometries.

Definition 12.15. A loxodromic isometry $f : \mathbb{H}^n \to \mathbb{H}^n$ is called a ***translation*** if it preserves some/any equidistant curve from its axis.

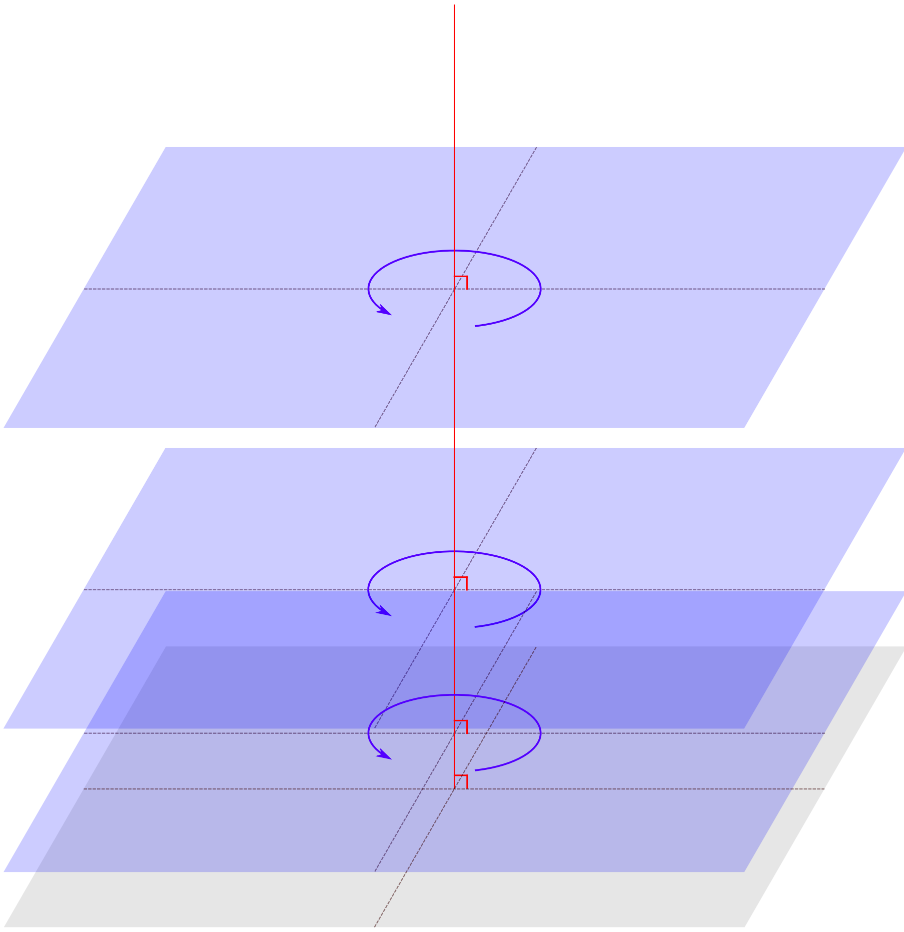


Figure 12.1: An elliptic isometry of the Poincaré half-space and its action by Euclidean rotations on horospheres centered at $\xi = \infty$.

The fact that "some" and "any" are equivalent in the definition above will be apparent in the proof of Proposition 12.18.

Remark 12.16. It is quite common in the literature to use the term *hyperbolic isometry* instead of *translation*. Such authors will also typically exclude translations from loxodromic isometries. I recommend not using this terminology (see Remark 12.2), or at least saying "purely hyperbolic" for translations and "purely loxodromic" for other loxodromic elements, to avoid any confusion.

Example 12.17. For any $\lambda > 0$, the map $z \mapsto \lambda z$ defines a translation in the Poincaré half-plane. In fact, the next characterization of translations shows that any translation is conjugate to a map of this form.

Proposition 12.18. *Let $f : \mathbb{H}^n \to \mathbb{H}^n$. The following are equivalent:*

(i) f is a loxodromic isometry and preserves some/any equidistant curve from its axis.

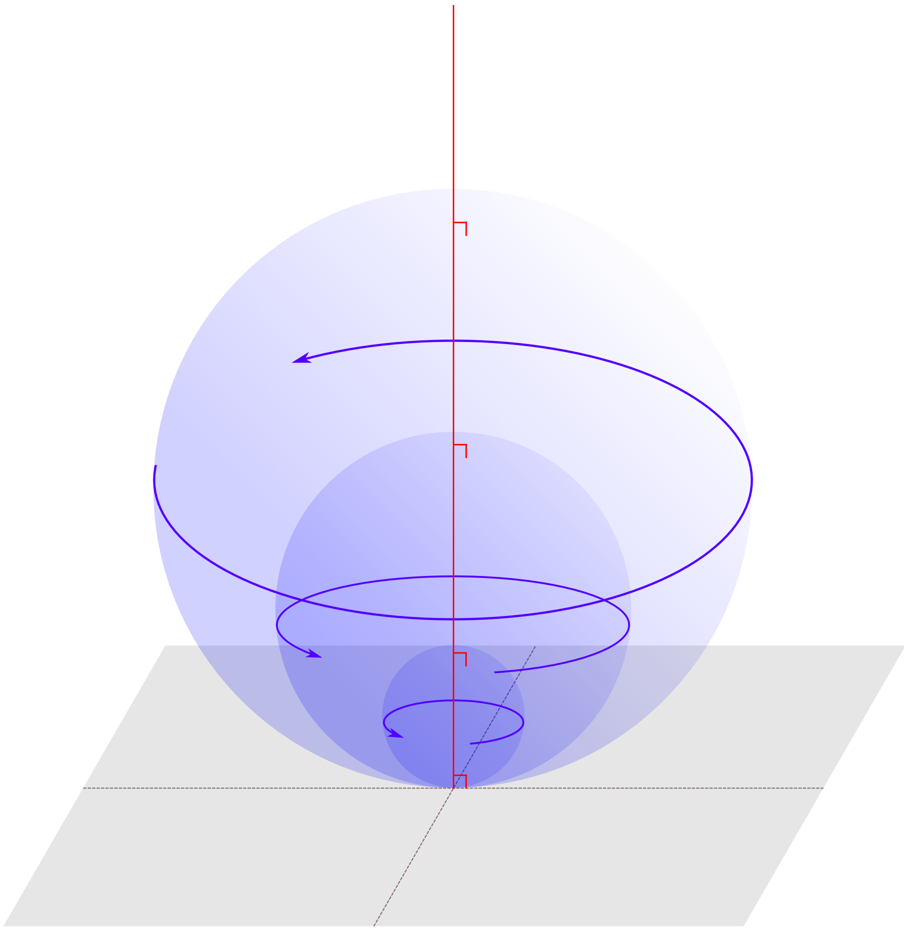


Figure 12.2: The same elliptic isometry as in Figure 12.1, acting on horospheres centered at $\xi = 0$.

(ii) f is conjugate to the transformation of the Poincaré half-space given by $x \in H^n \mapsto e^l x$, where l is the translation length of f.

Proof. This is the content of Exercise 12.4. ∎

A translation is uniquely determined by its axis and translation length:

Theorem 12.19. *For any oriented geodesic line $L \subseteq \mathbb{H}^n$ and $l > 0$, there exists a unique translation with axis L and translation length l.*

Proof. Denote by f_0 the transformation of the Poincaré half-space given by $x \in H^n \mapsto e^l x$ (as in Proposition 12.18). This is a translation with axis L_0, the geodesic line with endpoints 0 and ∞, and with translation length l. Let $\varphi\colon \mathbb{H}^n \to H^n$ be any isometry that sends the endpoints of L to L_0, preserving orientation (why does this exist?). Then $\varphi f \varphi^{-1}$ is a translation with axis L and translation length l . This shows existence.

For uniqueness, assume that f_1 and f_2 are two translations with same axis L and translation length l. $g := f_2 \circ f_1^{-1}$ fixes L pointwise, so that g is an elliptic transformation whose set of fixed points contains L. In particular, g preserves the horospheres centered at ∞, which are the horizontal hyperplanes in $H^n \subseteq \mathbb{R}^n$. On the other hand, g must preserve the equidistant lines from L_0, which are the Euclidean straight half-lines starting from 0. Since any such half-line intersects any aforementioned horosphere exactly once, ■

General loxodromic transformations

A general loxodromic transformation is determined by the data of an axis, a translation length, and a Euclidean isometry. More precisely:

Theorem 12.20. *Let $f\colon H^n \to H^n$. The following are equivalent:*

(i) f is a loxodromic isometry with axis L and translation length l.

(ii) $f = t \circ r$ where t is the translation with axis L and translation length l, and r is an elliptic isometry whose set of fixed points contains L.

Remark 12.21. The decomposition $f = t \circ r$ in Theorem 12.20 is unique, since t is uniquely determined by L and l (Theorem 12.19).

Proof. Let f be a loxodromic isometry with axis L and translation length l and let t be the unique translation with axis L and length t. It is immediate that $r := f \circ t^{-1}$ is an isometry that fixes L pointwise, therefore r is an elliptic isometry.

Conversely, assume $f = r \circ t$ where r and t are as before. Clearly, f is translates by l in restriction to L. By Theorem 12.5, since f has an axis, it is a loxodromic isometry. More precisely, f is a loxodromic isometry with translation length l by Lemma 12.12. ■

We shall see examples of loxodromic isometries in § 12.3 and § 12.4, e.g. Figure 12.6.

12.2.3 Parabolic isometries

A parabolic isometry is determined by the choice of an ideal fixed point and a Euclidean isometry without fixed points. More precisely:

Theorem 12.22. *Any parabolic isometry with ideal fixed point $\xi \in \partial_\infty \mathbb{H}^n$ acts as a Euclidean isometry in any horosphere with center ξ. Conversely, given a Euclidean isometry f_0 in some horosphere S_0 centered at ξ, without any fixed points, there exist a unique parabolic isometry whose restriction to S_0 coincides with f_0.*

Proof. Let f be a parabolic isometry with fixed point $\xi \in \partial_\infty X$. By Theorem 12.6, f preserves any horosphere with center ξ. Recall that any horosphere with its induced metric is isometric to Euclidean space (Theorem 11.40). It follows that f must act as a Euclidean isometry in any horosphere with center ξ.

Conversely, let us show that any Euclidean isometry f_0 of some horosphere S_0 centered at ξ uniquely extends as a parabolic isometry. For any $x \in \mathbb{H}^n$ and $t \in \mathbb{R}$, let $\varphi_t(x)$ denote the point through which the unit geodesic starting from x and with endpoint ξ goes at time t. Such a geodesic is orthogonally transverse to all horospheres centered at ξ (see Exercise 11.5). Moreover, for any horosphere S centered at ξ, there exists a unique $t \in \mathbb{R}$ such that $\varphi_t(S_0) = S$ (t is the signed distance between S_0 and S). One can show that any parabolic isometry f with fixed point ξ commutes with φ_t for any $t \in \mathbb{R}$, let us leave this claim as an exercise. It easily follows that f is uniquely determined by its restriction to any horosphere S centered at ξ. ■

12.3 Isometries of $\mathbb{H}^2$

We shall now describe isometries even more concretely in dimensions 2 and 3. For simplicity, we shall only consider orientation-preserving isometries. We recall that in the Poincaré models of hyperbolic space, isometries can be described as Möbius transformations; moreover in dimensions 2 and 3 these are identified to fractional linear transformations.

12.3.1 Isometries of the Poincaré half-plane

Let us favor the Poincaré half-plane model $\mathbb{H} \subseteq \mathbb{C}$. The group of orientation-preserving isometries of $\mathbb{H}^2$ is identified to $\mathrm{PSL}(2, \mathbb{R})$, acting on $\mathbb{H}$ by fractional linear transformations. Let us briefly recall how this works: any matrix

$$M = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in \mathrm{SL}(2, \mathbb{R})$$

induces an isometry of $\mathbb{H}$ given by

$$f_M : z \mapsto \frac{az + b}{cz + d}.$$

The assignment $M \to f_M$ is a group homomorphism from $\mathrm{SL}(2, \mathbb{R})$ to $\mathrm{Isom}(\mathbb{H})$, whose image is $\mathrm{Isom}^+(\mathbb{H})$ and whose kernel is $\{-I_2, I_2\}$, so that it induces an isomorphism $\mathrm{PSL}(2, \mathbb{R}) = \mathrm{SL}(2, \mathbb{R})/\{\pm I_2\} \xrightarrow{\sim} \mathrm{Isom}^+(\mathbb{H})$.

As a consequence of this discussion, the trace of an orientation-preserving isometry of $\mathbb{H}$ is well-defined *up to sign.*

12.3.2 Elliptic isometries

Theorem 12.23. *Let $f : \mathbb{H} \to \mathbb{H}$ be an orientation-preserving isometry, represented by $M \in \mathrm{SL}(2, \mathbb{R})$. Assume $f \neq \mathrm{id}$. The following are equivalent:*

(i) f *is an elliptic isometry.*

(ii) f *has a unique fixed point in* $\mathbb{H}$.

(iii) $\operatorname{tr} M \in (-2, 2)$.

(iv) M *is conjugate in* $\mathrm{SL}(2,\mathbb{R})$ *to* $\pm R_\theta$ *for some* $\theta \in \mathbb{R} - 2\pi\mathbb{Z}$, *where* $R_\theta = \begin{bmatrix} \cos(\frac{\theta}{2}) & \sin(\frac{\theta}{2}) \\ -\sin(\frac{\theta}{2}) & \cos(\frac{\theta}{2}) \end{bmatrix}$.

(v) f *is conjugate in* $\mathrm{Isom}^+(\mathbb{H})$ *to* f_θ *for some* $\theta \in \mathbb{R} - 2\pi\mathbb{Z}$, *where* $f_\theta(z) = \frac{(\cos(\frac{\theta}{2}))z+\sin(\frac{\theta}{2})}{-(\sin(\frac{\theta}{2}))z+\cos(\frac{\theta}{2})}$.

Before proving this theorem, let us establish an elementary yet useful lemma.

Lemma 12.24. *Let* $M = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in \mathrm{SL}(2,\mathbb{R})$ *and denote* $f\colon z \mapsto \frac{az+b}{cz+d}$ *the associated fractional linear transformation of* $\hat{\mathbb{C}}$.

- *If* $(\operatorname{tr} M)^2 > 4$, *then* f *has two fixed points, both of which lie in* $\hat{\mathbb{R}} \subseteq \hat{\mathbb{C}}$.
- *If* $(\operatorname{tr} M)^2 < 4$, *then* f *has two fixed points, one in* $\mathbb{H}$ *and the other is its complex conjugate.*
- *If* $(\operatorname{tr} M)^2 = 4$, *then either* f *is the identity, or* f *has a unique fixed point, which lies in* $\hat{\mathbb{R}} \subseteq \hat{\mathbb{C}}$.

Proof. This is a nice exercise: see Exercise 12.5. ■

Proof of Theorem 12.23. The fact that (i) ⇔ (ii) ⇔ (iii) immediately follows from Lemma 12.24. The implication (iv) ⇒ (iii) is trivial, as is (iv) ⇔ (v).

Finally, let us prove (ii) ⇒ (v). Assume f has a unique fixed point $z_0 \in \mathbb{H}$. Since $\mathrm{Isom}^+(\mathbb{H})$ acts transitively on $\mathbb{H}$, there exists $g \in \mathrm{Isom}^+(\mathbb{H})$ such that $g(z_0) = i$. Then $f_1 := gfg^{-1}$ is a fractional linear transformation of $\mathbb{H}$ that fixes i. It is elementary to check by direct computation that $z \mapsto \frac{a_1z+b_1}{c_1z+d_1}$ fixes i if and only if $d_1 = a_1$ and $b_1 = -c_1$. Since $a_1d_1 - b_1c_1 = 1 = a_1^2 + c_1^2$, there exists $\theta \in \mathbb{R}$ such that $a_1 = \cos(\frac{\theta}{2})$ and $c_1 = -\sin(\frac{\theta}{2})$. We conclude that $f_1 = f_\theta$. ■

Remark 12.25. We can alternatively write a proof of (iii) ⇒ (iv) using only linear algebra. The characteristic polynomial of $M \in \mathrm{SL}(2,\mathbb{R})$ is $\chi_M(\lambda) = \lambda^2 - (\operatorname{tr} M)\lambda + 1$, with discriminant $\Delta = (\operatorname{tr} M)^2 - 4$. If $\operatorname{tr} M \in (-2, 2)$, then χ_M has two non-real complex conjugate roots, and since their product is 1 they must be $\lambda = e^{\pm i\frac{\theta}{2}}$ for some $\theta \in \mathbb{R} - 2\pi\mathbb{Z}$. It follows that M has two distinct eigenvalues $\lambda = e^{\pm i\frac{\theta}{2}}$, therefore M is conjugate in $\mathrm{SL}(2,\mathbb{C})$ to $D_\theta = \operatorname{diag}(e^{i\frac{\theta}{2}}, e^{-i\frac{\theta}{2}})$. On the other hand, the matrix R_θ is also conjugate to D_θ in $\mathrm{SL}(2,\mathbb{C})$. We therefore find that M is conjugate to R_θ in $\mathrm{SL}(2,\mathbb{C})$. Conclude with the standard–albeit non-trivial–fact of linear algebra that two matrices in $\mathrm{SL}(2,\mathbb{R})$ are conjugate in $\mathrm{SL}(2,\mathbb{C})$ if and only if they are conjugate in $\mathrm{SL}(2,\mathbb{R})$.

A representation of the "standard" elliptic isometry f_θ is shown in Figure 12.3.

Corollary 12.26. *The conjugacy class of an elliptic element of* $\mathrm{Isom}^+(\mathbb{H})$ *is uniquely determined by its trace (which is a real number defined up to sign).*

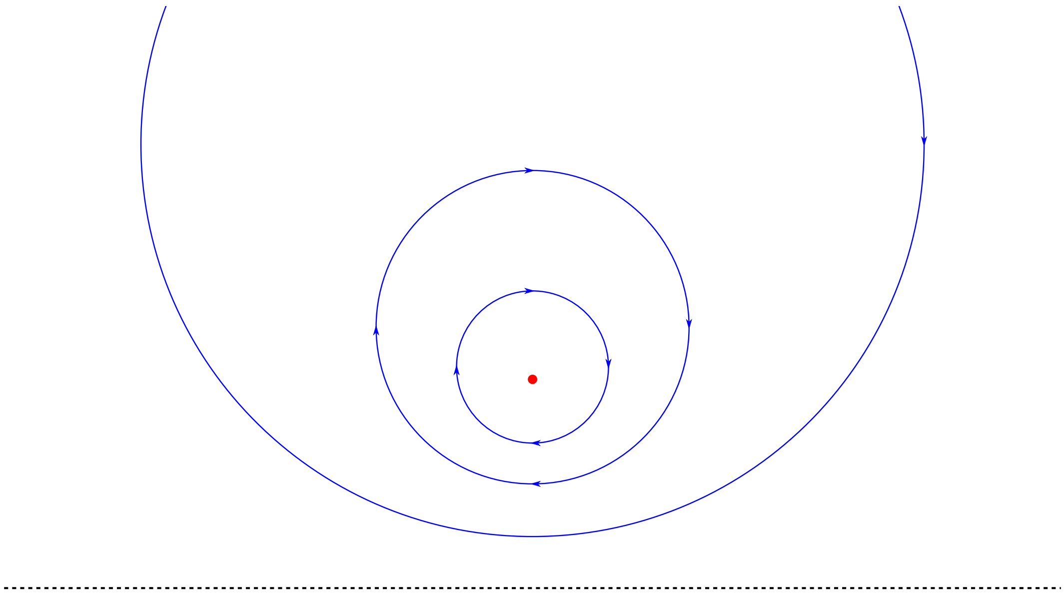


Figure 12.3: An elliptic isometry of the Poincaré half-plane.

12.3.3 Loxodromic isometries

Theorem 12.27. *Let $f\colon \mathbb{H} \longrightarrow \mathbb{H}$ be an orientation-preserving isometry, represented by $M \in \mathrm{SL}(2,\mathbb{R})$. The following are equivalent:*

(i) f *is a loxodromic isometry.*

(ii) f *has no fixed points in* $\mathbb{H}$, *and two distinct fixed points in* $\partial_\infty \mathbb{H} = \hat{\mathbb{R}}$.

(iii) $\operatorname{tr} M \in \mathbb{R} - [-2, 2]$.

(iv) M *is conjugate in* $\mathrm{SL}(2,\mathbb{R})$ *to* $\pm T_l$ *for some* $l \in \mathbb{R} - \{0\}$, *where* $T_l = \begin{bmatrix} e^{l/2} & 0 \\ 0 & e^{-l/2} \end{bmatrix}$.

(v) f *is conjugate in* $\mathrm{Isom}^+(\mathbb{H})$ *to* f_l *for some* $l \in \mathbb{R} - \{0\}$, *where* $f_l(z) = e^l z$.

(vi) f *is a translation.*

The absolute value of the number l in (iv) and (v) is the translation length of f.

Proof. The equivalence (i) $\Leftrightarrow$ (ii) is general: see Theorem 12.5. Lemma 12.24 shows that (ii) $\Leftrightarrow$ (iii). The equivalence (iii) $\Leftrightarrow$ (iv) is elementary linear algebra (it is an easier version of Remark 12.25, since M is diagonalizable over $\mathbb{R}$). The equivalence (iv) $\Leftrightarrow$ (v) is immediate. To prove that (v) implies (vi), it suffices to check that f_l is a translation, since the conjugate of any translation is a translation. The fact that f_l is a translation is a special case of Proposition 12.18. Finally, (vi) $\Rightarrow$ (i) is trivial. ∎

Remark 12.28. We emphasize that there are no "purely loxodromic" isometries of $\mathbb{H}^2$: this is just a way to rephrase (i) $\Leftrightarrow$ (vi).

A representation of the "standard" translation f_l is shown in Figure 12.4.

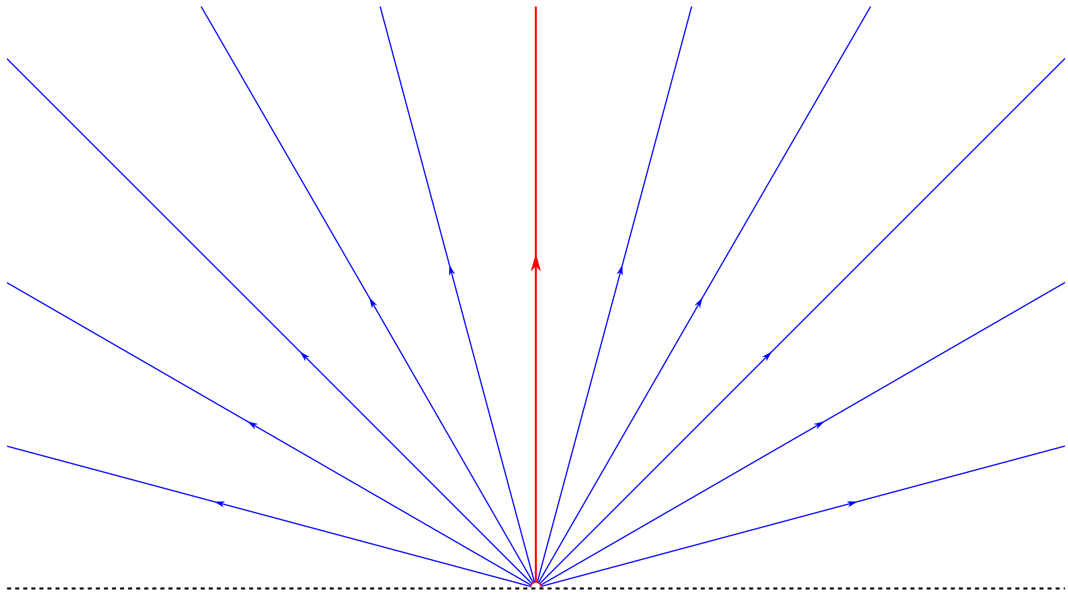


Figure 12.4: A translation of the Poincaré half-plane.

Corollary 12.29. *The conjugacy class of a loxodromic element of* $\mathrm{Isom}^+(\mathbb{H})$ *is uniquely determined by its trace (which is a real number defined up to sign).*

12.3.4 Parabolic isometries

Theorem 12.30. *Let* $f\colon \mathbb{H} \longrightarrow \mathbb{H}$ *be an orientation-preserving isometry, represented by* $M \in \mathrm{SL}(2,\mathbb{R})$. *Assume* $f \neq \mathrm{id}$. *The following are equivalent:*

(i) f *is a parabolic isometry.*

(ii) f *has no fixed points in* $\mathbb{H}$, *and one fixed point in* $\partial_\infty \mathbb{H} = \hat{\mathbb{R}}$.

(iii) $\operatorname{tr} M = \pm 2$.

(iv) M *is conjugate in* $\mathrm{SL}(2,\mathbb{R})$ *to* $\pm P$, *where* $P = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$.

(v) f *is conjugate in* $\mathrm{Isom}^+(\mathbb{H})$ *to* $z \longmapsto z + 1$.

Proof. The equivalence (i) $\Leftrightarrow$ (ii) is general: see Theorem 12.6. Lemma 12.24 shows that (ii) $\Leftrightarrow$ (iii). The fact that (iii) implies (iv) is elementary linear algebra: one quickly shows that M

has $+1$ as a repeated eigenvalue (or -1, but in that case consider $-M$), moreover M cannot be diagonalizable otherwise we would have $M = I_2$ and $f = \mathrm{id}$, therefore the Jordan normal form of M must be P. The converse (iv) $\Longrightarrow$ (iii) is trivial. Finally, the equivalence (iv) $\Leftrightarrow$ (v) is trivial. ∎

We emphasize that there is only one conjugacy class of parabolic isometries of $\mathbb{H}^2$:

Corollary 12.31. *Any parabolic element of* $\mathrm{Isom}^+(\mathbb{H})$ *has trace* ± 2*, and is conjugate to* $z \mapsto z+1$*. Conversely, any* $f \in \mathrm{Isom}^+(\mathbb{H})$ *of trace* ± 2 *is parabolic,* provided $f \neq \mathrm{id}$.

A representation of the "standard" parabolic isometry $z \mapsto z + 1$ is shown in Figure 12.5.

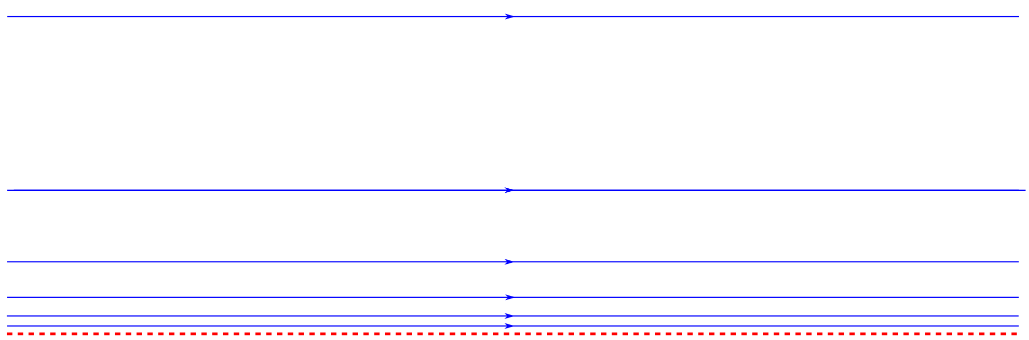


Figure 12.5: A parabolic isometry of the Poincaré half-plane.

12.3.5 Conjugacy classes and trace

As a consequence of Corollary 12.26, Corollary 12.29, and Corollary 12.31, we obtain:

Theorem 12.32. *The conjugacy class of an element of* $f \in \mathrm{Isom}^+(\mathbb{H}) - \{\mathrm{id}\}$ *is uniquely determined by its trace (which is a real number defined up to sign).*

More precisely, summarizing previous results:

- If $\mathrm{tr}(f) = \pm 2\cos(\frac{\theta}{2}) \in [2, 2]$, then f is elliptic and conjugate to $z \mapsto \frac{(\cos(\frac{\theta}{2}))z+\sin(\frac{\theta}{2})}{-(\sin(\frac{\theta}{2}))z+\cos(\frac{\theta}{2})}$.
- If $\mathrm{tr}(f) = \pm 2\cosh(l/2) \in \mathbb{R} - [2, 2]$, then f is a translation and conjugate to $z \mapsto e^l z$.
- If $\mathrm{tr}(f) = \pm 2$ and $f \neq \mathrm{id}$, then f is parabolic and conjugate to $z \mapsto z + 1$.

Remark 12.33. Although it is unambiguous from the definition that $f = \mathrm{id}$ is an elliptic isometry, it is quite special: it has the same trace as parabolic isometries. Moreover, it can be approached by translations as well as by non-trivial rotations. Informally speaking, $f = \mathrm{id}$ is at the junction between elliptic, loxodromic, and parabolic isometries.

12.4 Isometries of $\mathbb{H}^3$

12.4.1 Isometries of the Poincaré half-space

Let us favor the Poincaré half-space model $H^3 = \mathbb{C} \times [0, \infty) \subseteq \mathbb{R}^3$. We shall use coordinates $(z = x_1 + ix_2, x_3)$. The group of orientation-preserving isometries of $\mathbb{H}^3$ is identified to $\mathrm{PGL}(2, \mathbb{C})$, acting on $\partial_\infty H^3 = \hat{\mathbb{C}}$ by fractional linear transformations, and acting in H^3 by via the Poincaré extension. Let us recall how this works: any matrix

$$M = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in \mathrm{GL}(2, \mathbb{C})$$

induces a orientation-preserving Möbius transformation of $\hat{\mathbb{C}}$ given by

$$f_M \colon z \mapsto \frac{az + b}{cz + d} \,.$$

Any such Möbius transformation uniquely extends as a Möbius transformation of H^3 (this is called the Poincaré extension, see Theorem 9.30), which we still denote f_M, and which is an isometry of H^3. Conversely, any orientation-preserving isometry of H^3 is a Möbius transformation, and is uniquely determined by its continuous extension to $\partial_\infty H^3 = \hat{\mathbb{C}}$, which is an orientation-preserving Möbius transformation of $\hat{\mathbb{C}}$. The latter coincides with a fractional linear transformation as above.

The assignment $M \to f_M$ is a group homomorphism from $\mathrm{GL}(2, \mathbb{C})$ to $\mathrm{Isom}(H^3)$, whose image is $\mathrm{Isom}^+(H^3)$ and whose kernel is the group of homotheties $\mathbb{C}^* I_2$, so that it induces an isomorphism $\mathrm{PGL}(2, \mathbb{C}) = \mathrm{GL}(2, \mathbb{C})/\mathbb{C}^* I_2 \xrightarrow{\sim} \mathrm{Isom}^+(H^3)$.

Instead of $\mathrm{PGL}(2, \mathbb{C})$, in this section we will favor $\mathrm{PSL}(2, \mathbb{C}) := \mathrm{SL}(2, \mathbb{C})/\{-I_2, I_2\}$, which is basically the same group (there is a natural isomorphism $\mathrm{PSL}(2, \mathbb{C}) \xrightarrow{\sim} \mathrm{PGL}(2, \mathbb{C})$). Essentially, any matrix in $\mathrm{GL}(2, \mathbb{C})$ can be multiplied by some $\lambda \in \mathbb{C}^*$ so that the resulting matrix has determinant 1, and the associated fractional linear transformations are the same. More precisely, the story above can be repeated for $\mathrm{SL}(2, \mathbb{C})$: the assignment $M \to f_M$ is a group homomorphism from $\mathrm{SL}(2, \mathbb{C})$ to $\mathrm{Isom}(H^3)$, whose image is still $\mathrm{Isom}^+(H^3)$ and whose kernel is $\mathbb{C}^* I_2 \cap \mathrm{SL}(2, \mathbb{C}) = \{-I_2, I_2\}$, so that it induces an isomorphism $\mathrm{PSL}(2, \mathbb{C}) = \mathrm{SL}(2, \mathbb{C})/\{-I_2, I_2\} \xrightarrow{\sim} \mathrm{Isom}^+(H^3)$.

The benefits of $\mathrm{SL}(2, \mathbb{C})$ over $\mathrm{GL}(2, \mathbb{C})$ is that not only it will be useful in this section to assume that all matrices have determinant 1, it is especially convenient that we can associate a matrix $M \in \mathrm{SL}(2, \mathbb{C})$ unique up to sign to any $f \in \mathrm{Isom}^+(H^3)$. In particular, the trace of $f \in \mathrm{Isom}^+(H^3)$ is well-defined complex number up to sign.

12.4.2 Elliptic isometries

Theorem 12.34. *Let $f \colon H^3 \to H^3$ be an orientation-preserving isometry, represented by $M \in \mathrm{SL}(2, \mathbb{C})$. Assume $f \neq \mathrm{id}$. The following are equivalent:*

(i) f *is an elliptic isometry.*

(ii) *The set of fixed points of* f *in* H^3 *is a geodesic.*

(iii) $\operatorname{tr} M \in (-2,2) \subseteq \mathbb{R} \subseteq \mathbb{C}$.

(iv) M *is conjugate in* $\mathrm{SL}(2,\mathbb{C})$ *to* $\pm R_\theta$ *for some* $\theta \in \mathbb{R} - 2\pi\mathbb{Z}$, *where* $R_\theta = \begin{bmatrix} e^{i\frac{\theta}{2}} & 0 \\ 0 & e^{-i\frac{\theta}{2}} \end{bmatrix}$.

(v) f *is conjugate in* $\mathrm{Isom}^+(\mathbb{H})$ *to* f_θ *for some* $\theta \in \mathbb{R} - 2\pi\mathbb{Z}$, *where* f_θ *is given by* $(z, x_3) \mapsto (e^{i\theta} z, x_3)$.

Before writing the proof of this theorem, we show the useful lemma:

Lemma 12.35. *Let* $M = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in \mathrm{SL}(2,\mathbb{C})$ *and denote* $f\colon z \mapsto \frac{az+b}{cz+d}$ *the associated fractional linear transformation of* $\hat{\mathbb{C}}$.

- *If* $\operatorname{tr} M \in \mathbb{C} - (-2,2)$, *then* f *has exactly two fixed points in* $\hat{\mathbb{C}}$, *one attracting and one repelling.*
- *If* $\operatorname{tr} M \in (-2,2)$, *then* f *has exactly two fixed points in* $\hat{\mathbb{C}}$, *both neutral.*
- *If* $(\operatorname{tr} M)^2 = 4$, *then either* f *is the identity, or* f *has a unique fixed point in* $\hat{\mathbb{C}}$, *which is neutral.*

Proof. Firstly, one readily shows that a fixed point $z_0 \in \mathbb{C}$ is attracting [resp. repelling, resp. neutral] in the sense defined in § 12.1 if and only if the "multiplier" $|f'(z_0)|$ is < 1 [resp. > 1, resp. $= 1$]. For $z_0 = \infty$, take $|g'(0)|$ instead, where $g(z) = 1/f(1/z)$.

If $c \neq 0$, a fixed point of f is a root of the quadratic polynomial $cz^2 + (d-a)z - b$, with discriminant $\Delta = (\operatorname{tr} M)^2 - 4$. The derivative of f is $f'(z) = \frac{1}{(cz+d)^2}$, and at the two fixed points we have $f'(z) = \frac{4}{(\operatorname{tr} M \pm \sqrt{\Delta})^2}$. Notice that the product $f'(z_1)f'(z_2)$ is equal to 1, so that z_1 and z_2 are either distinct and attracting/repelling, or distinct and neutral, or equal and neutral. The conclusion quickly follows.

If $c = 0$, one must have $d = \frac{1}{a} \neq 0$, and the fixed points of f solve $a(az+b) = z$. If $a \neq 1$, f has two fixed points: $z_1 = \frac{ab}{a^2-1}$ and $z_2 = \infty$, with multipliers a^2 and $1/a^2$. Therefore z_1 and z_2 are either attracting and repelling, or repelling and attracting, or both neutral, depending on whether $|a| > 1$, $|a| < 1$, or $|a| = 1$. The first two cases correspond to $\operatorname{tr} M = a + \frac{1}{a} \in \mathbb{C} - (-2,2)$, and the third case to $\operatorname{tr} M = a + \frac{1}{a} \in (-2,2)$. Finally, if $a = d = 1$: we have $\operatorname{tr} M = 2$, and either $b = 0$ and f is the identity map, or $b \neq 0$ and f admits ∞ as a unique fixed point with multiplier 1. ∎

Proof of Theorem 12.34. Assume that f is elliptic. By Lemma 12.11, the set of fixed points F of f in H^3 is either a point, or a geodesic, or a 2-dimensional hyperbolic subspace. If f had a unique fixed point $p_0 \in H^3$, then f could not have any ideal fixed point $z \in \hat{\mathbb{C}}$, for otherwise the geodesic through p_0 with endpoint z would be fixed. However Lemma 12.35 shows that

f has at least one ideal fixed point. If F was a two-dimensional hyperbolic subspace, then f would have infinitely many fixed points in $\hat{\mathbb{C}}$; again this is ruled out by Lemma 12.35. Thus we proved that (i) implies (ii), and the converse is trivial.

The equivalence (i) $\Leftrightarrow$ (iii) is an immediate consequence of Theorem 12.4 and Lemma 12.35. The equivalence (iii) $\Leftrightarrow$ (iv) is elementary linear algebra.

Finally, let us prove (iv) $\Leftrightarrow$ (v). It is clear that M is conjugate to R_θ if and only if the restriction of f to $\hat{\mathbb{C}}$ is conjugate to $z \mapsto e^{i\theta}$. It remains to show that $f_\theta\colon (z, x_3) \mapsto (e^{i\theta}z, x_3)$ is the Poincaré extension of $z \mapsto e^{i\theta}$. It is enough to realize that f_θ is an isometry of H^3 (it clearly preserves the Riemannian metric), and that its continuous extension to $\hat{\mathbb{C}}$ is indeed $z \mapsto e^{i\theta}$. ■

The "standard" elliptic isometry f_θ is shown in Figure 12.1 and Figure 12.2. I also recommend checking out the website [Nel20] for cool animations of elliptic, "hyperbolic", loxodromic, and parabolic isometries in the Poincaré half-space model.

Corollary 12.36. *The conjugacy class of an elliptic element of* $\mathrm{Isom}^+(H^3)$ *is uniquely determined by its trace (which is a complex number defined up to sign).*

12.4.3 Loxodromic isometries

Theorem 12.37. *Let $f\colon H^3 \to H^3$ be an orientation-preserving isometry, represented by $M \in$* $\mathrm{SL}(2,\mathbb{C})$. *The following are equivalent:*

(i) *f is a loxodromic isometry.*

(ii) *f has no fixed points in H^3, and two distinct fixed points in $\partial_\infty H^3 = \hat{\mathbb{C}}$, one attracting and one repelling.*

(iii) $\mathrm{tr}\, M \in \mathbb{C} - [-2, 2]$.

(iv) *M is conjugate in* $\mathrm{SL}(2,\mathbb{R})$ *to $\pm T_l$ for some $l \in \mathbb{C} - i\mathbb{R}$, where* $T_l = \begin{bmatrix} e^{l/2} & 0 \\ 0 & e^{-l/2} \end{bmatrix}$.

(v) *f is conjugate in* $\mathrm{Isom}^+(H^3)$ *to f_l for some $l \in \mathbb{C} - i\mathbb{R}$, where $f_l(z, x_3) = (e^l z, x_3)$.*

The absolute value of the complex number l in (iv) and (v) is the translation length of f.

Proof. The fact that (i) $\Leftrightarrow$ (ii) $\Leftrightarrow$ (iii) is an application of Theorem 12.20 and Lemma 12.35. The equivalence (iii) $\Leftrightarrow$ (iv) is elementary linear algebra. Finally, the proof of (iv) $\Leftrightarrow$ (v) is the same as in Theorem 12.34. ■

Among loxodromic isometries, translations are special:

Corollary 12.38. *Let $f\colon H^3 \to H^3$ be an orientation-preserving isometry, represented by $M \in$* $\mathrm{SL}(2,\mathbb{C})$. *The following are equivalent:*

(i) f *is a translation.*

(ii) $\operatorname{tr} M \in \mathbb{R} - [-2, 2] \subseteq \mathbb{C}$.

(iii) M *is conjugate in* $\mathrm{SL}(2, \mathbb{C})$ *to* $\pm T_l$ *for some* $l \in \mathbb{R} - \{0\}$*, where* $T_l = \begin{bmatrix} e^{l/2} & 0 \\ 0 & e^{-l/2} \end{bmatrix}$.

(iv) f *is conjugate in* $\mathrm{Isom}^+(H^3)$ *to* f_l *for some* $l \in \mathbb{R} - \{0\}$*, where* $f_l(z, x_3) = (e^l z, e^{\mathrm{Re}(l)} x_3)$.

The absolute value of the real number l *in (iv) and (v) is the translation length of* f*.*

A representation of the "standard" loxodromic isometry f_l ($l \in \mathbb{C} - (i\mathbb{R} \cup \mathbb{R})$) and the "standard" translation f_l ($l \in \mathbb{R} - \{0\}$) are shown in Figure 12.6 and Figure 12.7. I also recommend checking out the website [Nel20].

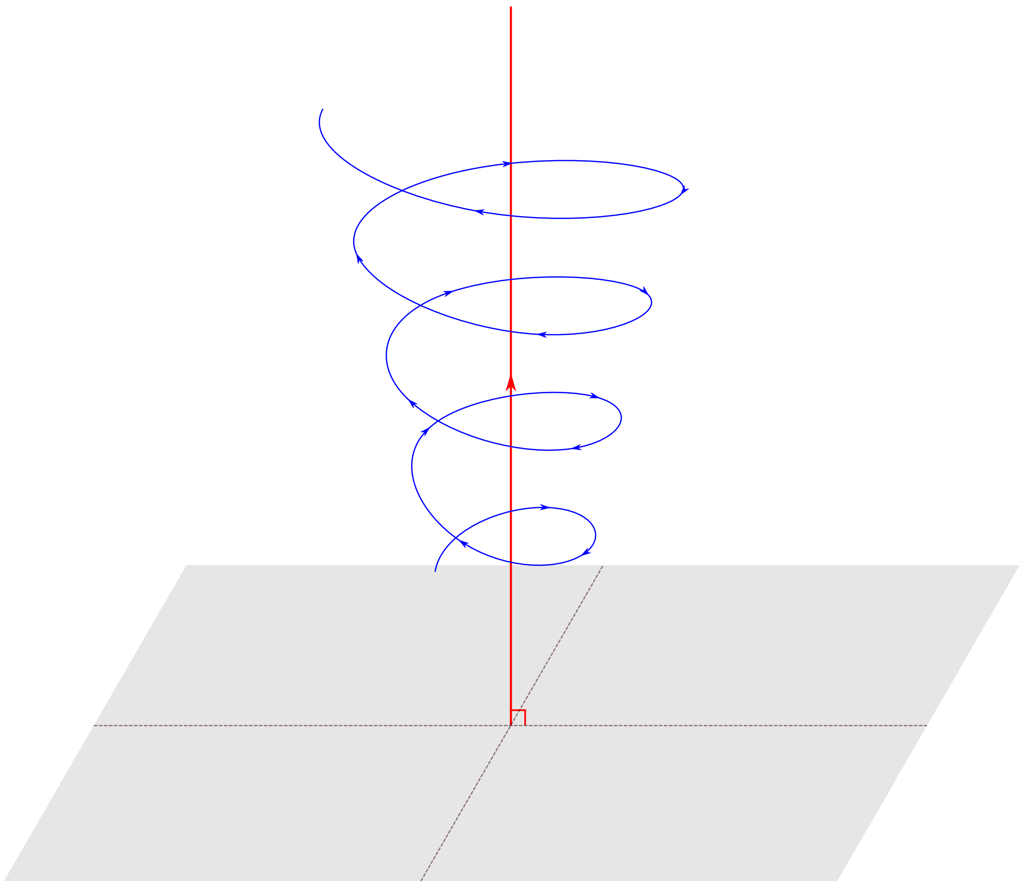


Figure 12.6: A loxodromic isometry of the Poincaré half-plane.

Corollary 12.39. *The conjugacy class of a loxodromic element of* $\mathrm{Isom}^+(H^3)$ *is uniquely determined by its trace (which is a complex number defined up to sign).*

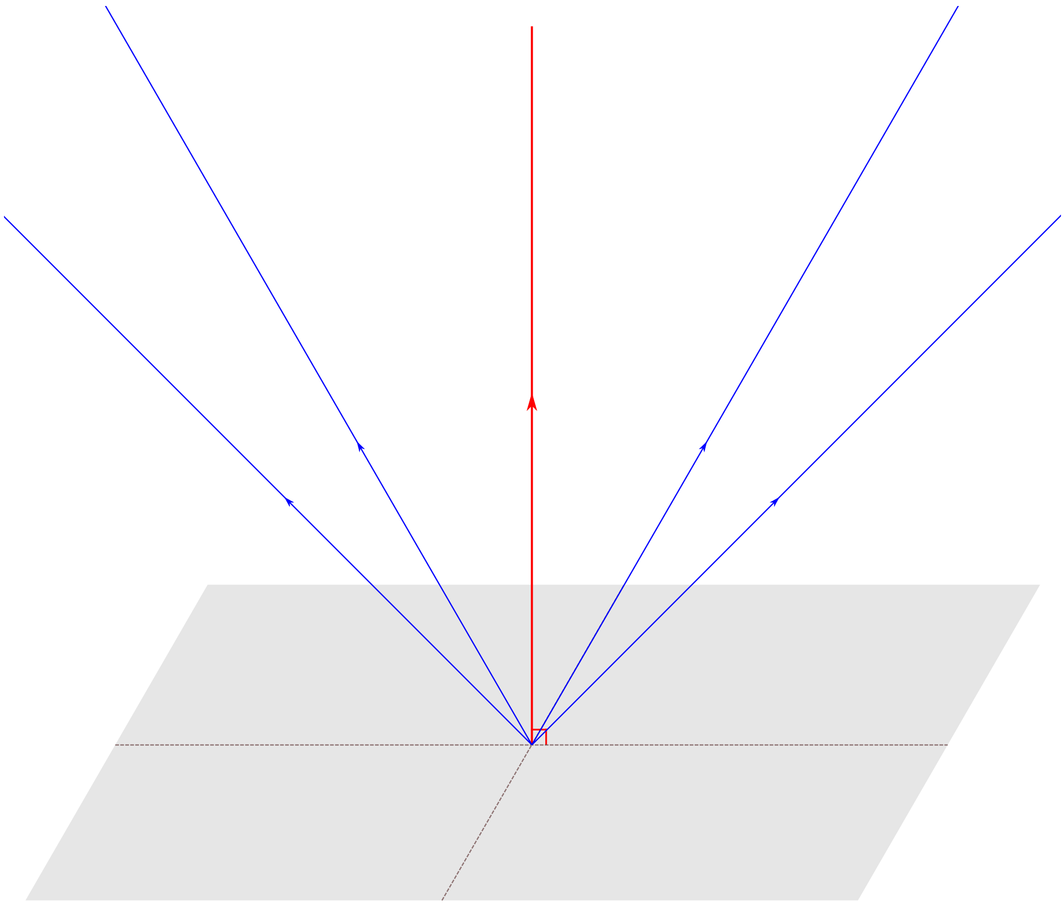


Figure 12.7: A translation of the Poincaré half-space.

12.4.4 Parabolic isometries

Theorem 12.40. *Let $f \in \mathrm{Isom}^+(H^3)$, represented by $M \in \mathrm{SL}(2, \mathbb{C})$. Assume $f \neq \mathrm{id}$. The following are equivalent:*

(i) *f is a parabolic isometry.*

(ii) *f has no fixed points in H^3, and one fixed point in $\partial_\infty H^3 = \hat{\mathbb{C}}$.*

(iii) $\mathrm{tr}\, M = \pm 2$.

(iv) *M is conjugate in $\mathrm{SL}(2, \mathbb{C})$ to $\pm P$, where $P = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$.*

(v) *f is conjugate in $\mathrm{Isom}^+(H^3)$ to $(z, x_3) \mapsto (z + 1, x_3)$.*

Proof. The fact that (i) ⇔ (ii) ⇔ (iii) is an application of Theorem 12.6 and Lemma 12.35. The proof of (iii) ⇔ (iv) is elementary linear algebra, it is the same as in Theorem 12.30. Finally, the proof of (iv) ⇔ (v) is the same as in Theorem 12.34. ∎

As in the two-dimensional case, there is only one conjugacy class of orientation-preserving parabolic isometries of $\mathbb{H}^3$:

Corollary 12.41. *Any parabolic element of* $\mathrm{Isom}^+(H^3)$ *has trace* ± 2*, and is conjugate to* $(z, x_3) \mapsto (z + 1, x_3)$*. Conversely, any* $f \in \mathrm{Isom}^+(H^3)$ *of trace* ± 2 *is parabolic,* provided $f \neq \mathrm{id}$.

A representation of the "standard" parabolic isometry $(z, x_3) \mapsto (z + 1, x_3)$ is shown in Figure 12.8.

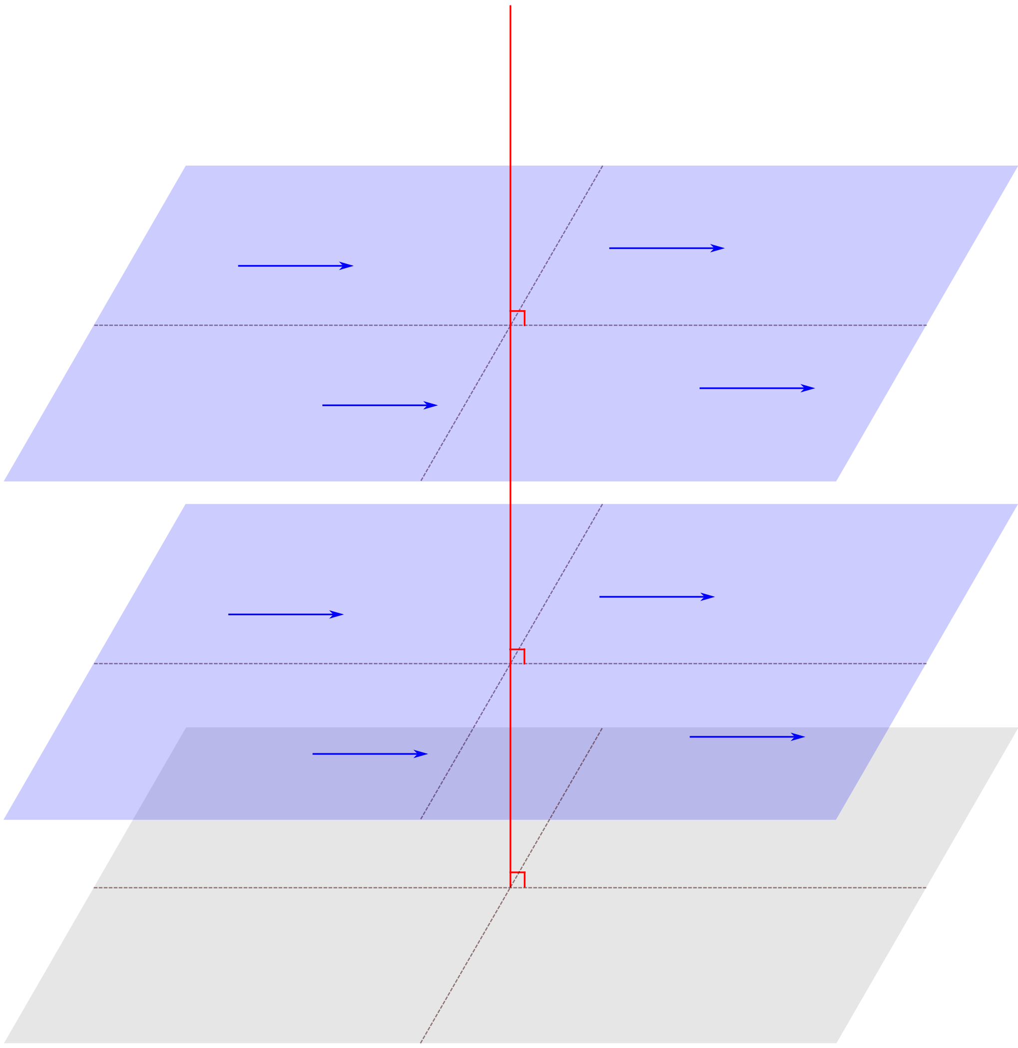


Figure 12.8: A parabolic isometry of the Poincaré half-space.

12.4.5 Conjugacy classes and trace

As a consequence of Corollary 12.36, Corollary 12.39, and Corollary 12.41, we obtain:

Theorem 12.42. *The conjugacy class of an element of $f \in \mathrm{Isom}^+(H^3) - \{\mathrm{id}\}$ is uniquely determined by its trace (which is a real number defined up to sign).*

More precisely, summarizing previous results:

- If $\mathrm{tr}(f) = \pm 2\cos(\frac{\theta}{2}) \in [2, 2]$, then f is elliptic and conjugate to $(z, x_3) \mapsto (e^{i\theta} z, x_3)$.
- If $\mathrm{tr}(f) = \pm 2\cosh(l/2) \in \mathbb{C} - [2, 2]$, then f is a translation and conjugate to $(z, x_3) \mapsto (e^{l} z, x_3)$.
- If $\mathrm{tr}(f) = \pm 2$ and $f \neq \mathrm{id}$, then f is parabolic and conjugate to $(z, x_3) \mapsto (e^{l} z, e^{\mathrm{Re}(l)} x_3)$.

Remark 12.43. Remark 12.33 also holds for H^3.

12.5 Exercises

Exercise 12.1.

Characterization of translation length (*borrowed from [BH99, Chap. II.6].*)

Let X be a metric space and let $f: X \to X$.

(1) Show that for any $x \in X$, the sequence $\frac{1}{n}d(x, f^n(x))$ converges in $[0, +\infty)$. *Hint: First show that $d(x, f^n(x))$ is a sub-additive function of n. Then show that $\frac{g(n)}{n}$ converges for any sub-additive function $g: \mathbb{N} \to \mathbb{R}$.*

(2) Show that $\lim_{n\to+\infty} \frac{1}{n}d(x, f^n(x))$ is independent of x.

(3) Show that if f is semi-simple (elliptic or hyperbolic), then $l_f = \lim_{n\to+\infty} \frac{1}{n}d(x, f^n(x))$.

Exercise 12.2.

Parabolic fixed point

Let f be a parabolic isometry of $X = \mathbb{H}^n$. Denote $\xi \in \partial_\infty X$ its ideal endpoint.

(1) Show that for any $x \in X \cup \partial_\infty X$, $\lim_{n\to+\infty} f^n(x) = \xi$. Is ξ an attracting fixed point?

(2) Show that for any compact set $K \subseteq \partial_\infty X - \{\xi\}$ and for any neighborhood U of ξ in $\partial_\infty X$, $f^n(K) \subseteq U$ for n sufficiently large. Is ξ an attracting fixed point?

Exercise 12.3.

Translation length of a parabolic

Let f be a parabolic isometry of $X = \mathbb{H}^n$. Show that f has zero translation length.

Exercise 12.4.

Equidistant curves and translations

(1) Let $L \subseteq \mathbb{H}^n$ be a geodesic line. How would you define an equidistant curve from L? Show that for any $x_0 \in \mathbb{H}^n$, there exists a unique equidistant curve from L.

(2) Let L be the geodesic line with ideal endpoints 0 and ∞ in the Poincaré half-space H^n. Show that the equidistant curves from L are the Euclidean straight half-lines starting from 0.

(3) Prove Proposition 12.18: $f\colon \mathbb{H}^n \to \mathbb{H}^n$ is a translation if and only if there exists an isometry $\varphi\colon \mathbb{H}^n \to H^n$ such that $\varphi f \varphi^-1$ is $x \in H^n \mapsto e^l x$, where l is the translation length of f.

Exercise 12.5.

Fixed points and trace

Recall Lemma 12.24: Let $M = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in \mathrm{SL}(2,\mathbb{R})$ and denote $f\colon z \mapsto \frac{az+b}{cz+d}$ the associated fractional linear transformation of $\hat{\mathbb{C}}$.

- If $(\operatorname{tr} M)^2 > 4$, then f has two fixed points, both of which lie in $\hat{\mathbb{R}} \subseteq \hat{\mathbb{C}}$.
- If $(\operatorname{tr} M)^2 < 4$, then f has two fixed points, one in $\mathbb{H}$ and the other is its complex conjugate.
- If $(\operatorname{tr} M)^2 = 4$, then either f is the identity, or f has a unique fixed point, which lies in $\hat{\mathbb{R}} \subseteq \hat{\mathbb{C}}$.

(1) Prove the lemma by direct computation, solving the equation $\frac{az+b}{cz+d} = z$.

(2) Consider the projective transformation $\hat{f}\colon \mathbb{C}P^1 \to \mathbb{C}P^1$ associated to M. Explain why the fixed points of $\hat{f}$ are the eigenlines of M. Recover the lemma.

Exercise 12.6.

Limits of loxodromics

(1) Recall the "standard form" of orientation-preserving elliptic, loxodromic, and parabolic isometries of $\mathbb{H}^3$ in the Poincaré half-space model.

(2) Using the previous question, show that any elliptic element of $\mathrm{Isom}^+(\mathbb{H}^3)$ can be obtained as a limit of loxodromic elements.

(3) Prove more generally that any elliptic isometry of $\mathbb{H}^n$ can be obtained as a limit of loxodromic isometries.

(4) Going back to $\mathbb{H}^3$, write a different proof using matrices. Prove in fact that loxodromic elements are dense in $\mathrm{Isom}^+(\mathbb{H}^3)$.

Exercise 12.7.

A baby character variety

Let us work in the Poincaré half-space model $\mathbb{H} \subseteq \mathbb{C}$ of the hyperbolic plane $\mathbb{H}^2$. We denote $G = \mathrm{Isom}^+(\mathbb{H})$ the group of orientation-preserving isometries, which can be identified to $\mathrm{PSL}(2, \mathbb{R}) = \mathrm{SL}(2, \mathbb{R})/\{\pm I_2\}$ equipped with the quotient topology.

(1) Show that $f_0 = \mathrm{id} \in G$ is in the closure of the conjugacy class $C \subseteq \mathrm{Isom}^+(\mathbb{H})$ of some/any parabolic isometry.

(2) Let G act on itself by conjugation. Derive from the previous question that the quotient $\mathcal{R}$ is not Hausdorff.

(3) (*) We recall that an element of G is called ***semisimple*** (or *completely reducible*, or *polystable*, depending on context) if it is not parabolic. Let $\mathcal{X} \subseteq \mathcal{R}$ denote the subset of conjugacy classes of semisimple elements. Show that $\mathcal{X}$ is Hausdorff.

Exercise 12.8.

Trace relations

We let $G = \mathrm{SL}(2, \mathbb{C})$ in this exercise.

(1) Show that for any $A, B \in G$, $\mathrm{tr}(AB) + \mathrm{tr}(AB^{-1}) = \mathrm{tr}\, A \,\mathrm{tr}\, B$.

(2) Show that the trace of any element of the subgroup of G generated by A and B can be expressed as a polynomial in $\mathrm{tr}\, A$, $\mathrm{tr}\, B$, and $\mathrm{tr}\, AB$ with integer coefficients.

(3) *Optional.* Show that any polynomial function of $(A, B) \in G \times G$ that is invariant by conjugation (that is, invariant by $(A, B) \mapsto (gAg^{-1}, gBg^{-1})$ for all $g \in G$) can be expressed as a polynomial function of $\mathrm{tr}\, A$, $\mathrm{tr}\, B$, and $\mathrm{tr}\, AB$.

Exercise 12.9.

Classification in $\mathrm{O}^+(n, 1)$

Recall that $\mathrm{Isom}(\mathbb{H}^n) \approx \mathrm{O}^+(n, 1)$, e.g. via the hyperboloid model. Using linear algebra, find a characterization of elliptic, loxodromic, and parabolic elements of $\mathrm{O}^+(n, 1)$.

Part VI

Plane hyperbolic geometry

[…] the way in which I have proceeded does not lead to the desired goal, the goal that you declare you have reached, but instead to a doubt of the validity of [Euclidean] geometry. I have certainly achieved results which most people would look upon as proof, but which in my eyes prove almost nothing; if, for example, one can prove that there exists a right triangle whose area is greater than any given number, then I am able to establish the entire system of [Euclidean] geometry with complete rigor. Most people would certainly set forth this theorem as an axiom; I do not do so, though certainly it may be possible that, no matter how far apart one chooses the vertices of a triangle, the triangle's area still stays within a finite bound. I am in possession of several theorems of this sort, but none of them satisfy me.

– Carl Friedrich Gauß[5]

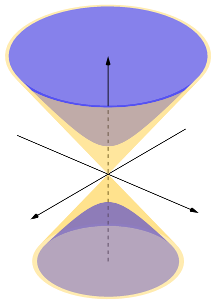
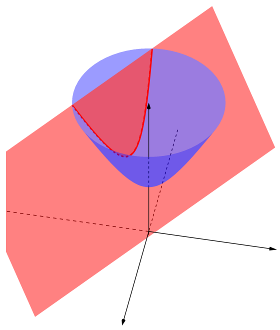
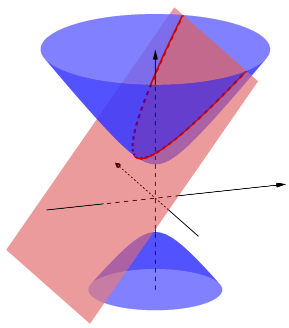
[5] 1799, Answer to a letter from Farkas Bolyai in which Bolyai claimed to have proved Euclid's fifth postulate.

CHAPTER 13

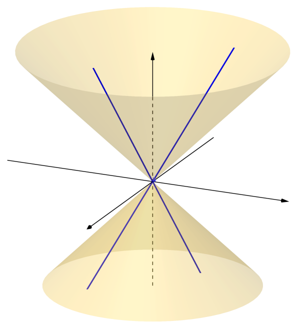
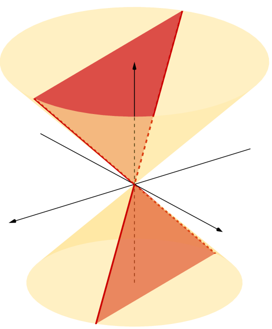
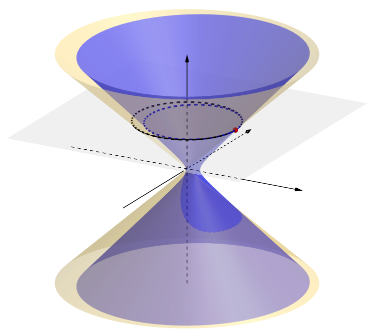
Recap of 2D models

This is not a chapter.

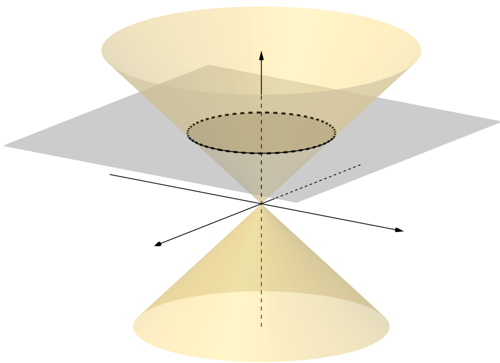
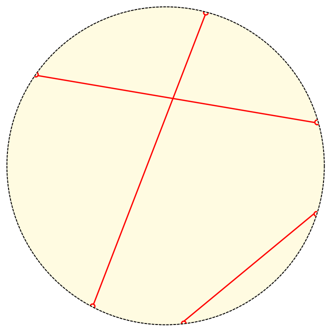
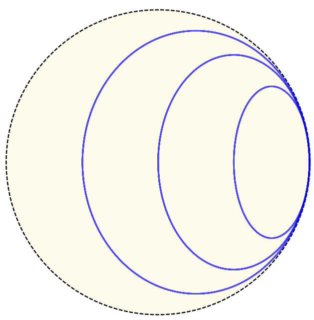
13.1 Hyperboloid model

Name	**Hyperboloid**
Definition	$\mathcal{H}^+ = \mathcal{H} \cap \{z > 0\}$ is the upper sheet of the hyperboloid $\mathcal{H} = \left\{(x, y, z) \mid x^2 + y^2 - z^2 = -1\right\}$ $= \{p \in \mathbb{R}^{2,1} \mid \langle p, p\rangle = -1\} \subseteq \mathbb{R}^{2,1}$ 
Riem. metric	$\mathrm{d}s^2 = \mathrm{d}x^2 + \mathrm{d}y^2 - \mathrm{d}z^2$ (restricted to the tangent plane to $\mathcal{H}^+$, given by $\mathrm{T}_p\,\mathcal{H}^+ = \{p\}^\perp$)
Distance	$d(p, q) = \measuredangle(p, q)$ (hyperbolic angle) i.e. $d(p, q) = \operatorname{arcosh}(-\langle p, q\rangle)$
Geodesics	Hyperbolas $\gamma = \mathcal{H}^+ \cap P$ where P is a vector plane in $\mathbb{R}^{2,1}$ They are nicely parametrized: $\gamma(t) = \cosh(\lVert v\rVert t)p + \sinh(\lVert v\rVert t)v$ 
Isometries	$\mathrm{O}^+(2, 1)$ acting linearly on $\mathbb{R}^{2,1}$ (in restriction to $\mathcal{H}^+$) Orientation-preserving isometries: $\mathrm{SO}^+(2, 1)$ (= $\mathrm{O}_0(2, 1)$)
Curvature	$K \equiv -1$
Ideal boundary	The hyperboloid model is not best suited to see the ideal boundary. It can be described as the set of future-directed isotropic half-lines in $\mathbb{R}^{2,1}$, which is essentially the same as the projectivized light cone (see § 13.2).
Horocycles	Parabolas $C = \mathcal{H}^+ \cap P$ where P is an affine plane with isotropic normal (See Exercise 5.4, Exercise 11.6) 

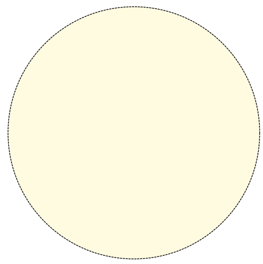
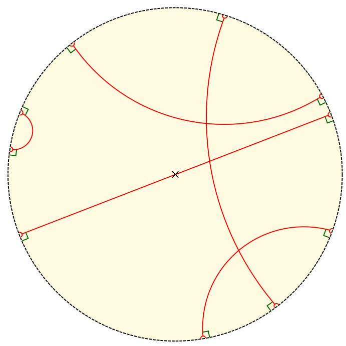
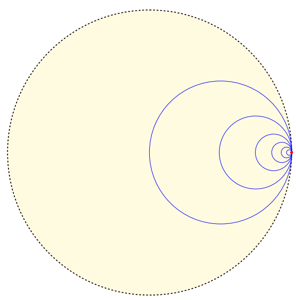
13.2 Cayley–Klein model

Name	**Cayley–Klein model** or **Projective model**
Definition	$\Omega^- = \mathbf{P}\left(\{q<0\}\right) \subseteq \mathbf{P}\left(\mathbb{R}^{2,1}\right)$ i.e. $\Omega^- = \{\text{lines inside the light cone}\}$ 
Riem. metric	(See § 13.3 for expression in affine chart)
Distance	$d(p,q) = \frac{1}{2}\ln[p,q,J,I]$ i.e. $d([u],[v]) = \operatorname{arcosh}\left(\frac{-b(u,v)}{\sqrt{q(u)q(v)}}\right)$
Geodesics	Projective lines $l \subseteq \mathbf{P}(\mathbb{R}^{2,1})$ intersected with Ω^- (i.e. *chords* in Ω^-) 
Isometries	$\mathrm{PO}(2,1)$ acting projective linearly on $\mathbf{P}(\mathbb{R}^{2,1})$ (in restriction to Ω^-) Orientation-preserving isometries: $\mathrm{PSO}(2,1)$
Curvature	$K \equiv -1$
Ideal boundary	$\partial\Omega^- = \mathbf{P}\left(\{q=0\}\right) \subseteq \mathbf{P}\left(\mathbb{R}^{2,1}\right)$ (projectivized light cone) Note: this is a circle (more precisely, a projective ellipse)
Horocycles	Projective ellipses tangent to $\partial\mathbb{D}$ to order 2 (See Exercise 11.7, § 13.3) 

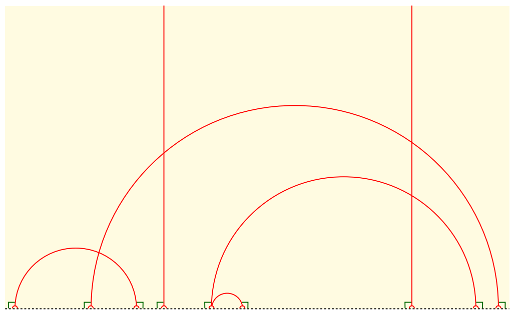
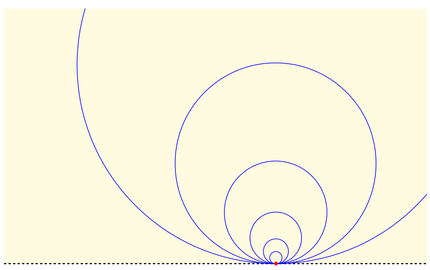
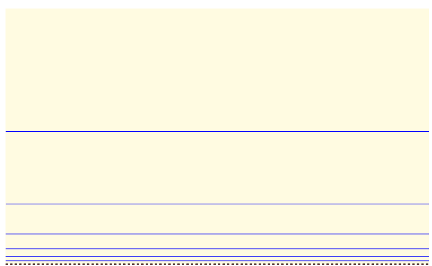
13.3 Beltrami–Klein model

Name	Beltrami–Klein disk or Klein disk
Definition	$\{x^2+y^2-z^2<0\}\cap\{z=1\}\subseteq\mathbb{R}^3$ $\approx\mathbb{D}=\{x^2+y^2<1\}\subseteq\mathbb{R}^2$ 
Riem. metric	$ds^2=\dfrac{dx^2+dy^2}{1-x^2-y^2}+\dfrac{(x\,dx+y\,dy)^2}{(1-x^2-y^2)^2}$
Distance	$d(p,q)=\frac{1}{2}\ln[p,q,J,I]$ i.e. $d(p,q)=\operatorname{arcosh}\left(\dfrac{1-\langle p,q\rangle}{\sqrt{(1-\Vert p\Vert^2)(1-\Vert q\Vert^2)}}\right)$
Geodesics	Chords in $\mathbb{D}$ 
Isometries	$\mathrm{PO}(2,1)$ acting by fractional linear transformations on $\mathbb{R}^2$ Orientation-preserving isometries: $\mathrm{PSO}(2,1)$
Curvature	$K\equiv-1$
Ideal boundary	$\partial_\infty\mathbb{D}=\partial\mathbb{D}=\{\lvert z\rvert=1\}$
Horocycles	Ellipses tangent to $\partial\mathbb{D}$ to order 2 (See Exercise 11.7) 

13.4 Poincaré disk model

Name	**Poincaré disk**
Definition	$\mathbb{D} = \left\{(x,y) \mid x^2+y^2<1\right\} \subseteq \mathbb{R}^2$ $= \{z \in \mathbb{C} \mid \lvert z\rvert < 1\} \subseteq \mathbb{C}$ 
Riem. metric	$\mathrm{d}s^2 = 4\dfrac{\mathrm{d}x^2+\mathrm{d}y^2}{(1-x^2-y^2)^2} = 4\dfrac{\lvert\mathrm{d}z\rvert^2}{(1-\lvert z\rvert^2)^2}$
Distance	$d(z_1,z_2) = \operatorname{arcosh}\left(1+\dfrac{2\lvert z_1-z_2\rvert^2}{(1-\lvert z_1\rvert^2)(1-\lvert z_2\rvert^2)}\right) = \ln[z_1,z_2,J,I]$
Geodesics	Circle arcs $\perp$ to $\partial\mathbb{D}$ (including diameters) 
Isometries	$\mathrm{PSU}(1,1) = \mathrm{PU}(1,1)$ acting by fractional linear transformations: $z \mapsto \dfrac{az+b}{\bar{b}z+\bar{a}}$ with $\lvert a\rvert^2 - \lvert b\rvert^2 = 1$ i.e. $z \mapsto u\dfrac{z-a}{1-\bar{a}z}$ with $\lvert u\rvert = 1$, $\lvert a\rvert < 1$ (for orientation-reversing isometries, replace z by $\bar{z}$)
Curvature	$K \equiv -1$
Ideal boundary	$\partial_\infty\mathbb{D} = \partial\mathbb{D} = \{\lvert z\rvert = 1\}$
Horocycles	Euclidean circles tangent to $\partial\mathbb{D}$ 

13.5 Poincaré half-plane model

Name	**Poincaré half-plane**
Definition	$\mathbb{H} = \{(x,y) \mid y > 0\} \subseteq \mathbb{R}^2$ $= \{z \in \mathbb{C} \mid \operatorname{Im} z > 0\} \subseteq \mathbb{C}$
Riem. metric	$ds^2 = \dfrac{dx^2 + dy^2}{y^2} = \dfrac{\lvert dz\rvert^2}{(\operatorname{Im} z)^2}$
Distance	$d(z_1, z_2) = \operatorname{arcosh}\left(1 + \dfrac{\lvert z_1 - z_2\rvert^2}{2y_1y_2}\right) = \ln[z_1, z_2, J, I]$
Geodesics	Circle arcs $\perp$ to $\partial\mathbb{H}$ (including vertical lines) 
Isometries	$\mathrm{PSL}(2,\mathbb{R})$ acting by fractional linear transformations: $z \mapsto \dfrac{az+b}{cz+d}$ with $a,b,c,d \in \mathbb{R}$, $ad - bc = 1$ (for orientation-reversing isometries, replace z by $-\bar{z}$)
Curvature	$K \equiv -1$
Ideal boundary	$\partial_\infty\mathbb{H} = \hat{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$
Horocycles	  Euclidean circles tangent to $\partial_\infty\mathbb{H}$ (including horizontal lines)

13.6 Exercises

Exercise 13.1.

Comparison of 2D models

Discuss the advantages and disadvantages of each of the 2-dimensional models. Do you have a favorite?

CHAPTER 14

Hyperbolic trigonometry

Disclaimer: This chapter is a draft.

Trigonometry, in the literal sense[1], is the study of measurements in triangles, especially the relations between side lengths and angles at the vertices. Such relations are fundamental because not only they are inherent to the geometry of the "universe" (e.g. Euclidean, spherical, or hyperbolic space), they completely characterize it.

After reviewing the basics of triangles in the hyperbolic plane, we shall see that relations between sides and angles are incarnated by the *hyperbolic law of cosines*. Two direct applications of this formula set hyperbolic geometry uniquely apart from Euclidean geometry: the fact that two triangles with the same angles are congruent, and the notion of angle of parallelism. Next, we turn to the strikingly simple relation between the area of a hyperbolic triangle and the sum of its interior angles. This is a trivial consequence of the Gauss—Bonnet theorem, but we present an elegant alternative proof, also due to Gauss. We conclude the chapter by showing that $\mathbb{H}^2$ is a hyperbolic metric space in the sense of Gromov, a feature that we have used in Chapter 11. It will appear in these discussions that the notion of *ideal triangle*, i.e. triangle with vertices are "at infinity", is very useful in hyperbolic geometry.

A chapter on hyperbolic trigonometry could very well be the first in a course of hyperbolic geometry. It is therefore somewhat amusing (or suspicious!) that it arrives so late in our presentation[2]. This is a consequence of my decision to go for a "clean and modern" presentation of hyperbolic geometry, which assumes notions of Riemannian geometry, Minkowski spaces,

[1]The word *trigonometry* is derived from the Greek $\tau\rho\tilde{\iota}\gamma o\nu o\nu$ (*trigōnon*), triangle, and $\mu\acute{\epsilon}\tau\rho o\nu$ (*metron*), measure.

[2]This is the last chapter of the course that I taught at TU Darmstadt, although I initially planned for two additional chapters: the next would contain more plane hyperbolic geometry, including tessellations of $\mathbb{H}^2$, and the final chapter would discuss hyperbolic structures on surfaces.

projective geometry and Möbius transformations. Among the benefits of this approach, beyond the fact that all the theorems of this chapter will be given concise and rigorous proofs, it will be elegant and effective to juggle the different models of the hyperbolic plane. For instance, the hyperbolic law of cosines is easily derived from the hyperboloid model, while the dual law of cosines can be understood in the Klein model via projective duality, and the Poincaré models are best suited to study ideal triangles and compute areas.

14.1 Hyperbolic triangles

In the whole chapter, let $\mathbb{H}^2$ denote the hyperbolic plane, not favoring a particular model unless otherwise stated.

Basic definitions

By definition, a ***hyperbolic triangle*** consists of three points typically denoted A, B, C, the ***vertices***, and the three geodesic segments between them, denoted AB, BC, CA, the ***sides*** (or ***edges***). We allow ***degenerate*** triangles, where the three vertices are collinear (lie on a geodesic), including the cases where two or three vertices are equal. We typically denote the ***side lengths*** by $a = d(B, C)$, $b = d(C, A)$, $c = d(A, B)$, and the ***interior angles*** by $\hat{A}$, $\hat{B}$, $\hat{C}$, i.e. the unoriented angles between the sides[3]. See Figure 14.1.

Remark 14.1. If two of the vertices are equal, say $A = B$, then the angles $\hat{A}$ and $\hat{B}$ are undefined (and $\hat{C} = 0$, unless $C = A = B$). In the rest of this chapter, any identity involving $\hat{A}$ implicitly assumes that it only applies to triangles where A is distinct from B and C.

As in the Euclidean plane, we call ***right triangle*** a triangle that has a right angle, ***isoceles triangle*** a triangle that has two equal side lengths, etc.

Congruent triangles

Two triangles are called ***congruent*** if there exists an isometry that takes one to the other. It is enough to require that the isometry maps the vertices of the first triangle to the vertices of the second; such an isometry automatically maps the sides to the sides. It is clear that congruence is an equivalence relation on the set of hyperbolic triangles.

Theorem 14.2. *Given $a, b, c \in [0, +\infty)$, there exists a hyperbolic triangle with side lengths a, b, c if and only if the triangle inequalities $a \leqslant b + c$, $b \leqslant c + a$, $c \leqslant a + b$ are satisfied. Moreover, any two hyperbolic triangles are congruent if and only if they have the same side lengths.*

[3] Let us recall that the angle between two intersecting geodesics, in fact between any two intersecting curves, is defined as the angle between their tangent vectors at the intersection. This works in any Riemannian manifold, as we saw in § 9.1.

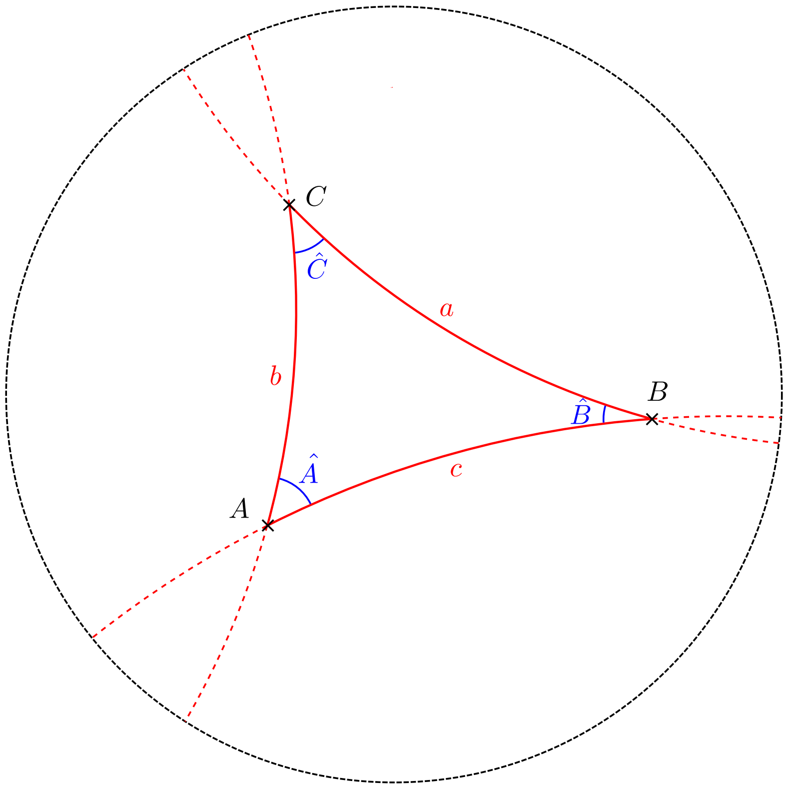


Figure 14.1: A typical hyperbolic triangle in the Poincaré disk model.

Proof. It is clear that if there exists a triangle with side lengths a, b, c, then the triangle inequalities are satisfied. This is because $\mathbb{H}^2$ is a genuine metric space, as is any connected Riemannian manifold with the induced distance.

Conversely, assume that a, b, c satisfy the triangle inequalities. Let us show both the existence of a triangle ABC with side lengths a, b, c and its uniqueness up to congruence at the same time. We shall work in the Poincaré disk model $\mathbb{D} \subseteq \mathbb{C}$ of the hyperbolic plane. First choose the position of A and B in $\mathbb{D}$ so that $d(A, B) = c$. After applying a translation, we can assume that A is the origin $0 \in \mathbb{D}$, and after applying a rotation we can assume that B lies on a the ray $[0, 1)$. It is clear that under these conditions, the position of B is completely determined by the condition $d(A, B) = c$.

Now let us look for the position of C. After applying the reflection $z \mapsto \bar{z}$ if necessary, we can assume that $\mathrm{Im}(C) \geqslant 0$. Let us show that the position of C is now completely determined by $d(A, C) = b$ and $d(B, C) = a$. In other words, we need to show that the circles $C(A, b)$ and $C(B, a)$ have a unique point of intersection C with $\mathrm{Im}(C) \geqslant 0$. These two circles are Euclidean circles by Lemma 14.3, whose centers lie on the same line as A and B.

There is a limited number of configurations of two circles in the Euclidean plane: either one is contained in the interior of the other, or they are each contained in the exterior of the other, or they intersect in two (possibly equal) points. In our situation, $a \leqslant b + c$ and $b \leqslant a + c$ rule out the first configuration, and $c \leqslant a + b$ rules out the second configuration. Therefore we must be in the third configuration where the circles intersect. Moreover, their two (possibly equal) points of intersection are symmetric with respect to the line $(-1, 1)$, due to the invariance of our configuration under the isometry $z \mapsto \bar{z}$. The conclusion follows. ∎

Lemma 14.3. *Let $C = C(A, R)$ denote the circle with center A and radius $R \geqslant 0$ in the Poincaré disk, i.e. the set of points in $\mathbb{D}$ at distance R from A. Then C is a Euclidean circle.*

Remark 14.4. Be careful: the Euclidean center of C is different from A. That is, unless $A = 0$. Also, the Euclidean radius of C is different from R.

Proof of Lemma 14.3. If $A = 0$, it is easy to see from the expression of the hyperbolic distance that C is a Euclidean circle centered at 0 (and with Euclidean radius $r = \tanh(R/2)$). If $A \neq 0$, one can always use an isometry to send A to 0 (isometries act transitively). The conclusion follows from the fact that isometries of $\mathbb{D}$ are Möbius transformations, and Möbius transformations map Euclidean circles to Euclidean circles. ∎

Triangles with ideal vertices

It is convenient to allow hyperbolic triangles to have one or more ideal vertices. For instance, if $A \in \partial_\infty \mathbb{H}^2$ is an ideal point and $B, C \in \mathbb{H}^2$ are “interior” points, the triangle ABC still consists of the three vertices A, B, C and the three sides AB, BC, CA; however now the sides AB and CA are semi-infinite geodesic lines (i.e. geodesic rays) with ideal endpoint A. A triangle with 1 ideal vertex is called ***1/3 ideal***. Similarly, we have obvious definitions of triangles having 2 ideal vertices, called ***2/3 ideal triangles***, and 3 ideal vertices, called ***ideal triangles***. An ideal triangle is shown in Figure 14.2 (and another in Figure 14.3).

Clearly, there is only one sensible (i.e. continuous) to extend the notion of side length and interior angles for triangles with one or more ideal vertices: the sides adjacent to an ideal vertex have side length $+\infty$, and the interior angle at an ideal vertex is zero. Indeed, in the Poincaré disk model, recall that angles in $\mathbb{H}^2$ are equal to Euclidean angles (the Poincaré disk model is conformal). At an ideal vertex, the two adjacent sides are both orthogonal to the boundary, therefore the Euclidean angle between them is zero.

14.2 The hyperbolic law of cosines

Review: the Euclidean case

Before we jump to the hyperbolic law of cosines, let us quickly review the Euclidean case. A good starting point is the celebrated Pythagorean theorem: a triangle ABC has a right angle at C if and only if we have the identity $c^2 = a^2 + b^2$. The Euclidean law of cosines is a generalization:

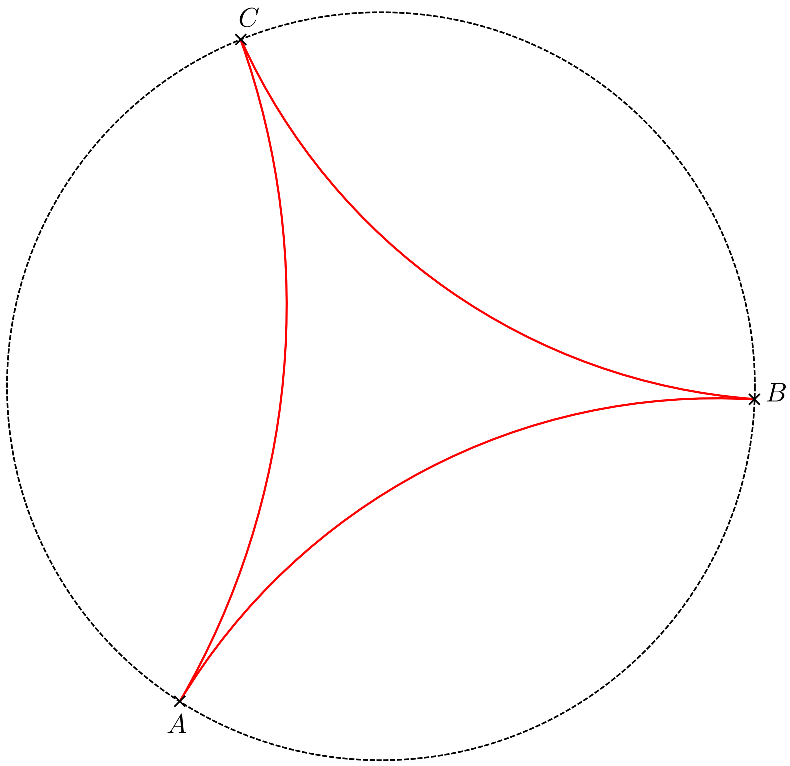


Figure 14.2: An ideal triangle in the Poincaré disk model.

Theorem 14.5 (Euclidean law of cosines). *For any triangle ABC, with angles denoted $\hat{A}$, $\hat{B}$, $\hat{C}$ and opposite side lengths a, b, c, we have:*

$$c^2 = a^2 + b^2 - 2ab\cos\hat{C}\ .$$

Proof. There are many proofs of the Euclidean law of cosines. A modern proof with vector calculus is elementary: starting with $c = \|\overrightarrow{AB}\|$, we have

$$\begin{aligned} c^2 &= \|\overrightarrow{AC} + \overrightarrow{CB}\|^2 \\ &= \|\overrightarrow{AC}\|^2 + \|\overrightarrow{CB}\|^2 + 2\langle\overrightarrow{AC}, \overrightarrow{CB}\rangle\ . \end{aligned}$$

The conclusion follows, since $\langle\overrightarrow{AC}, \overrightarrow{CB}\rangle = -\langle\overrightarrow{CA}, \overrightarrow{CB}\rangle = -ba\cos\hat{C}$. ∎

Remark 14.6. The Euclidean law of cosines is also known as *Al-Kashi's theorem* (for instance, this is the name that I learned as a high-schooler in France in the early 2000s), after the Persian mathematician Jamshid al-Kashi who proved the theorem in 1427[4]. It must be noted an

[4] It is contained in Al Kashi's main mathematical work, *Miftāḥ al-Ḥisab* (*Key to Arithmetic*). This work, which consists of five books, is recently being translated to English with commentary: [AH19].

equivalent version of this theorem is proved in Euclid's *Elements*[5] (3rd century BC), although without using trigonometric functions.

Next we have the law of sines. First, for a triangle ABC with a right angle at C, we have

$$\sin \hat{A} = \frac{a}{c} \qquad \sin \hat{B} = \frac{b}{c}$$

so that $\frac{\sin \hat{A}}{a} = \frac{\sin \hat{B}}{b} = \frac{1}{c}$. More generally, the ***law of sines*** says that for any triangle ABC, we have

$$\frac{\sin \hat{A}}{a} = \frac{\sin \hat{B}}{b} = \frac{\sin \hat{C}}{c} .$$

We leave the proof of the law of sines as an exercise of elementary Euclidean geometry.

Remark 14.7. It can be useful to memorize the additional equalities:

$$\frac{\sin \hat{A}}{a} = \frac{\sin \hat{B}}{b} = \frac{\sin \hat{C}}{b} = \frac{1}{2R} = \frac{2S}{abc}$$

where R is the radius of the circumscribed circle and S is the area of the triangle.

Hyperbolic law of cosines, dual law of cosines, and law of sines

Let us go back to triangles the hyperbolic plane $\mathbb{H}^2$.

Theorem 14.8. *For any hyperbolic triangle ABC with $C \neq A$ and $C \neq B$, we have the **hyperbolic law of cosines**:*

$$\cosh c = \cosh a \cosh b - \sinh a \sinh b \cos \hat{C} . \tag{14.1}$$

Proof. It is easiest to prove the hyperbolic law of cosines in the hyperboloid model $\mathcal{H}^+ \subseteq \mathbb{R}^{2,1}$. Let γ_u [resp. γ_v] be the unit geodesic from C to A [resp. from C to B]. Our notation indicates that u [resp. v] is the initial tangent vector to the geodesic. Since $\gamma_u(t)$ is a length-minimizing geodesic parametrized by arclength, it reaches A when $t = d(C, A)$, in other words: $A = \gamma_u(b)$. For the same reason, $B = \gamma_v(a)$. Given the explicit expression of geodesics in the hyperboloid model (see Theorem 5.8), namely $\gamma_u(t) = (\cosh t)C + (\sinh t)u$, we find that:

$$\begin{aligned} A &= (\cosh b)C + (\sinh b)u \\ B &= (\cosh a)C + (\sinh a)v . \end{aligned}$$

On the other hand, we have $\cosh c = \cosh d(A, B) = -\langle A, B\rangle$ where $\langle \cdot , \cdot \rangle$ indicates the inner product in Minkowski space $\mathbb{R}^{2,1}$ (see Theorem 5.12). Substituting the expressions of A and B above, we find:

$$\cosh c = - \langle (\cosh b)C + (\sinh b)u \, , \, (\cosh a)C + (\sinh a)v \rangle \; .$$

[5][Euc56, Book 2, Propositions 12 and 13].

Now simply expand this inner product, noticing that: $\langle C, C\rangle = -1$ since C is on the hyperboloid, $\langle C, u\rangle = \langle C, v\rangle = 0$ since u and v are tangent vectors at C, and $\langle u, v\rangle = \cos\hat{C}$ by definition of the angle at $\hat{C}$. What comes out is the desired identity. ■

Next we have the dual hyperbolic law of cosines and the hyperbolic law of sines:

Theorem 14.9. *For any hyperbolic triangle ABC, we have the* ***dual hyperbolic law of cosines****:*

$$\cos\hat{C} = -\cos\hat{A}\cos\hat{B} + \sin\hat{A}\sin\hat{B}\cosh c\,.$$

And the ***hyperbolic law of sines****:*

$$\frac{\sin\hat{A}}{\sinh a} = \frac{\sin\hat{B}}{\sinh b} = \frac{\sin\hat{C}}{\sinh c}\,.$$

Proof. The dual hyperbolic law of cosines can be derived from the three hyperbolic law of cosines in the triangle ABC and basic calculus; we leave out the details. Alternatively, one can derive it from the hyperbolic law of cosines (14.1) through projective duality in the Cayley–Klein model, see remark below. The reader may also refer to [Thu97, Chap. 2.4] for a different proof.

To prove the hyperbolic law of sines, first assume that the triangle ABC has a right angle at C. By the law of cosines, we have

$$\cosh b = \cosh a\cosh c - \sinh a\sinh c\cos\hat{B}\,.$$

Substituting $\cosh c = \cosh a\cosh b$ (by the law of cosines) and $\cos\hat{B} = \cosh b\sin\hat{A}$ (by the dual law of cosines), we find that $1 = \cosh^2 a - \sinh a\sinh c\sin\hat{A}$, therefore

$$\sin\hat{A}\sinh c = \sinh a\,.$$

Now for a generic triangle ABC, let H be the orthogonal projection of C on the geodesic line AB. Applying the previous identity in the triangles AHC and BHC, we find

$$\sin\hat{A}\sinh b = \sinh h = \sin\hat{B}\sinh a$$

where $h = d(H, C)$. In particular, we have $\frac{\sin\hat{A}}{\sinh a} = \frac{\sin\hat{B}}{\sinh b}$ as desired. The same argument can be repeated after relabeling the vertices A, B, C, so the second equality follows. ■

Remark 14.10. The most elegant proof of the dual hyperbolic law of cosines is through projective duality in the Cayley–Klein model: essentially, distances between points in $\mathbb{H}^2$ corresponds to angles between lines under projective duality, and the dual law of cosines is nothing more than the law of cosines in the projective dual. Making this argument rigorous is a great exercise, but it turns out to be a bit tricky. The subtlety is that projective duality sends lines contained in $\mathbb{H}^2$ (i.e. secant to the quadric $\mathcal{Q}$) to points *outside* the dual conic $\mathcal{Q}^*$ (such points are sometimes called ***ultra-ideal***). Nevertheless, the Cayley–Klein metric is still defined outside of $\mathcal{Q}^*$, and the fact that it is purely imaginary (before taking the absolute value) accounts for the presence of regular cosines and sines in the dual law instead of hyperbolic cosines and sines[6].

[6]Check out [MvG] for an equivalent explanation, which is more detailed but not completely clean in my

Consequences

It is easy to derive many formulas in hyperbolic triangles from the law of cosines, the dual law of cosines, and the law of sines. For instance, the ***hyperbolic Pythagorean theorem*** reads: if ABC has a right angle at C, then $\cosh c = \cosh a \cosh b$.

Remark 14.11. Observe that the second-order expansion of the hyperbolic Pythagorean theorem $\cosh c = \cosh a \cosh b$ is $c^2 = a^2 + b^2$, i.e. the Euclidean Pythagorean theorem. This is not a coincidence: informally speaking, small hyperbolic triangles look almost Euclidean. More generally, hyperbolic geometry limits to Euclidean geometry on a small scale. It is a good exercise to make a precise interpretation of this statement.

Still assuming that ABC has a right angle at C, the sine and cosine of the angle at A can be computed as:

$$\sin \hat{A} = \frac{\sinh a}{\sinh c}$$
$$\cos \hat{A} = \frac{\tanh b}{\tanh c}\,.$$

Let us spare all these silly calculations.

One very interesting consequence of the hyperbolic law of cosines is the following:

Theorem 14.12. *The congruence class of a hyperbolic triangle with distinct vertices is uniquely determined by its interior angles.*

Proof. It follows from the dual law of cosines that the three side lengths a, b, c, are uniquely determined by the three angles $\hat{A}$, $\hat{B}$, $\hat{C}$. Conclude with Theorem 14.2. ∎

We leave it as an exercise that conversely, given any three numbers $\alpha, \beta, \gamma \geqslant 0$ such that $\alpha + \beta + \gamma < \pi$, there exists a hyperbolic triangle whose interior angles are equal to α, β, γ: see Exercise 14.2.

Remark 14.13. It is important to realize that Theorem 14.12 this is drastically different from the Euclidean situation, where two homothetic triangles have same interior angles but different side lengths. In other words, Euclidean triangles can be conformally equivalent without being isometric. By contrast, any conformal automorphism of the hyperbolic plane is an isometry, as we have seen in the Poincaré models.

Another application of the hyperbolic law of cosines is the easy computation of the angle of parallelism (see Figure 1.3):

Theorem 14.14. *Let l be a line in the hyperbolic plane and A be a point at distance $a > 0$ from l. The angle of parallelism at A is the angle $\Pi(a) \in (0, \pi/2)$ given by*

$$\sin \Pi(a) = \frac{1}{\cosh a}\,. \tag{14.2}$$

opinion: distances between points and angles between lines should not be defined independently; the point is to show the relation between them.

Proof. To avoid confusion below, let us rename c the distance between A and l. Let B be the nearest-point projection of A on l and let $C \in \partial_\infty \mathbb{H}^2$ be an ideal endpoint of l. Clearly, the angle of parallelism at A is the angle $\hat{A}$ in the triangle ABC. By the dual law of cosines, which extends to triangles with one or more ideal vertices by continuity, we have $\cos \hat{C} = -\cos \hat{A} \cos \hat{B} + \sin \hat{A} \sin \hat{B} \cosh c$. We find $1 = \sin \hat{B} \cosh c$ and the conclusion follows. ∎

Remark 14.15. The formula (14.2) can also be written

$$\Pi(a) = \frac{\pi}{2} - gd(a)$$

where $gd(x) = \int_0^x \frac{\mathrm{d}t}{\cosh t}$ is the *Gudermannian function.*

14.3 Area of hyperbolic triangles

The goal of this section is to prove the following theorem:

Theorem 14.16. *Let ABC be a hyperbolic triangle with three distinct vertices, one or more possibly ideal. Denote by* $\mathrm{Area}(ABC)$ *the hyperbolic area enclosed by the triangle. We have the identity:*

$$\mathrm{Area}(ABC) = \pi - (\hat{A} + \hat{B} + \hat{C})\,. \tag{14.3}$$

Proof with the Gauss–Bonnet theorem

Theorem 14.16 is an immediate consequence of the Gauss–Bonnet theorem. The Gauss–Bonnet theorem is a deep theorem of Riemannian geometry that we shall not discuss; nevertheless, we mention this proof out of interest.

Theorem 14.17 (Gauss–Bonnet theorem)**.** *Let* (S, g) *be a compact* 2*-dimensional Riemannian manifold with boundary. Then*

$$\int_S K_g \,\mathrm{d}\sigma_g \;+\; \int_{\partial S} k_g \,\mathrm{d}s = 2\pi\chi(S)$$

where K_g *denotes the Gaussian curvature in* S, $\mathrm{d}\sigma_g$ *the area element in* S, k_g *the geodesic curvature along* ∂S, $\mathrm{d}s$ *the line element along* ∂S*, and* $\chi(S)$ *the Euler characteristic of* S.

Let us not explain precisely all these terms, and only mention what they are when S is the interior of a hyperbolic triangle:

- The Gaussian curvature (a.k.a sectional curvature) K_g is constant equal to -1 inside S, since it is an open subset of the hyperbolic plane.
- The area element $\mathrm{d}\sigma_g$ is the hyperbolic area element $\mathrm{d}A$.

- The geodesic curvature k_g vanishes along the sides of the triangle, because the sides are geodesic by assumption. However, when the boundary ∂S is only piecewise smooth, one must add to $\int_{\partial S} k_g \,\mathrm{d}s$ the exterior angle at each point of discontinuity. Thus in our situation, we have $\int_{\partial S} k_g \,\mathrm{d}s = (\pi-\hat{A})+(\pi-\hat{B})+(\pi-\hat{C})$, that is $\int_{\partial S} k_g \,\mathrm{d}s = 3\pi-(\hat{A}+\hat{B}+\hat{C})$.
- The Euler characteristic of the triangle is $\chi(S) = 1$: that is +3 (vertices) −3 (edges) +1 (face).

Putting all this together, the Gauss–Bonnet formula reads:

$$\int_{ABC} (-1)\,\mathrm{d}A + (3\pi - (\hat{A} + \hat{B} + \hat{C})) = 2\pi$$

and the formula (14.3) follows.

Ideal triangles

Before turning to an alternative proof, let us examine the case of ideal triangles.

Theorem 14.18. *All ideal triangles are congruent, and have area π.*

Proof. The fact that all ideal triangles are congruent is an immediate consequence of the fact that isometries of $\mathbb{H}^2$ act 3-transitively on the ideal boundary. Indeed, we have seen that the projective linear group $\mathrm{PGL}_2(\mathbb{R})$ acts 3-transitively on $\mathbb{R}P^1$ (Theorem 7.70), in other words it acts 3-transitively on $\hat{\mathbb{R}}$ by fractional linear transformation. The Poincaré extension of any such transformation of $\hat{\mathbb{R}} \approx \partial_\infty \mathbb{H}^2$ is an isometry of $\mathbb{H}^2$ (in the Poincaré disk or half-plane model), so we are done.

Since all ideal triangles are isometric, they all have the same area, so we can pick our favorite to check that its area is equal to π. Let us choose the ideal triangle with vertices $A = -1$, $B = 1$, $C = \infty$ in the Poincaré half-plane: see Figure 14.3. Computing its area is now elementary calculus:

$$\begin{aligned}\mathrm{Area}(ABC) &= \int_{ABC} \mathrm{d}A \\ &= \int_{x=-1}^{1} \int_{y=\sqrt{1-x^2}}^{+\infty} \frac{\mathrm{d}x\,\mathrm{d}y}{y^2} \\ &= \int_{x=-1}^{1} \frac{\mathrm{d}x}{\sqrt{1-x^2}}\,.\end{aligned}$$

The change of variables $x = \sin\theta$ yields

$$\begin{aligned}\mathrm{Area}(ABC) &= \int_{\theta=-\pi/2}^{\pi/2} \mathrm{d}\theta \\ &= \pi\,.\end{aligned}$$

∎

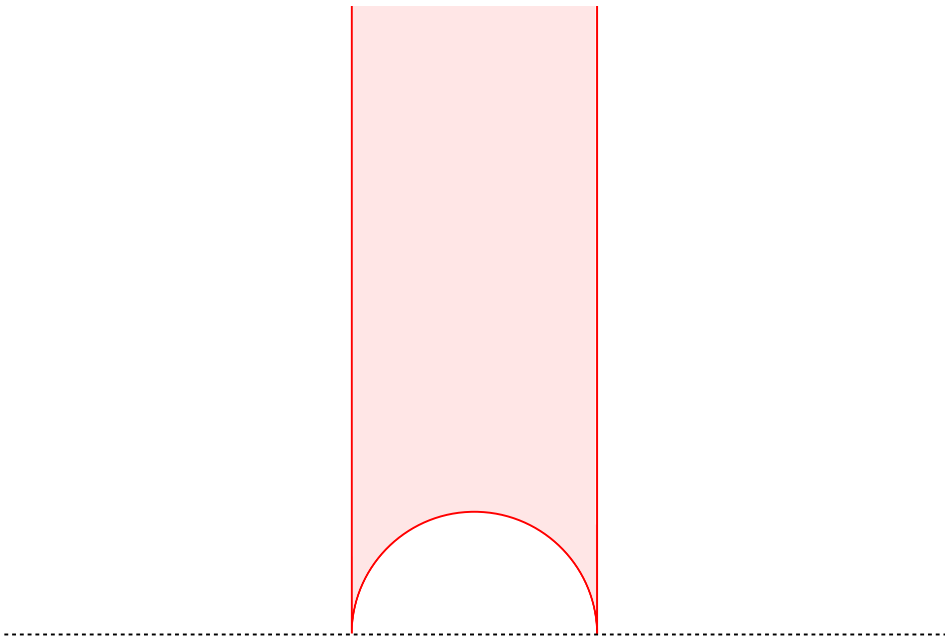


Figure 14.3: The ideal triangle with vertices -1, 1, ∞ in the Poincaré disk model.

Gauss's proof

We now propose an alternative and elegant proof of Theorem 14.16, also due to Gauss (according to [Thu97]). What follows is based on [Thu97, Prop. 2.4.13].

First consider a 2/3-ideal triangle. The congruence class of such a triangle is completely determined by the angle at the interior vertex. Indeed, after applying a translation, we can assume that the interior vertex is the origin in the Poincaré disk. It is then clear that any two such triangles with same interior angle are related by a rotation. Denote by $A(\theta)$ the area of any 2/3-ideal triangle whose angle at the interior vertex is $\pi - \theta$. Per our discussion, $A(\theta)$ is a well-defined function of $\theta \in (0, \pi)$.

Gauss's clever observation is that $A(\theta)$ is an additive function of θ: we have $A(\theta_1 + \theta_2) = A(\theta_1) + A(\theta_2)$ whenever $\theta_1, \theta_2, \theta_1 + \theta_2 \in (0, \pi)$. To see this, consider Figure 14.4. The triangles BOA, BOB', and $A'OB'$ have areas $A(\theta_1) = \mathcal{A}_1$, $A(\theta_2) = \mathcal{A}_2 + \mathcal{A}_3$, and $A(\theta_1 + \theta_2) = \mathcal{A}_3 + \mathcal{A}_4$ respectively. On the other hand, the half-turn (π-rotation) through Ω takes the triangle ΩAB to $\Omega A'B'$, so we have $\mathcal{A}_4 = \mathcal{A}_1 + \mathcal{A}_2$. Therefore $A(\theta_1 + \theta_2) = \mathcal{A}_3 + (\mathcal{A}_1 + \mathcal{A}_2) = A(\theta_1) + A(\theta_2)$.

The function $\theta \in (0, \pi) \mapsto A(\theta)$ being additive and continuous, it must be linear. Moreover, it extends continuously at π by $A(\pi) = \pi$ as a consequence of Theorem 14.18. This forces $A(\theta) = \theta$ for all $\theta \in [0, \pi]$. Hence we have proved Theorem 14.16 for 2/3-ideal triangles.

The case of 1/3-ideal triangles easily follows by a cut-and-paste procedure: any such triangle can be written as the difference of two 2/3-ideal triangles. The case of triangles with no ideal vertices is derived from the 1/3-ideal case with the same trick, writing any triangle with no ideal vertices as the difference of two 1/3-ideal triangles. We leave it to the reader to draw the appropriate sketches.

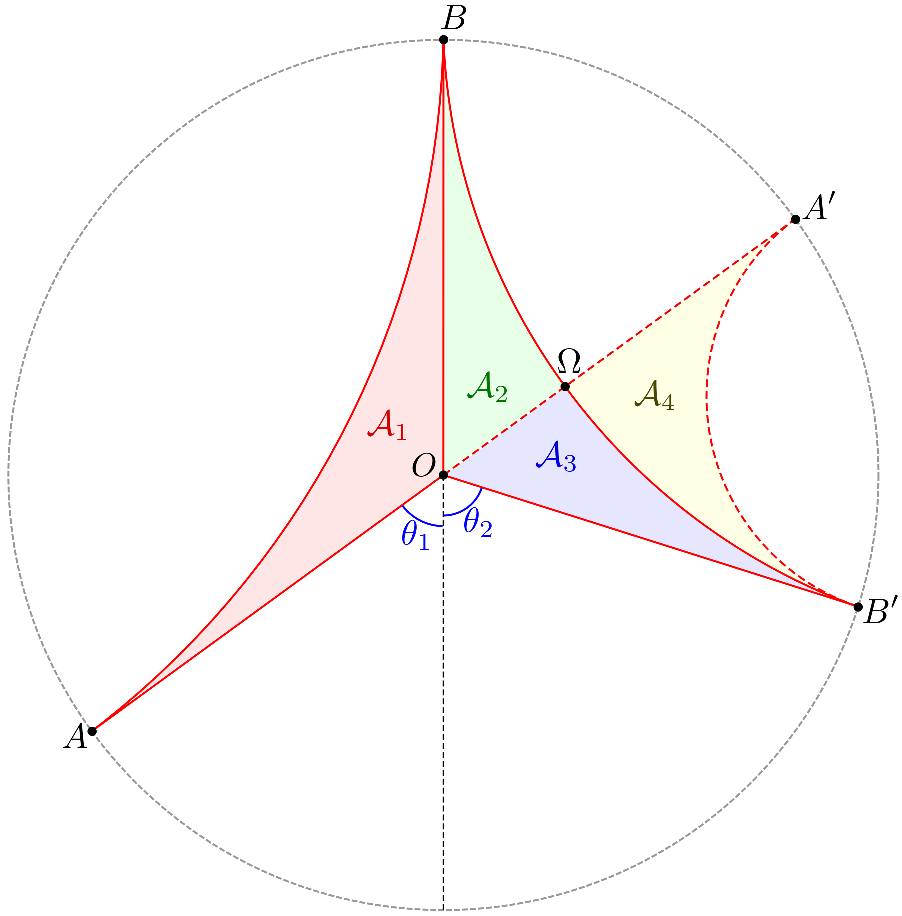


Figure 14.4: Gauss's trick to compute the area of 2/3-ideal triangles.

14.4 Gromov hyperbolicity of the hyperbolic plane

We conclude this chapter by showing that the hyperbolic plane $\mathbb{H}^2$ is a hyperbolic space in the sense of Gromov, which is a property regarding hyperbolic triangles. It readily follows that hyperbolic space $\mathbb{H}^n$ is also Gromov hyperbolic for all $n \geqslant 2$.

We have already discussed Gromov hyperbolicity in general metric spaces in Chapter 11 (see § 11.1.3), where we mentioned that the notion of ideal boundary is well-suited to such spaces (e.g., we used it for Lemma 11.16).

By definition, $\mathbb{H}^2$ being Gromov hyperbolic means that there exists $\delta \geqslant 0$ such that all hyperbolic triangles are δ-***slim***: any point on one side of the triangle is within distance $\leqslant \delta$ of some point on another side. In other words, any side is contained in the δ-neighborhood of the union of the two other sides: see Figure 11.2. In the case of $\mathbb{H}^2$, taking $\delta = 1$ is sufficient; in fact the best δ can be computed as $\delta = \operatorname{arsinh}(1) \approx 0.88137\ldots$

Theorem 14.19. *The hyperbolic plane $\mathbb{H}^2$ is hyperbolic in the sense of Gromov. The smallest constant $\delta \geqslant 0$ such that all hyperbolic triangles are δ-hyperbolic is $\delta = \operatorname{arsinh}(1)$.*

Proof. A detailed proof is proposed in Exercise 14.6. ∎

14.5 Exercises

Exercise 14.1.

Congruent triangles with ideal vertices

We have seen (Theorem 14.2) that two hyperbolic triangles are congruent if and only if they have the same side lengths. State and prove a generalization for triangles having one or more ideal vertices.

Exercise 14.2.

Congruent triangles and angles

Show that for any three numbers $\alpha, \beta, \gamma \geqslant 0$ such that $\alpha + \beta + \gamma < \pi$, there exists a hyperbolic triangle whose interior angles are equal to α, β, γ. Show that moreover, any two such triangles are congruent. Is this true for Euclidean triangles?

Exercise 14.3.

Inscribed and Circumscribed circles

(1) Show that not all hyperbolic triangles admit a circumscribed circle.

(2) In Chapter 12, we saw that any bounded set in $\mathbb{H}^n$ has a well-defined *minimum bounding ball*, which some authors call “circumball”. Is that not a contradiction with the previous question?

(3) Show that any hyperbolic triangle admits a uniquely defined inscribed circle.

(4) Show that there exists a finite upper bound for the radii of the inscribed circle of all hyperbolic triangles.

Exercise 14.4.

Unified law of cosines

For $R \in \mathbb{C} - 0$, define the generalized cosine and sine functions by:

$$\cos_R(x) = \cos\left(\frac{x}{R}\right)$$
$$\sin_R(x) = R\sin\left(\frac{x}{R}\right) .$$

Consider the “unified law of cosines for curvature $k = \frac{1}{R^2}$”:

$$\cos_R c = \cos_R a \cos_R b + \frac{1}{R^2} \sin_R a \sin_R b \cos \hat{C} .$$

(1) Check that in the case $k = -1$, i.e. $R = \pm i$, one recovers the hyperbolic law of cosines.

(2) Prove the hyperbolic law of cosines in the hyperbolic space of constant curvature $k < 0$.

(3) Predict the spherical law of cosines. Prove it.

(4) Show that the Euclidean law of cosines may be obtained asymptotically from the unified law of cosines when $k \to 0$.

(5) *Optional.* Can you come up with a heuristic explanation for the existence of a unified law of cosines that works in any constant curvature?

Exercise 14.5.

Area of hyperbolic polygons

How would you define a hyperbolic polygon? Find a formula for the area of any hyperbolic polygon, and prove it.

Exercise 14.6.

Gromov hyperbolicity of hyperbolic space

Let $n \geqslant 2$. The goal of this exercise is to show that hyperbolic space $\mathbb{H}^n$ is Gromov hyperbolic (see Definition 11.8): there exists $\delta > 0$ such that any triangle in $\mathbb{H}^n$ is δ-slim.

(1) Argue that it is enough to do the case $n = 2$.

(2) Argue that it is enough to show that some ideal triangle is δ-slim.

(3) Consider the ideal triangle with vertices $A = 0$, $B = \infty$, and $C = 1$ in the Poincaré half-plane. What are the sides (AB), (BC), and (CA) of this triangle? Draw a picture.

(4) Let $p = (0, y) \in (AB)$. Show that the distance from p to (BC) is achieved at $p' = (1, \sqrt{1 + y^2})$. Derive that $d(p, (BC)) = \operatorname{arsinh}(1/y)$.

(5) Find an isometry that maps $A \mapsto B$, $B \mapsto C$, $C \mapsto A$. Show that $d(p, (CA)) = \operatorname{arsinh}(y)$.

(6) Conclude that $d(p, (BC) \cup (CA)) \leqslant \delta$ where $\delta = \operatorname{arsinh}(1)$ and conclude the exercise.

(7) Is the constant $\delta = \operatorname{arsinh}(1)$ optimal?

CHAPTER 15

More plane hyperbolic geometry

Disclaimer: This chapter is a draft.

CHAPTER 16

Tessellations of the hyperbolic plane

Disclaimer: This chapter is a draft.

Part VII

Hyperbolic geometry and data science

The greatest mathematicians, as Archimedes, Newton, and Gauss, always united theory and applications in equal measure.

– Felix Klein[1]

[1] [Kle67].

CHAPTER 17

Graph embeddings in hyperbolic space

Disclaimer: This chapter is a draft.

CHAPTER 18

Hyperbolic neural networks

Disclaimer: This chapter is a draft.

APPENDIX A

Basic notions

Disclaimer: This chapter is a draft.

A.1 Algebra

Sets, equivalence relations, partitions, quotient sets maps, functions, notation. $\subseteq$ and $\subset$. When I write $A \subseteq B$, I do not mean to insist that A and B can be equal, but simply that I have no interest in thinking about whether they could be or not.

coloneqq notation Set notations: such as $S = \{f(x), x \in X\}$ or $S = \{x \in X | P(x)\}$.

Forall notation $\forall x, y \in X$.

Group, field, vector space

subgroup $H < G$. Quotient group.

A.1.1 Group actions

A group action is the same things as a group homomorphism.

A.1.2 Linear algebra

Assume known: vector spaces, dimension, bases, subspaces, complements. Notation $V = W + W'$, $\mathbb{R}v + \mathbb{R}w$, etc. orientation of a vector space.

linear maps, endomorphisms, projections and symmetries, matrices, duality, bidual eigenvalues, eigenspaces. Determinants.

bilinear algebra. inner products, gram-schmidt, linear and quadratic forms

Inner product spaces

orthogonal group, spectral theorem, Gram-Schmidt

Affine spaces

Notation E and $\vec{E}$ Affine frame

Complexification

A.2 Analysis

A.2.1 Multivariable calculus

A.2.2 Complex functions

A.2.3 Hyperbolic functions

A.3 Geometry

A.3.1 Metric spaces

A.3.2 Notions of topology

A.3.3 Lengths and geodesics in Euclidean spaces

Let E be a Euclidean affine space. Consider a smooth curve $\gamma \colon I \to E$. We call γ a ***geodesic*** if it satisfies the clearly equivalent conditions:

- γ has vanishing acceleration: $\gamma'' = 0$.
- γ has constant velocity: $\gamma'(t) = u$ for some $u \in \vec{E}$.
- γ is a constant speed parametrization of a straight line.

Furthermore, we have the following characterizations:

Theorem A.1. *Let $\gamma \colon I = [a,b] \to E$ be a regular smooth curve. The following are equivalent:*

(i) γ is a geodesic.

(ii) γ is a critical point of the ***energy functional*** $\mathcal{E}(\gamma) := \dfrac{1}{2}\displaystyle\int_I \|\gamma'(t)\|^2 \,\mathrm{d}t$.

(iii) γ has constant speed and is a crit. point of the ***length functional*** $\ell(\gamma) := \displaystyle\int_I \|\gamma'(t)\| \,\mathrm{d}t$.

Before proving this theorem, let us explain what it means for γ to be a critical point. A ***variation of*** γ is a family of curves $(\gamma_s)_{s\in J}$ with $\gamma_0 = \gamma$, where J is some open interval of $\mathbb{R}$ contain 0. It is a ***smooth variation*** if $(s,t) \mapsto \gamma_s(t)$ is a smooth map on $J \times I$. In the setting of Theorem A.1, we implicitly assume that any variation has fixed endpoints: $\gamma_s(a) = \gamma(a)$ and $\gamma_s(b) = \gamma(b)$ for all $s \in J$. By definition, γ is a ***critical point*** of a function $F\colon \gamma \mapsto F(\gamma) \in \mathbb{R}$ if $\frac{\mathrm{d}}{\mathrm{d}s}_{|s=0} F(\gamma_s) = 0$ for any smooth variation (γ_s).

Given a smooth variation (γ_s), the corresponding ***infinitesimal variation*** is the smooth map $X\colon I \to V$ defined by $X(t) := \frac{\mathrm{d}}{\mathrm{d}t}_{|s=0}\gamma_s(t)$. Think of $X(t)$ as a vector based at $\gamma(t)$, so that X is a *vector field along* γ. One can expect that the first variation $\frac{\mathrm{d}}{\mathrm{d}s}_{|s=0}F(\gamma_s)$ only depends on X if the function F is "differentiable": $\frac{\mathrm{d}}{\mathrm{d}s}_{|s=0}F(\gamma_s)$ should just be the differential $\mathrm{d}F(X)$. There is no need for us to further discuss this idea in general, but we shall see that it is verifed for the energy and for the length functionals.

Lemma A.2. *The first variation of the energy is given by* $\dfrac{\mathrm{d}}{\mathrm{d}s}_{|s=0}\mathcal{E}(\gamma_s) = -\displaystyle\int_I \langle \gamma''(t), X(t)\rangle\,\mathrm{d}t.$

Proof. There is no regularity obstacle to differentiate under the integral:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}s}_{|s=0}\mathcal{E}(\gamma_s) &= \frac{1}{2}\int_I \frac{\mathrm{d}}{\mathrm{d}s}_{|s=0}\left\langle \gamma_s'(t), \gamma_s'(t)\right\rangle \mathrm{d}t \\ &= \int_I \left\langle \frac{\mathrm{d}}{\mathrm{d}s}_{|s=0}\gamma_s'(t), \gamma'(t)\right\rangle \mathrm{d}t\,.\end{aligned}$$

By Schwarz's theorem (symmetry of second derivatives), $\dfrac{\mathrm{d}}{\mathrm{d}s}_{|s=0}\gamma_s'(t) = \dfrac{\mathrm{d}}{\mathrm{d}t}\dfrac{\mathrm{d}}{\mathrm{d}s}_{|s=0}\gamma_s(t) = X'(t).$ Furthermore, one can write:

$$\langle X'(t), \gamma'(t)\rangle = \frac{\mathrm{d}}{\mathrm{d}t}\langle X(t), \gamma'(t)\rangle - \langle X(t), \gamma''(t)\rangle\ .$$

We integrate over $I = [a,b]$:

$$\frac{\mathrm{d}}{\mathrm{d}s}_{|s=0}\mathcal{E}(\gamma_s) = \Big[\ \langle X(t), \gamma'(t)\rangle\ \Big]_{t=a}^{t=b} - \int_I \langle X(t), \gamma''(t)\rangle\ .$$

Since the variation (γ_s) is assumed to have fixed endpoints, we have $X(a) = X(b) = 0$, therefore the first term vanishes and we are left with $\frac{\mathrm{d}}{\mathrm{d}s}_{|s=0}\mathcal{E}(\gamma_s) = -\int_I \langle X(t), \gamma''(t)\rangle$. ■

Remark A.3. In the language of "the calculus of variations", the geodesic equation $\gamma'' = 0$ is the ***Euler-Lagrange equation*** for the energy functional.

Remark A.4. Lemma A.2 can be interpreted as: $\operatorname{grad} E(\gamma) = -\gamma''$. It holds in great generality that the gradient of the energy functional is (minus) the Laplacian operator: see e.g. [Lou19].

Proof of Theorem A.1. The lemma shows that the first variation of the energy only depends on the infinitesimal variation $X(t) = \frac{\mathrm{d}}{\mathrm{d}s}_{|s=0}\gamma_s(t)$. Any smooth vector field along γ is an infinitesimal variation: given $X(t)$, just put $\gamma_s(t) := \gamma(t) + sX(t)$. Thus, by Lemma A.2, γ is a critical point of the energy if and only if $\int_I \langle \gamma''(t), X(t) \rangle \, \mathrm{d}t = 0$ for any $X(t)$. This implies that γ'' vanishes (consider $X(t) = \gamma''(t)$). Thus we have proven that (i) is equivalent to (ii).

The first variation of length can also be computed by differentiating under the integral:

$$\frac{\mathrm{d}}{\mathrm{d}s}_{|s=0} \ell(\gamma_s) = \int_I \frac{\frac{\mathrm{d}}{\mathrm{d}s}_{|s=0} \left\langle \gamma_s'(t), \gamma_s'(t) \right\rangle}{\langle \gamma'(t), \gamma'(t) \rangle} \, \mathrm{d}t \,.$$

When γ has constant speed $v := \|\gamma'(t)\|$, this is equal to the first variation of the energy $\frac{\mathrm{d}}{\mathrm{d}s}_{|s=0} \mathcal{E}(\gamma_s)$ multiplied by the constant $\frac{1}{v^2}$. It follows that (ii) is equivalent to (iii). ∎

Corollary A.5. *For any $A, B \in E$, the line segment $[A, B]$ is the unique geodesic from A to B (up to parametrization) and is uniquely length-minimizing among all $\mathcal{C}^1$ curves from A to B.*

Proof. Since geodesics are affine parametrizations of straight lines, it is clear that $\gamma_0(t) = A + t(B - A)$ is the unique geodesic from A to B up to reparametrization. Its length is $\ell(\gamma_0) = \int_0^1 \|B - A\| \, \mathrm{d}t = \|B - A\|$. On the other hand, for any smooth curve γ, the triangle inequality for integrals says that $\left\| \int_I \gamma'(t) \, \mathrm{d}t \right\| \leqslant \int_I \|\gamma'(t)\| \, \mathrm{d}t$, that is $\|B - A\| \leqslant \ell(\gamma)$. ∎

A.3.4 Elementary Riemannian geometry

I think this should be a part of Chapter 2 maybe.

Exercises hints and solutions

Chapter 1

Exercise 1.2. (2) First prove it for a triangle that has a vertex at the center of the disk. Then play a game of cut and paste.

Chapter 2

Exercise 2.3. (4) Straightforward calculations lead to $\kappa = \frac{|a|}{a^2+b^2}$ and $\tau = \frac{b}{a^2+b^2}$. The helix with parameters $a = 2$ and $b = 3$ is shown in Figure A.1.

(5) Isometries of E act transitively on affine frames, therefore one can assume that $\gamma(0)$ is fixed as well as the Frenet–Serret frame (T, N, B) at $t = 0$. The Frenet–Serret formulas define a linear system of ODEs for (T, N, B), therefore it has a unique solution by Picard–Lindelöf given fixed initial conditions. Conclude by noting that the curve γ is recovered from integrating T. A more detailed proof can be found anywhere, e.g. [Car16, Chap. 1-5], [Pre10, Thm. 2.3.6], or [Spi99, Vol. II, Chap. 1].

(6) Note that the formulas giving κ and τ in terms of a and b can easily be inverted: $a = \frac{\pm\kappa}{\kappa^2+\tau^2}$ and $b = \frac{\tau}{\kappa^2+\tau^2}$. By (4), the helix with parameters a and b has constant curvature κ and torsion τ. By the previous question, any other curve with such curvature and torsion is an image of that helix by an isometry of E.

Exercise 2.4. Let $U_p \subseteq \mathrm{T}_p S$ denote the unit circle, consisting of unit tangent vectors at p. Given an orthonormal basis of $\mathrm{T}_p M$, U_p can be identified to the unit circle in $\mathbb{R}^2$, parametrized by the angle $\theta \in \mathbb{R}/2\pi\mathbb{Z}$. Let us denote $\vec{v}(\theta)$ the unit tangent vector with angle θ and $\rho_p(\theta) := \rho_p(\vec{v}(\theta))$ its extrinsic curvature. We claim that $H_p = \frac{1}{2\pi}\int_0^{2\pi} \rho_p(\theta)\,\mathrm{d}\theta$.

Indeed, the average of any symmetric bilinear form on $\mathbb{R}^2$ on the unit circle is equal to one half of its trace. We leave the proof of this fact as a subsidiary exercise to the reader, with the hint: use the spectral theorem.

Exercise 2.5. (1) We recall that by definition, a Riemannian isometry is a map whose derivative at any point is a linear isometry between tangent spaces.

(2) (b) Use Proposition 2.10.

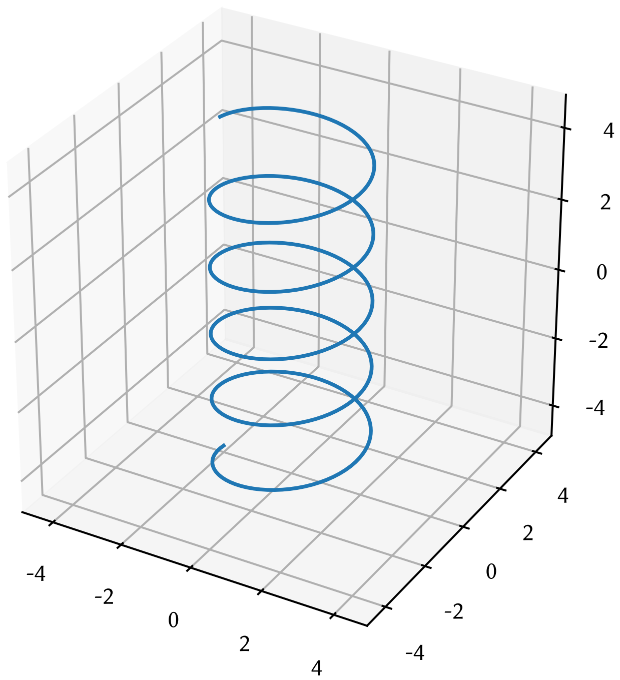


Figure A.1: The helix with $a = 2$ and $b = 3$

Note: This figure and others were created with Python using the `matplotlib` library [Hun07].

Exercise 2.6. (1) See Figure A.2 for a picture of the tractrix.

(3) See Figure A.3 for a picture of the tractricoid with meridians and parallels. It is a general fact that if a regular curve is the set of fixed points of an isometry, then this curve must be a geodesic, up to reparametrization: otherwise, it is easy to see that uniqueness of geodesics with a given initial velocity would be violated. In our case, consider the reflection through a vertical plane.

(4) (a) Since there exists an isometry of $\mathbb{R}^3$ preserving S and taking $p_\theta := f(\theta, t_0)$ to $p_0 := f(0, t_0)$, namely the rotation of angle θ around the z-axis, S must have same Gaussian curvature at p_θ and p_0. Indeed, this isometry transports everything from p_θ to p_0: geodesics, normal to S, etc, so S must have same principal curvatures at p_θ and p_0, and have same Gaussian curvature. Note that this is an illustration, in an easy case, of the Theorema Egregium.

(b) $\gamma_0'(t) = (-\operatorname{sech} t \tanh t, 0, \tanh^2 t) =: u$ and $c_t'(0) = (0, \operatorname{sech} t, 0) =: v$. To get a vector normal to S at p, we can take the cross-product of u and v, and renormalize to get a unit vector. One finds $\vec{N} = \pm(\tanh t, 0, \operatorname{sech} t)$. We take + for the "exterior" normal.

(c) Calculations yield $\gamma_0''(t) = (\operatorname{sech} t(1-2\operatorname{sech} t), 0, 2\operatorname{sech}^2 t \tanh t)$ so we find the normal curvature $\langle \gamma_0''(t), \vec{N} \rangle = \operatorname{sech} t \tanh t$. This is the extrinsic curvature $\rho_p(u)$ where $u = \gamma_0'(t)$; in order to get the extrinsic curvature $\rho_p(u_1)$ where $u_1 = \frac{u}{\|u\|}$ is the unit tangent, we have

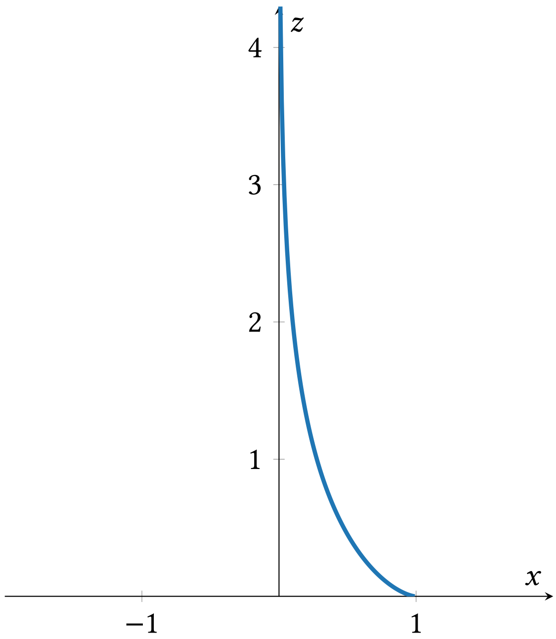


Figure A.2: The tractrix

Note: This figure was created with LaTeX using the package `Tikz`.

to divide by $\|u\|^2 = \tanh^2 t$, because ρ_p is quadratic. We obtain $\rho_p(u_1) = \frac{1}{\sinh t}$. Similar calculations yield $\rho_p(v_1) = -\sinh t$, where v_1 is the unit tangent to c_t at p.

(d) For the symmetry argument: consider the reflection through the vertical plane containing the curve γ_0. Show that up to sign, it preserves the unit vectors giving the principal directions of curvature.

(5) The arclength is easily computed as $\mathrm{d}s = \tanh t$, which gives $s = \ln(\cosh t)$. In particular, the arclength parameter stays bounded when $t \to 0$. This shows that the geodesic γ, or rather, its arclength parametrization, is incomplete. Thus S is not geodesically complete, equivalently it is not a complete metric space by the Hopf-Rinow theorem. Note that if we try to extend the tractricoid by allowing t to take negative values, then the resulting surface is singular at points where $t = 0$.

Exercise 2.8. Let us interpret Euclid's postulates in the realm of surfaces equipped with a Riemannian metric. Note that this is not only anachronistic, but also too restrictive: Euclid's postulates could be interpreted in much more generality. Nevertheless, it is an interesting exercise.

Let (S, g) be a Riemannian surface. In this setting, a *line* must be understood as a geodesic.

First postulate. The first postulate of Euclid reads: there exists a geodesic segment between any two points in (S, g). Note that if we add uniqueness, this excludes S having closed geodesics and self-intersecting geodesics. In particular, S must be simply connected. If we

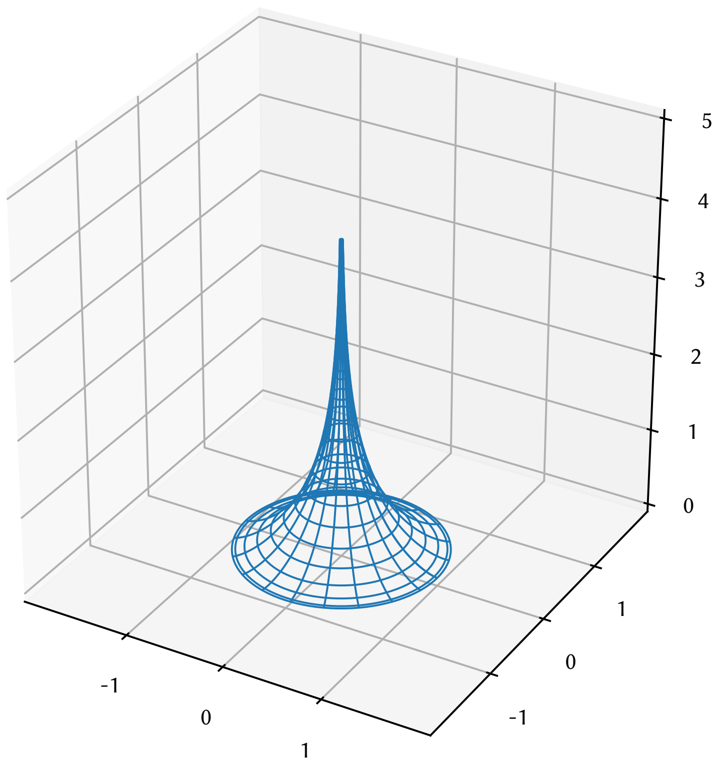


Figure A.3: The tractricoid

only require uniqueness of the maximal geodesic through two points, then the projective plane $S^2/\{\pm 1\}$, model of elliptic geometry, is acceptable.

Second postulate. The second postulate is precisely saying that (S, g) is geodesically complete. By the Hopf-Rinow theorem, this is equivalent to (S, g) being complete as a Riemannian manifold. Note that this implies the first postulate, without uniqueness.

At this point, any Hadamard 2-manifold is acceptable. That is a simply-connected, complete 2-manifold of nonpositive sectional curvature.

Third postulate. In this setting, the third postulate is trivially true: given a point p and a radius $r > 0$, the circle $C(p, R)$ is uniquely defined as the set of points at distance r from p. Although one way to interpret the postulate is that this circle is nonempty, or is a topological circle.

Fourth postulate. This is arguably the most important postulate. Firstly, it implies that (S, g) is *homogeneous*, i.e. that the group G of isometries of (S, g) acts transitively on S. The additional requirement on right angles is equivalent to (S, g) being *isotropic*: for any $p \in G$, G acts (via derivatives of its elements) transitively in $\mathrm{T}_p\, G$. Every complete isotropic Riemannian manifold is homogeneous, making the first requirement unnecessary.

At this stage, I think that up to isometry, there are no other models of Euclid's axioms than the space forms of constant curvature: the sphere S^2 and its analogs S^2_R of constant curvature $k = \frac{1}{R^2}$ for any $R > 0$, the Euclidean plane $\mathbb{R}^2$, and the hyperbolic plane $\mathbb{H}^2$ and its analogs $\mathbb{H}^2_R$ of constant curvature $k = -\frac{1}{R^2}$ for any $R > 0$. More precisely, depending on a more or less restrictive interpretation of the first postulate, we may exclude or include the spheres S^2_R and/or their quotients $S^2_k/\{\pm 1\}$.

Fifth postulate. Given a geodesic and a point not on it, there exists a unique geodesic through the point which does not intersect the first. This postulate excludes the hyperbolic planes $\mathbb{H}^2_R$ and, regardless of the interpretation of the first postulate, the spheres S^2_R and/or their quotients $S^2_R/\{\pm 1\}$.

We can therefore wrap up:

Theorem. *Let (S, g) be a smooth connected surface equipped with a Riemannian metric. Then*

(i) (S, g) satisfies the first four postulates of Euclid if and only if it isometric to either $\mathbb{R}^2$, or $\mathbb{H}^2_R$ for some $R > 0$. Depending on the interpretation of the first postulate, the spheres S^2_k and/or their quotients $S^2_R/\{\pm 1\}$ should also be included.

(ii) (S, g) satisfies the five postulates of Euclid if and only if it is isometric to $\mathbb{R}^2$.

Chapter 3

Chapter 4

Chapter 5

Exercise 5.1. (1) (b) We recall that by definition, a Riemannian isometry is a map whose derivative at any point is a linear isometry between tangent spaces.

(2) (c) Hint: identify this action to the action of $\mathrm{O}(n)$ in $\mathbb{R}^n$. (d) Hint: Recall that $\mathcal{H}^+$ is uniquely goedesic: for any $v \in \mathcal{H}^+$, there exists a unique geodesic from v_0 to v.

Exercise 5.2. (1) This is an immediate computation after recalling that the distance on the hyperboloid is $d(p, q) = \operatorname{arcosh}(-\langle p, q\rangle)$.

(2) $\operatorname{arcosh}(\cosh^2 x) = \sqrt{2}x + \frac{1}{6\sqrt{2}}x^3 + O(x^4)$.

(3) It follows from (1) and (2) that $d(\gamma_v(t), \gamma_w(t))^2 = 2t^2 - \frac{1}{3}t^4 + O(t^5)$, hence $K(v, w) = -1$. Since this holds for any p and any orthonormal pair $v, w \in \mathrm{T}_p\,\mathcal{H}^+$, we proved that $\mathcal{H}^+$ has constant sectional curvature $K = -1$.

(4) We now find $d(\gamma_v(t), \gamma_w(t)) = R \operatorname{arcosh}\left(\frac{\cosh^2 t}{R}\right)$. It follows from the Taylor expansion found in (2) that $d(\gamma_v(t), \gamma_w(t))^2 = 2t^2 - \frac{1}{3R^2}t^4 + O(t^5)$, hence $K(v, w) = -\frac{1}{R^2}$.

Chapter 7

Exercise 7.17. Hint: Show that for a surface $S \subseteq \mathbb{R}^3 \subseteq \mathbb{R}P^3$, the sign of any extrinsic curvature $\rho_p(v)$ is invariant under orientation-preserving projective linear transformations of $\mathbb{R}P^3$.

Chapter 8

Exercise 8.6. (1) By definition of the Cayley–Klein metric,

$$\begin{aligned} d(x,y) &= \frac{1}{2}\left|\ln([0, \pm\|x\|, -1, 1])\right| \\ &= \frac{1}{2}\left|\ln \frac{1 \mp \|x\|}{1 \pm \|x\|}\right| \\ &= \operatorname{artanh}(\|x\|)\,. \end{aligned}$$

(2)

$$\begin{aligned} d(x,y) &= \operatorname{arcosh}\left(\frac{1 - \langle x, 0\rangle}{\sqrt{(1-0)(1-\|x\|^2)}}\right) \\ &= \operatorname{arcosh}\left(\frac{1}{\sqrt{1-\|x\|^2}}\right) \\ &= \operatorname{artanh}(\|x\|)\,. \end{aligned}$$

(3) Let $\gamma(t) = tx$ for $t \in [0,1]$. Since the image of γ is a minimizing geodesic, $d(0,x) = L(\gamma) = \int_0^1 \|\gamma'(t)\|\,\mathrm{d}t$. Here we have $\gamma'(t) = tx$ and the expression of the Riemannian metric on the Beltrami–Klein disk gives (after a couple lines of calculations) $\|\gamma'(t)\| = \frac{\|x\|}{1-t^2\|x\|^2}$. Note that this is $\frac{\mathrm{d}}{\mathrm{d}t}\operatorname{artanh}(t\|x\|)$, so we find $d(0,x) = \operatorname{artanh}(\|x\|)$.

Exercise 8.7. Show that: (i) The result is true when $x_0 = 0$, (ii) $\mathrm{PO}(2,1)$ acts transitively on circles of radius R, and (iii) $\mathrm{PO}(2,1)$ sends ellipses to ellipses. Of course, you could also try a direct proof, let me know if you succeed that.

Chapter 9

Exercise 9.7. (3) You should find that the pullback metric on $S^n \subseteq \mathbb{R}^{n+1}$ is $\frac{\mathrm{d}x_1^2+\cdots+\mathrm{d}x_{n+1}^2}{(1-x_{n+1})^2}$. Clearly, this is conformal to the Euclidean metric of $\mathbb{R}^{n+1}$, which is the spherical metric in restriction to S^n.

Exercise 9.9. (3) (b) For a translation $z \mapsto z + b$, take two reflections having b as a normal vector. For a similarity $z \mapsto az$ with $a \in \mathbb{C}^*$, first write it as the composition of the rotation

$z \mapsto e^{i\theta} z$ and the homothety $z \mapsto \rho z$, where $a = \rho e^{i\theta}$. For the rotation, try two reflections whose axes intersect at the origin. For the homothety, try two inversions through spheres centered at the origin. Finally, write $z \mapsto \frac{1}{z}$ as the composition of the inversion through the unit circle and the reflection through the real axis.

Exercise 9.10. (5) Hint: Show that if a fractional linear transformation preserves $\mathbb{D}$, then it also preserves the unit circle $\partial\mathbb{D} = \{|z| = 1\}$. Then show that if the fractional linear action of $M = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ preserves $|z| = 1$, then $a\bar{b} - c\bar{d} = 0$ and $|a|^2 - |c|^2 = |d|^2 - |b|^2$. Conclude that, after multiplying M by a constant, it belongs to $\mathrm{U}(1,1)$.

Exercise 9.12. *Still works:* Defining inversions, Möbius transformations as product of inversions, both for $\widehat{\mathbb{R}}$ and S^1. It is still true that the $\mathrm{M\ddot{o}b}(\widehat{\mathbb{R}}) \approx \mathrm{M\ddot{o}b}(S^1) \approx \mathrm{PO}(1,1)$ and $\mathrm{M\ddot{o}b}^+(\widehat{\mathbb{R}}) \approx \mathrm{M\ddot{o}b}^+(S^1) \approx \mathrm{PSO}(2,1)$. In this case, one can also identify $\mathrm{M\ddot{o}b}^+(\widehat{\mathbb{R}})$ to $\mathrm{PGL}_2^+(\mathbb{R})$, acting by fractional linear transformations on $\widehat{\mathbb{R}} \approx \mathbb{R}P^1$. What also works is the Poincaré extension from dimension 1 to 2: any Möbius transformation of $\mathrm{M\ddot{o}b}(\widehat{\mathbb{R}})$ [resp. S^1] extends to a unique transformation of H^2 [resp. B^2]. In fact we see directly that $\mathrm{PGL}_2^+(\mathbb{R})$ acts both on $\mathrm{M\ddot{o}b}(\widehat{\mathbb{R}})$ and $\mathbb{H}$ by fractional linear transformations; similarly $\mathrm{PSU}(1,1)$ acts both on $S^1 \subseteq \mathbb{C}$ and $\mathbb{D} \subseteq \mathbb{C}$ by fractional linear transformations.

Breaks down: Any diffeomorphism $\mathbb{R} \to \mathbb{R}$ is conformal, therefore the Liouville theorem does not hold: these are not all Möbius transformations. It is also not true that Möbius transformations can be characterized as sphere-preserving, because in this case lower dimensional spheres are pairs of points, so any injective map is sphere-preserving. The Poincaré extension from dimension 0 to 1 also fails: it is not true that any Möbius transformation of $\widehat{\mathbb{R}}$ is uniquely determined by its restriction to $\widehat{\mathbb{R}^0} = \{0, \infty\}$. This is because while it is still true that the subgroup of $\mathrm{PO}(2,1)$ preserving $\widehat{\mathbb{R}^0}$ is $\mathrm{PO}(1,1)$, the latter does not act faithfully on $\widehat{\mathbb{R}^0}$; in other words, what fails is that $\mathrm{M\ddot{o}b}(\widehat{\mathbb{R}^n}) = \mathrm{PO}(n+1,1)$ is not correct for $n = 0$.

Chapter 10

Exercise 10.9. If we define a hyperbolic space *à la Euclid*, axiomatically, then we could define a hyperbolic subspace of X as a subset $X' \subseteq X$ where the axioms still hold. In order for this to make sense, we should assume that X' is stable under taking the line through two points. It turns out that this condition is sufficient.

Let us take instead the modern definition of a hyperbolic space as a complete, simply-connected Riemannian manifold of constant sectional curvature -1. A hyperbolic subspace is a complete and totally geodesic submanifold $X' \subseteq X$. Equivalently, X' is a subset of X stable under taking the complete geodesic through any two points. Equivalently, X' is a totally geodesically embedded copy of $\mathbb{H}^k$ in $\mathbb{H}^n$ for some $k \leqslant n$.

Hyperbolic subspaces have very natural incarnations in the different models. In the hyperboloid model, a hyperbolic subspace is the intersection with a subspace of Minkowski space (see Proposition 5.1). In the Cayley–Klein model, it is the intersection with a projective subspace. In the Beltrami–Klein ball, it is the intersection with an affine subspace. In the Poincaré models, it is the intersection with a half-sphere orthogonal to the boundary. We leave it to the reader to prove all these descriptions.

Chapter 11

Exercise 11.1. (2) Hint: Compare the distance between two geodesics from the same point in $\mathbb{R}^2$ versus in $\mathbb{H}^2$.

(3) Hint 1: Show that any quasi-isometric (i.e. coarsely surjective) map $r\colon [0,+\infty) \longrightarrow \mathbb{H}^n$ is at finite distance from a geodesic ray. Hint 2: Show that if there exists a quasi-isometry $X \longrightarrow Y$, then there exists a quasi-isometry $Y \longrightarrow X$.

Exercise 11.2.. Hint: start by recalling the relation between the hyperboloid model and the Cayley–Klein model.

Chapter 12

Exercise 12.1. (1) By definition, a function $g\colon \mathbb{N} \longrightarrow \mathbb{R}$ is subadditive if $g(x+y) \leqslant g(x) + g(y)$ for all $x, y \in \mathbb{N}$. For such a function, $\lim_{n\to+\infty} \frac{g(n)}{n}$ always exists. Indeed, for a fixed integer $d > 0$, the Euclidean division of n by d is written $n = qd + r$ with $0 \leqslant r < d$. The subadditivity condition implies that $\frac{g(n)}{n} \leqslant \frac{g(d)}{d} + \frac{g(r)}{n}$, hence $\limsup_{n\to+\infty} \frac{g(n)}{n} \leqslant \frac{g(d)}{d}$. Therefore $\limsup_{n\to+\infty} \frac{g(n)}{n} \leqslant \liminf_{d\to+\infty} \frac{g(d)}{d}$.

Exercise 12.8. (1) Hint: Derive from the Cayley-Hamilton theorem that $B + B^{-1} = \mathrm{tr}(B)I$.

(2) Hint: Start by words of length 1, 2, 3, etc. (in the generators A, B, A^{-1}, and B^{-1}).

Exercise 12.9. For instance, try to prove the following classification:

- $M \in \mathrm{O}^+(n,1)$ is elliptic if and only if M has a timelike eigenvector. In this case, all complex eigenvalues of M have unit modulus.
- $M \in \mathrm{O}^+(n,1)$ is loxodromic if and only if M has a complex eigenvalue λ of modulus $\neq 1$. In this case, λ and λ^{-1} are the only complex eigenvalues of M of modulus $\neq 1$.
- $M \in \mathrm{O}^+(n,1)$ is elliptic if and only if all complex eigenvalues of M have unit modulus, and M has no timelike eigenvector.

For additional guidance, you can check out [Thu97, Problem 2.5.24].

Index of notations

Index of terminology

References

[AP15] Norbert A'Campo and Athanase Papadopoulos. "On Klein's *so-called non-Euclidean geometry*". In: *Sophus Lie and Felix Klein: the Erlangen program and its impact in mathematics and physics.* Vol. 23. IRMA Lect. Math. Theor. Phys. Eur. Math. Soc., Zürich, 2015, pp. 91–136 (cited pp. 7, 131).

[Ahl78] Lars V. Ahlfors. *Complex analysis.* Third. An introduction to the theory of analytic functions of one complex variable, International Series in Pure and Applied Mathematics. McGraw-Hill Book Co., New York, 1978. xi+331 (cited p. 170).

[Ale51] A. D. Aleksandrov. "A theorem on triangles in a metric space and some of its applications". In: *Trudy Mat. Inst. Steklov., v 38.* Trudy Mat. Inst. Steklov., v 38. Izdat. Akad. Nauk SSSR, Moscow, 1951, pp. 5–23 (cited p. 197).

[Aud03] Michèle Audin. *Geometry.* Universitext. Translated from the 1998 French original. Springer-Verlag, Berlin, 2003. vi+357. DOI: `10.1007/978-3-642-56127-6` (cited pp. 81, 101).

[Aud06] Michèle Audin. *Géométrie.* Enseignement SUP-Maths. EDP Sciences, 2006 (cited pp. 81, 101).

[AH19] Nuh Aydin and Lakhdar Hammoudi. *Al-Kāshī's Miftāh al-Hisab. Vol. I. Arithmetic.* Translation and commentary. Birkhäuser/Springer, Cham, 2019. vii+241 (cited p. 253).

[Bal20] Christopher Baltus. *Collineations and conic sections—an introduction to projective geometry in its history.* Springer, Cham, 2020. ix+187. DOI: `10.1007/978-3-030-46287-1` (cited p. 81).

[Bea95] Alan F. Beardon. *The geometry of discrete groups.* Vol. 91. Graduate Texts in Mathematics. Corrected reprint of the 1983 original. Springer-Verlag, New York, 1995. xii+337 (cited pp. 154, 160).

[Bel86] E.T. Bell. *Men of Mathematics.* A Fireside book. Touchstone, 1986 (cited p. 79).

[Bel68a] Eugenio Beltrami. *Saggio di interpretazione della geometria Non-Euclidea.* ita. Napoli: s.n., 1868 (cited pp. 7, 131, 177).

[Bel68b] Eugenio Beltrami. "Teoria fondamentale degli spazii di curvatura costante". In: *Annali di Matematica Pura ed Applicata (1867-1897)* 2.1 (Aug. 1868), pp. 232–255. DOI: `10.1007/BF02419615` (cited pp. 7, 131, 177).

[Ber77] Marcel Berger. *Géométrie I, II.* CEDIC, Paris; Nathan Information, Paris, 1977 (cited pp. 33, 81, 101).

[Ber09] Marcel Berger. *Geometry I, II.* Universitext. Translated from the 1977 French original by M. Cole and S. Levy, Fourth printing of the 1987 English translation. Springer-Verlag, Berlin, 2009 (cited pp. 33, 81, 101, 105).

[Ber16] Marcel Berger. *Géométrie I, II.* Nouvelle Bibliothèque Mathématique. Cassini, 2016 (cited p. 33).

[BR98] Albrecht Beutelspacher and Ute Rosenbaum. *Projective geometry: from foundations to applications.* Cambridge University Press, Cambridge, 1998. x+258 (cited pp. 81, 85, 96, 97).

[BR04] Albrecht Beutelspacher and Ute Rosenbaum. *Projektive Geometrie.* Second. Vieweg Studium: Aufbaukurs Mathematik. [Vieweg Studies: Mathematics Course]. Von den Grundlagen bis zu den Anwendungen. [From the foundations to the applications]. Friedr. Vieweg & Sohn, Wiesbaden, 2004. x+265. DOI: `10.1007/978-3-322-80329-0` (cited p. 85).

[Bob19] Alexander I. Bobenko. *Projective Geometry.* `https://page.math.tu-berlin.de/~bobenko/Lehre/Skripte/projective_part.pdf`. 2019 (cited p. 127).

[Bou71] N. Bourbaki. *Éléments de mathématique. Topologie générale. Chapitres 1 à 4.* Hermann, Paris, 1971. xv+357 pp. (not consecutively paged) (cited pp. 83, 106).

[BH99] Martin R. Bridson and André Haefliger. *Metric spaces of non-positive curvature.* Vol. 319. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Springer-Verlag, Berlin, 1999. xxii+643. DOI: `10.1007/978-3-662-12494-9` (cited pp. 194–197, 201, 205, 214, 220, 237).

[BBI01] Dmitri Burago, Yuri Burago, and Sergei Ivanov. *A course in metric geometry.* Vol. 33. Graduate Studies in Mathematics. American Mathematical Society, Providence, RI, 2001. xiv+415. DOI: `10.1090/gsm/033` (cited p. 194).

[Cal13] Danny Calegari. "Liouville illiouminated". Blog post on the math web blog *Geometry and the imagination.* URL: `https://lamington.wordpress.com/2013/10/28/liouville-illiouminated/`. 2013 (cited p. 160).

[Can+97] James W. Cannon et al. "Hyperbolic geometry". In: *Flavors of geometry.* Vol. 31. Math. Sci. Res. Inst. Publ. Cambridge Univ. Press, Cambridge, 1997, pp. 59–115 (cited p. v).

[Car16] Manfredo P. do Carmo. *Differential geometry of curves & surfaces.* Revised & updated second edition of [MR0394451]. Dover Publications, Inc., Mineola, NY, 2016. xvi+510 (cited pp. 12, 275).

[Car88] Élie Cartan. *Leçons sur la géométrie des espaces de Riemann.* Les Grands Classiques Gauthier-Villars. [Gauthier-Villars Great Classics]. Reprint of the second (1946) edition. Éditions Jacques Gabay, Sceaux, 1988. viii+379 (cited pp. 197, 218).

[Cay59] Arthur Cayley. "A sixth Memoir upon Quantics". In: *Philos. Trans. Roy. Soc. London Ser. A* 149 (1859), pp. 61–90 (cited pp. 7, 131).

[Chr19] Piotr T. Chruściel. *Elements of general relativity.* Compact Textbooks in Mathematics. Birkhäuser/Springer, Cham, 2019, pp. ix+283. DOI: `10.1007/978-3-030-28416-9` (cited p. 39).

[CDP90] M. Coornaert, T. Delzant, and A. Papadopoulos. *Géométrie et théorie des groupes.* Vol. 1441. Lecture Notes in Mathematics. Les groupes hyperboliques de Gromov. [Gromov hyperbolic groups], With an English summary. Springer-Verlag, Berlin, 1990. x+165 (cited p. 194).

[Cox89] H. S. M. Coxeter. *Introduction to geometry.* Wiley Classics Library. Reprint of the 1969 edition. John Wiley & Sons, Inc., New York, 1989. xxii+469 (cited pp. 81, 85, 126, 127).

[Cox94] H. S. M. Coxeter. *Projective geometry.* Revised reprint of the second (1974) edition. Springer-Verlag, New York, 1994. xii+162 (cited pp. 81, 85).

[Cox98] H. S. M. Coxeter. *Non-Euclidean geometry.* Sixth. MAA Spectrum. Mathematical Association of America, Washington, DC, 1998. xviii+336 (cited p. 3).

[Dap] Dap (`https://mathoverflow.net/users/112284/dap`). *Regularity of conformal maps.* MathOverflow. URL: `https://mathoverflow.net/q/351206` (version: 2020-01-26) (cited p. 160).

[DSU17] Tushar Das, David Simmons, and Mariusz Urbański. *Geometry and dynamics in Gromov hyperbolic metric spaces.* Vol. 218. Mathematical Surveys and Monographs. With an emphasis on non-proper settings. American Mathematical Society, Providence, RI, 2017. xxxv+281 (cited p. 216).

[Die96] Jean Dieudonné. *Abrégé d'histoire des mathématiques: 1700–1900.* Hermann, 1996. 518 pp. (cited p. 3).

[Dob61] Jay Doblin. *Perspective: a new system for designers.* Whitney Library of Design New York, 1961 (cited pp. 111, 121).

[Duc18] Bruno Duchesne. "Groups acting on spaces of non-positive curvature". In: *Handbook of group actions. Vol. III.* Vol. 40. Adv. Lect. Math. (ALM). Int. Press, Somerville, MA, 2018, pp. 101–141 (cited p. 196).

[Efi64] N. V. Efimov. "Generation of singularites on surfaces of negative curvature". In: *Mat. Sb. (N.S.)* 64 (106) (1964), pp. 286–320 (cited p. 29).

[Ein15] Albert Einstein. *Relativity: the Special and the General theory.* 100th anniversary. With commentaries and background material by Hanoch Gutfreund and Jürgen Renn. Princeton, N.J.: Princeton University Press, 2015. xviii + 300 (cited p. 31).

[Euc56] Euclid. *The thirteen books of Euclid's Elements translated from the text of Heiberg. Vol. I: Introduction and Books I, II. Vol. II: Books III–IX. Vol. III: Books X–XIII and Appendix.* Translated with introduction and commentary by Thomas L. Heath, 2nd ed. Dover Publications, Inc., New York, 1956. xi+432 pp.; i+436 pp.; i+546 (cited pp. 4, 254).

[FS19] François Fillastre and Andrea Seppi. "Spherical, hyperbolic, and other projective geometries: convexity, duality, transitions". In: *Eighteen essays in non-Euclidean geometry.* Vol. 29. IRMA Lect. Math. Theor. Phys. Eur. Math. Soc., Zürich, 2019, pp. 321–409 (cited pp. 131, 143).

[GHL04] Sylvestre Gallot, Dominique Hulin, and Jacques Lafontaine. *Riemannian geometry.* Third. Universitext. Springer-Verlag, Berlin, 2004. xvi+322. DOI: 10.1007/978-3-642-18855-8 (cited p. 12).

[GLM19] Jonah Gaster, Brice Loustau, and Léonard Monsaingeon. "Computing harmonic maps between Riemannian manifolds". In: *Preprint: arXiv:1910.08176* (2019) (cited p. 23).

[GH90] Étienne Ghys and Pierre de la Harpe. *Sur les groupes hyperboliques d'après Mikhael Gromov.* Ed. by Etienne Ghys and Pierre de la Harpe. Vol. 83. Progress in Mathematics. Boston, MA: Birkhäuser Boston, 1990. DOI: 10.1007/978-1-4684-9167-8 (cited pp. 194, 214, 216).

[Gra04] Jeremy Gray. *János Bolyai, non-Euclidean geometry, and the nature of space.* Vol. 1. Burndy Library Publications. New Series. With a foreword by Benjamin Weiss, a facsimile of Bolyai's ıt Appendix, and an 1891 English translation by George Bruce Halsted. Burndy Library, Cambridge, MA, 2004. viii+246 (cited p. 1).

[Gra11] Jeremy Gray. *Worlds out of nothing.* Springer Undergraduate Mathematics Series. A course in the history of geometry in the 19th century, Second edition [of MR2305283]. Springer-Verlag London, Ltd., London, 2011. xxvi+384 (cited p. 3).

[Gre93] Marvin Jay Greenberg. *Euclidean and non-Euclidean geometries.* Third. Development and history. W. H. Freeman and Company, New York, 1993. xviii+483 (cited pp. 2, 3).

[Gro87] M. Gromov. "Hyperbolic groups". In: *Essays in group theory.* Vol. 8. Math. Sci. Res. Inst. Publ. Springer, New York, 1987, pp. 75–263. DOI: 10.1007/978-1-4613-9586-7_3 (cited p. 196).

[HOS00] Francisco J. Herranz, Ramón Ortega, and Mariano Santander. "Trigonometry of spacetimes: a new self-dual approach to a curvature/signature (in)dependent trigonometry". In: *J. Phys. A* 33.24 (2000), pp. 4525–4551. DOI: 10.1088/0305-4470/33/24/309 (cited p. 143).

[Her18] Alain Herreman. "Aux sources du théorème du perroquet". In: *Images des Mathématiques (CNRS)* (2018). URL: https://images.math.cnrs.fr/Aux-sources-du-theoreme-du-perroquet.html (cited p. 4).

[Hil99] David Hilbert. *Grundlagen der Geometrie.* Fourteenth. Vol. 6. Teubner-Archiv zur Mathematik. Supplement [Teubner Archive on Mathematics. Supplement]. B. G. Teubner Verlagsgesellschaft mbH, Stuttgart, 1999. xxiv+408 (cited pp. 84, 85, 126).

[Hoh+18] M. Hohenwarter et al. *GeoGebra – Dynamic Mathematics for Everyone*. Version: 6.0.507.0-w. `http://www.geogebra.org`. Oct. 2018 (cited p. 15).

[Hun07] J. D. Hunter. "Matplotlib: A 2D graphics environment". In: *Computing in Science & Engineering* 9.3 (2007), pp. 90–95. DOI: `10.1109/MCSE.2007.55` (cited p. 276).

[Ida10] Brigham Young University – Idaho. "Basic Perspective: Minor Axis". `https://courses.byui.edu/art110_new/art110/week01/minor.html`. 2010 (cited p. 111).

[IM01] Tadeusz Iwaniec and Gaven Martin. *Geometric function theory and non-linear analysis*. Oxford Mathematical Monographs. The Clarendon Press, Oxford University Press, New York, 2001. xvi+552 (cited p. 160).

[Jon86] Alexander Jones, ed. *Pappus of Alexandria. Book 7 of the Collection*. Translation and commentary by Alexander Jones. Springer-Verlag, New York, 1986. xvi+749. DOI: `10.1007/978-1-4612-4908-5` (cited p. 126).

[Kap] Vitali Kapovitch. *Conformally-flat*. MathOverflow. URL: `https://mathoverflow.net/q/85542` (version: 2012-01-13) (cited p. 188).

[Kle71] Felin Klein. "Ueber die sogenannte Nicht-Euklidische Geometrie". In: *Mathematische Annalen* 4.4 (Dec. 1871), pp. 573–625. DOI: `10.1007/BF02100583` (cited pp. 7, 131).

[Kle73] Felix Klein. "Ueber die sogenannte Nicht-Euklidische Geometrie". In: *Mathematische Annalen* 6.2 (June 1873), pp. 112–145. DOI: `10.1007/BF01443189` (cited pp. 7, 131).

[Kle93] Felix Klein. "Vergleichende Betrachtungen über neuere geometrische Forschungen". In: *Math. Ann.* 43.1 (1893), pp. 63–100. DOI: `10.1007/BF01446615` (cited p. 81).

[Kle67] Felix Klein. *Elementarmathematik vom Höheren Standpunkte aus*. Springer Berlin Heidelberg, 1967. DOI: `10.1007/978-3-662-11652-4` (cited p. 267).

[Laf15] Jacques Lafontaine. *An introduction to differential manifolds*. second. Springer, Cham, 2015. xx+395. DOI: `10.1007/978-3-319-20735-3` (cited pp. 12, 88).

[Lee11] John M. Lee. *Introduction to topological manifolds*. Second. Vol. 202. Graduate Texts in Mathematics. Springer, New York, 2011. xviii+433. DOI: `10.1007/978-1-4419-7940-7` (cited pp. 12, 88).

[Lee13] John M. Lee. *Introduction to smooth manifolds*. Second. Vol. 218. Graduate Texts in Mathematics. Springer, New York, 2013. xvi+708 (cited pp. 12, 88).

[Lee18] John M. Lee. *Introduction to Riemannian manifolds*. Vol. 176. Graduate Texts in Mathematics. Second edition of [MR1468735]. Springer, Cham, 2018. xiii+437 (cited pp. 12, 23, 69).

[Lou19] Brice Loustau. "Harmonic maps from Kähler manifolds". In: *Preprint* (2019) (cited p. 273).

[Luc96] C. G. Luckhardt. *Wittgenstein: Sources and Perspectives*. Wittgenstein studies. Thoemmes Press, 1996 (cited p. i).

[Mat17] Matt F. (https://mathoverflow.net/users/44143/matt-f). *What is needed to prove the consistency of Tarski's Euclidean geometry?* MathOverflow. `https://mathoverflow.net/q/261416` (version: 2017-02-05). 2017 (cited p. 9).

[Mes83] Bruce E. Meserve. *Fundamental Concepts of Geometry*. Dover Publications, 1983 (cited p. 81).

[Mil82] John Milnor. "Hyperbolic geometry: the first 150 years". In: *Bull. Amer. Math. Soc. (N.S.)* 6.1 (1982), pp. 9–24. DOI: `10.1090/S0273-0979-1982-14958-8` (cited p. 3).

[Mil72] Tilla Klotz Milnor. "Efimov's theorem about complete immersed surfaces of negative curvature". In: *Advances in Math.* 8 (1972), pp. 474–543. DOI: `10.1016/0001-8708(72)90007-2` (cited p. 29).

[Möb27] A. F. Möbius. *Der barycentrische Calcul.* Verlag von Johann Ambrosius Barth, Leipzig, 1827 (cited p. 87).

[MvG] MvG (`https://math.stackexchange.com/users/35416/mvg`). *Cosine Law Duality in Hyperbolic Trigonometry.* Mathematics Stack Exchange. URL: `https://math.stackexchange.com/q/1871478` (version: 2016-07-26) (cited p. 255).

[Nel20] Roice Nelson. *Isometry Classes of Hyperbolic 3-Space.* Website. `http://roice3.org/h3/isometries/`. 2020 (cited pp. 232, 233).

[New00] J.R. Newman. *The World of Mathematics, Vol. 1.* The World of Mathematics. Dover, 2000 (cited pp. 31, 79).

[Noi] Noix07. *the equivalence between paracompactness and second countablity in a locally Euclidean and T_2 space.* Mathematics Stack Exchange. URL: `https://math.stackexchange.com/q/2007292` (version: 2017-04-13) (cited p. 88).

[Nor99] Ernest R. Norling. *Perspective Made Easy.* Dover Publications, 1999 (cited p. 111).

[ONe83] Barrett O'Neill. *Semi-Riemannian geometry.* Vol. 103. Pure and Applied Mathematics. With applications to relativity. Academic Press, Inc. [Harcourt Brace Jovanovich, Publishers], New York, 1983. xiii+468 (cited pp. 39, 59).

[Pap14] Athanase Papadopoulos. *Metric spaces, convexity and non-positive curvature.* Second. Vol. 6. IRMA Lectures in Mathematics and Theoretical Physics. European Mathematical Society (EMS), Zürich, 2014. xii+309. DOI: `10.4171/132` (cited pp. 140, 149).

[Poi82] Henri Poincaré. "Théorie des groupes fuchsiens". In: *Acta Math.* 1.1 (1882). First article in a series published in the same volume. DOI: `10.1007/BF02391845` (cited pp. 7, 177).

[Poi02] Henri Poincaré. *La Science et l'Hypothèse.* Bibliothèque de Philosophie scientifique. Paris: Flammarion, 1902 (cited p. 3).

[Poi08] Henri Poincaré. *Science et méthode.* Paris: Flammarion, 1908 (cited p. 153).

[Pre10] Andrew Pressley. *Elementary differential geometry.* Second. Springer Undergraduate Mathematics Series. Springer-Verlag London, Ltd., London, 2010. xii+473. DOI: `10.1007/978-1-84882-891-9` (cited p. 275).

[Rey93] William F. Reynolds. "Hyperbolic geometry on a hyperboloid". In: *Amer. Math. Monthly* 100.5 (1993), pp. 442–455. DOI: `10.2307/2324297` (cited p. 64).

[Ric11] Jürgen Richter-Gebert. *Perspectives on projective geometry.* A guided tour through real and complex geometry. Springer, Heidelberg, 2011. xxii+571. DOI: `10.1007/978-3-642-17286-1` (cited pp. 81, 126, 131, 143).

[Rie13] Bernhard Riemann. *Bernhard Riemann "Über die Hypothesen, welche der Geometrie zu Grunde liegen".* Klassische Texte der Wissenschaft. [Classical Texts of Science]. Historical and mathematical commentary by Jürgen Jost. Springer Spektrum, 2013. x+156. DOI: `10.1007/978-3-642-35121-1` (cited p. 23).

[Rob04] Scott Robertson. *How to Draw Cars the Hot Wheels Way.* Motorbooks, 2004 (cited p. 111).

[RB13] Scott Robertson and Thomas Bertling. *How to Draw: Drawing and Sketching Objects and Environments from Your Imagination.* Design Studio Press, 2013 (cited p. 111).

[Ros88] B. A. Rosenfeld. *A history of non-Euclidean geometry.* Vol. 12. Studies in the History of Mathematics and Physical Sciences. Evolution of the concept of a geometric space, Translated from the Russian by Abe Shenitzer. Springer-Verlag, New York, 1988. xii+471. DOI: `10.1007/978-1-4419-8680-1` (cited p. 3).

[SW77] Rainer Kurt Sachs and Hung Hsi Wu. *General relativity for mathematicians.* Graduate Texts in Mathematics, Vol. 48. Springer-Verlag, New York-Heidelberg, 1977, pp. xii+291 (cited p. 39).

[Sam86] Pierre Samuel. *Géométrie projective.* Mathématiques. [Mathematics]. Presses Universitaires de France, Paris, 1986. 176 pp. (cited pp. 81, 101, 105).

[Sam88] Pierre Samuel. *Projective geometry.* Undergraduate Texts in Mathematics. Translated from the French by Silvio Levy, Readings in Mathematics. Springer-Verlag, New York, 1988. x+156. DOI: `10.1007/978-1-4612-3896-6` (cited pp. 81, 101, 105).

[Sid00] Jean-Claude Sidler. *Géométrie projective. Cours, exercices et problèmes corrigés.* 2nd ed. Dunod, 2000. xiii+206 (cited p. 81).

[ske21] sketching4ids. "Circles in 3D". `https://sketching4ids.wordpress.com/circles/`. 2021 (cited p. 111).

[Spi99] Michael Spivak. *A comprehensive introduction to differential geometry. Vol.I–V.* Third. Publish or Perish, Inc., Wilmington, Del., 1999 (cited pp. 12, 275).

[Sti96] John Stillwell. *Sources of hyperbolic geometry.* Vol. 10. History of Mathematics. American Mathematical Society, Providence, RI; London Mathematical Society, London, 1996. x+153. DOI: `10.1090/hmath/010` (cited p. 3).

[Sti10] John Stillwell. *Mathematics and its history.* Third. Undergraduate Texts in Mathematics. Springer, New York, 2010. xxii+660. DOI: `10.1007/978-1-4419-6053-5` (cited p. 81).

[Tho96] A. C. Thompson. *Minkowski geometry.* Vol. 63. Encyclopedia of Mathematics and its Applications. Cambridge University Press, Cambridge, 1996, pp. xvi+346. DOI: `10.1017/CBO9781107325845` (cited p. 32).

[Thu97] William P. Thurston. *Three-dimensional geometry and topology. Vol. 1.* Vol. 35. Princeton Mathematical Series. Edited by Silvio Levy. Princeton University Press, Princeton, NJ, 1997. x+311 (cited pp. 196, 255, 259, 282).

[Tru08] Richard J. Trudeau. *The non-Euclidean revolution.* Modern Birkhäuser Classics. With an introduction by H. S. M. Coxeter, Reprint of the 1987 edition. Birkhäuser Boston, Inc., Boston, MA, 2008. xiv+269 (cited p. 3).

[VY08] Oswald Veblen and John Wesley Young. "A Set of Assumptions for Projective Geometry". In: *Amer. J. Math.* 30.4 (1908), pp. 347–380. DOI: `10.2307/2369956` (cited pp. 84, 85).

[VY65a] Oswald Veblen and John Wesley Young. *Projective geometry. Vol. 1.* Blaisdell Publishing Co. Ginn and Co. New York-Toronto-London, 1965. x+345 (cited p. 85).

[VY65b] Oswald Veblen and John Wesley Young. *Projective geometry. Vol. 2 (by Oswald Veblen).* Blaisdell Publishing Co. Ginn and Co. New York-Toronto-London, 1965. x+511 (cited p. 85).

[Voe05] Jean-Daniel Voelke. *Renaissance de la géométrie non euclidienne entre 1860 et 1900.* Peter Lang, Bern, 2005. xiii+459 (cited p. 3).

[Wan37] Pierre-Laurent Wantzel. "Recherches sur les moyens de reconnaître si un problème de Géométrie peut se résoudre avec la règle et le compas." In: *J. Math. Pures Appl.* 1e série.tome 2 (1837), pp. 366–372 (cited p. 120).

[Whi60] A. N. Whitehead. *The axioms of projective geometry.* Reprinting of Cambridge Tracts in Mathematics and Mathematical Physics, No. 4. Hafner Publishing Co., New York, 1960. viii+64 (cited p. 84).

[Wik21a] Wikipedia contributors. *Cross-ratio — Wikipedia, The Free Encyclopedia.* `https://en.wikipedia.org/w/index.php?title=Cross-ratio&oldid=896577778`. 2021 (cited pp. 101, 128).

[Wik21b] Wikipedia contributors. *Pappus's hexagon theorem — Wikipedia, The Free Encyclopedia.* `https://tinyurl.com/ffbbp9vc`. 2021 (cited p. 126).