Undergraduate Texts in Mathematics

UTM

Nam-Hoon Lee

Geometry: from Isometries to Special Relativity

Springer

Undergraduate Texts in Mathematics

Undergraduate Texts in Mathematics

Undergraduate Texts in Mathematics are generally aimed at third- and fourth-year undergraduate mathematics students at North American universities. These texts strive to provide students and teachers with new perspectives and novel approaches. The books include motivation that guides the reader to an appreciation of interrelations among different aspects of the subject. They feature examples that illustrate key concepts as well as exercises that strengthen understanding.

More information about this series at http://www.springer.com/series/666

Nam-Hoon Lee

Geometry: from Isometries to Special Relativity

Springer

Nam-Hoon Lee
Department of Mathematics Education
Hongik University
Seoul, Korea (Republic of)

ISSN 0172-6056 ISSN 2197-5604 (electronic)
Undergraduate Texts in Mathematics
ISBN 978-3-030-42100-7 ISBN 978-3-030-42101-4 (eBook)
https://doi.org/10.1007/978-3-030-42101-4

Mathematics Subject Classification (2020): 51-01, 51M10, 51B20, 51F15, 83A05, 51P05, 53B30

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

To my daughter, Kyuri

Preface

This book is intended for undergraduate students who have some working knowledge of elementary calculus. It covers the geometries of various two-dimensional homogeneous spaces with metrics—i.e., the Euclidean plane, the sphere, the hyperbolic plane, and the Lorentz–Minkowski plane, which is the two-dimensional spacetime that arises from the theory of special relativity. Our main purpose is to introduce abstract spaces, such as the hyperbolic plane and the Lorentz–Minkowski plane, to undergraduate students as easily and intuitively as possible. We start with very intuitive spaces, the Euclidean plane and the sphere, and focus on a specific structure theorem for isometries (Theorems 1.6, 2.7, 4.20, and 5.7), in which every isometry is a composition of at most three (or four for the Lorentz–Minkowski plane) reflections. This theorem appears in very similar forms in all four spaces, and the proofs for them can also be presented in a similar fashion. We present geometric proofs of these reflection theorems for the first two intuitive spaces and gradually transit to the latter two abstract spaces. Although some terminologies are different and more abstract, the basic structure of the theorem remains the same. Our strategy is to familiarize the readers with the theorems and logic of the proofs first in the two intuitive spaces and to then study the theorems, having similar structures and logic, for the two abstract spaces so that the transition from concrete spaces to abstract spaces can be made easily and smoothly.

In addition to the concentration on the reflection theorem, another difference between this book and other similar books on geometry is that it treats the geometry of special relativity from a truly geometric viewpoint, employing tools that can be used by undergraduate students. Special relativity is an experimentally well confirmed and universally accepted physical theory that explains how space and time are linked. It replaces the primitive notion of an absolute universal time by the notion of a relative time that is not independent of the reference frame and the spatial position. Rather than treating the invariant time and invariant spatial intervals between two events separately, one needs to consider an invariant spacetime interval, which enables us to understand the spacetime from a geometric view of distances and isometries.

It is not easy for undergraduate students to approach the geometry of special relativity. The existing books exploit tools that are too difficult for beginning undergraduates to follow or do not provide a truly geometric viewpoint, depicting only some physical consequences of the theory. We amalgamate special relativity, as another interesting geometry, with classical geometries, noting the following two points.

First, a similar type of reflection theorem holds for the space of special relativity. We repeat a logically similar proof of the theorem given for classical spaces. We think that this similarity helps the reader approach the geometry of special relativity without reluctance. Although we mainly focus on the structure theorems of isometries, readers will also understand other aspects of the geometry of the spaces during their study.

Second, the three-dimensional spacetime of special relativity contains a hypersurface (a hyperboloid) that is isometric with the hyperbolic plane. It is always an interesting task to connect two subjects that initially seem completely different. We concretely explain this connection and give a thorough description. This will help readers prepare for more advanced subjects, such as higher dimensional hyperbolic geometry.

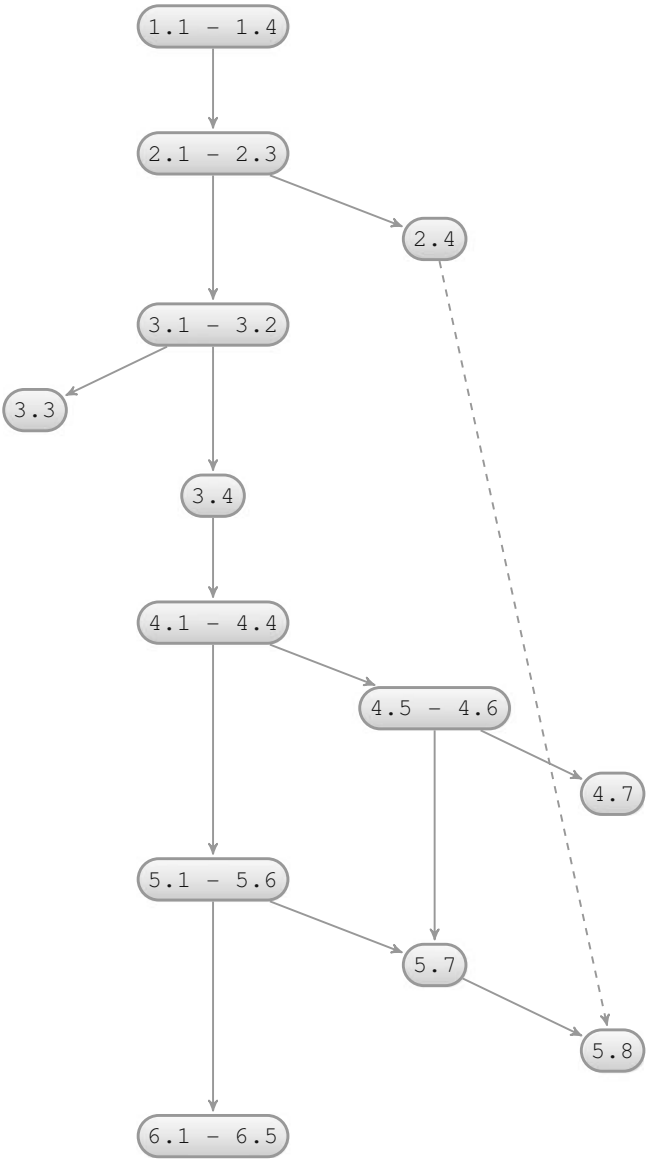
The structure of the book is as follows: In Chapter 1, we introduce the basic geometric terminologies in the familiar Euclidean geometry, such as distance, isometry, translation, reflection, rotation, orientation, and a fixed point. We also introduce the reflection theorem and other structure theorems for isometries. Since our purpose is to introduce abstract spaces to students using similarities to intuitive spaces, we try to keep the discussion within studies on isometries. In Chapter 2, we transit to another classical and intuitive geometry, the surface of the sphere. These theorems and the proofs continue to appeal to geometric intuition. At this point, the readers will start to note that the theorems and proofs for the sphere are very similar to those for the Euclidean plane. Chapter 3 deals with the stereographic projection and inversions. We introduce and prove some properties of inversions. Unlike other books, we try to provide a representation of the geometry of the sphere in the extended plane, following the perspectives of Henri Poincaré, which, we hope, helps the readers understand why we must use abstract metrics for some spaces. In Chapter 4, we deal with hyperbolic geometry. Although the treatment is standard, we try to keep the prerequisite for this chapter minimal. We do not use the notion of complex numbers. In Chapter 5, we introduce the Lorentz–Minkowski plane, which is the two-dimensional space of special relativity. We keep the same structure, i.e., we define similar terminologies and prove the theorem using a similar logical structure. After we are familiarized with the geometry of special relativity, we show that the hyperboloid in three-dimensional spacetime is actually isometric with the hyperbolic plane. Finally, in Chapter 6, we explore the geometry of special relativity. We define basic notions such as causality, worldline, four-vector, and four-momentum. Some basic principles of relativistic kinematics will be touched upon. Solving exercises is the most important part of learning mathematics. The results of some exercises are used in the text and some interesting facts are stated in exercises. The solutions for some selected exercises are provided at the end of the book.

This book has evolved from notes used in courses at Hongik University. I am grateful to all the students who have made various valuable comments on the notes. Particularly, Yeun Kim, Kayun Kim, Jungsoo Nah, and Daun Kim noted several critical errors. My niece, Da Hyee Choi, drew several figures. Several colleagues of mine read it and provided remarks, suggestions, and criticisms. I would like to express my sincere thanks to Prof. Gabjin Yun, Dr. Kyoung-Tark Kim, Dr. Hyejin Park, Sokmin Hong, and Hyelyn Choi. Suggestions, corrections, or comments are most welcome. These may be sent to me at nhlee@kias.re.kr.

Seoul, South Korea Nam-Hoon Lee
March 2020

Contents

Dependence Chart



Chapter 1
Euclidean Plane

"The study of mathematics, like the Nile, begins in minuteness but ends in magnificence."

Charles Caleb Colton (1780–1832)

"Equations are just the boring part of mathematics. I attempt to see things in terms of geometry."

Stephen Hawking (1942–2018)

The word "geometry" is derived from the Greek words geos and metron, meaning earth and measure, whose definition is generally attributed to the fact that the ancient Egyptians regularly utilized geometry to resurvey the fertile farmlands of the Nile river floodplain in late summer. The concepts of "distance" and "area" need not be defined; they are already given by nature. A plane with this concept of distance is called the *Euclidean plane*, denoted by $\mathbb{E}^2$. It does not have special points or directions. When this plane is equipped with a coordinate system, it is given the origin $\mathbf{0}$ and x- and y-axes. The Euclidean plane with a coordinate system can be identified by $\mathbb{R}^2$, the set of all ordered pairs (x, y) of real numbers. We will not distinguish $\mathbb{R}^2$ and $\mathbb{E}^2$ in this book.

1.1 Isometries

For two points $p_1 = (x_1, y_1),\ p_2 = (x_2, y_2) \in \mathbb{R}^2$, the distance between p_1 and p_2 is given by

N.-H. Lee, *Geometry: from Isometries to Special Relativity*, Undergraduate Texts in Mathematics, https://doi.org/10.1007/978-3-030-42101-4_1

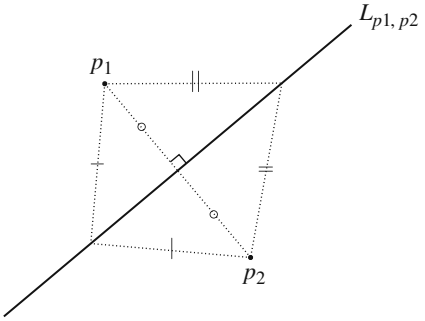


Fig. 1.1 Line L_{p_1,p_2} with respect to the points p_1 and p_2

$$d(p_1, p_2) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}.$$

Note that $d(p_1, p_2) = 0$ if and only if $p_1 = p_2$. Let us start with a definition.

Definition 1.1. A bijective map $\phi : \mathbb{R}^2 \to \mathbb{R}^2$ is called an *isometry* of $\mathbb{R}^2$ if it preserves the distance, i.e.,

$$d(\phi(p_1), \phi(p_2)) = d(p_1, p_2)$$

for any two points $p_1, p_2 \in \mathbb{R}^2$.

"Iso" means "same," and "metry" indicates "measurement," as in the word "geometry."

A line on $\mathbb{R}^2$ can be regarded as the set of points equidistant from the two distinct points p_1 and p_2 (Exercise 1.2, Figure 1.1):

$$L_{p_1,p_2} = \{p \in \mathbb{R}^2 \mid d(p_1, p) = d(p_2, p)\}.$$

Theorem 1.2. *An isometry maps a line to a line.*

Proof. Let $\phi : \mathbb{R}^2 \to \mathbb{R}^2$ be an isometry, and let L be a line on $\mathbb{R}^2$. Then, there are two distinct points p_1 and p_2 such that

$$L = \{p \in \mathbb{R}^2 \mid d(p_1, p) = d(p_2, p)\}.$$

Since ϕ is injective, $\phi(p_1)$ and $\phi(p_2)$ are distinct. Consider a line

$$L' = \{q \in \mathbb{R}^2 \mid d(\phi(p_1), q) = d(\phi(p_2), q)\}.$$

We will show that

$$L' = \phi(L).$$

Let $p \in L$. Note that

$$d(p_1, p) = d(p_2, p).$$

Since ϕ preserves distance,

$$d(\phi(p_1), \phi(p)) = d(p_1, p) = d(p_2, p) = d(\phi(p_2), \phi(p)),$$

i.e.,

$$d(\phi(p_1), \phi(p)) = d(\phi(p_2), \phi(p)),$$

so $\phi(p) \in L'$. Conversely, let $q \in L'$. Note that

$$d(\phi(p_1), q) = d(\phi(p_2), q).$$

Since ϕ is surjective, there exists a point p such that $q = \phi(p)$. Accordingly, the following holds:

$$d(\phi(p_1), \phi(p)) = d(\phi(p_2), \phi(p)).$$

However, as before,

$$d(p_1, p) = d(\phi(p_1), \phi(p)) = d(\phi(p_2), \phi(p)) = d(p_2, p),$$

i.e., $d(p_1, p) = d(p_2, p)$. Therefore, $p \in L$, and then,

$$q = \phi(p) \in \phi(L).$$

□

By Theorem 1.2, an isometry maps a triangle to a triangle. Since an isometry preserves distance, the lengths of the edges in the triangle do not change. Thus, an isometry maps a triangle to a congruent triangle. Because a polygon can be decomposed into a union of triangles, an isometry maps a polygon to a congruent polygon (Figure 1.2). In general, an isometry maps a geometric figure to a congruent one, which will be clear by the end of this chapter. One can consider isometries as maps that preserve all geometric properties of geometric figures.

Example 1.1. Let $\phi(x, y) = (2 - x, y)$. For two points $p_1 = (x_1, y_1)$, $p_2 = (x_2, y_2) \in \mathbb{R}^2$, we have

$$\begin{aligned} d(\phi(p_1), \phi(p_2)) &= \sqrt{((2 - x_1) - (2 - x_2))^2 + (y_1 - y_2)^2} \\ &= \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \\ &= d(p_1, p_2). \end{aligned}$$

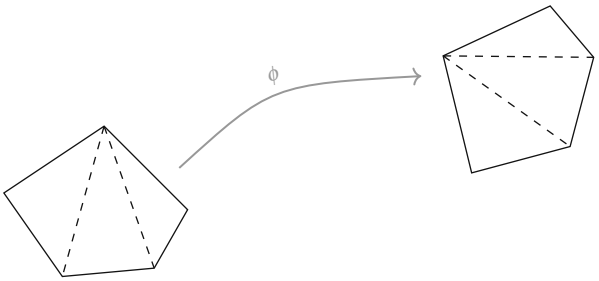


Fig. 1.2 An isometry maps a polygon to a congruent polygon

Therefore, ϕ is an isometry.

Example 1.2. Let $\phi(x, y) = (x, y^3)$. If $p_1 = (0, 0)$ and $p_2 = (0, 2)$, then

$$\begin{aligned} d(\phi(p_1), \phi(p_2)) &= \sqrt{(0-0)^2 + (0-8)^2} \\ &= 8 \neq 2 = \sqrt{(0-0)^2 + (0-2)^2} = d(p_1, p_2), \end{aligned}$$

i.e., $d(\phi(p_1), \phi(p_2)) \neq d(p_1, p_2)$. Therefore, ϕ is not an isometry.

A translation, which moves the entire plane, is also an isometry.

Example 1.3 ($t_{(a,b)}$: translation by vector (a, b)). Let $p = (a, b)$ be a point and

$$t_p(x, y) = t_{(a,b)}(x, y) = (x + a, y + b).$$

Then, t_p is an isometry.

A reflection in a certain line is also an isometry.

Example 1.4 ($\bar{r}$: reflection in the x-axis). Let $\bar{r}(x, y) = (x, -y)$; then, $\bar{r}$ is an isometry.

Example 1.5 (r_θ: rotation by an angle θ about the origin). Let

$$r_\theta(x, y) = (x \cos\theta - y \sin\theta, x \sin\theta + y \cos\theta);$$

then, r_θ is an isometry (Figure 1.3).

Example 1.6 ($r_{p,\theta}$: rotation by an angle θ about a point p). For a point p and an angle θ, let us denote by $r_{p,\theta}$ the counterclockwise rotation by angle θ about point p. Intuitively, this is also an isometry, which can be shown by direct calculation. Note

$$r_{p,\theta} = t_p \circ r_\theta \circ t_{-p}.$$

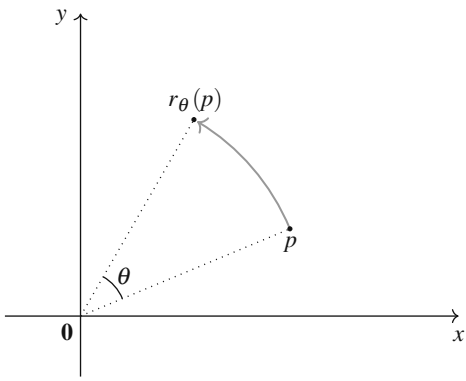


Fig. 1.3 Rotation r_θ by an angle θ about the origin

We leave the proof of the following theorem as an easy exercise for the reader (Exercise 1.7).

Theorem 1.3.

1. *The identity map*

$$\mathrm{id}_{\mathbb{R}^2} : \mathbb{R}^2 \to \mathbb{R}^2$$

   is an isometry.
2. *If ϕ is an isometry of $\mathbb{R}^2$, then ϕ^{-1} is also an isometry.*
3. *If ϕ and ψ are isometries of $\mathbb{R}^2$, then $\phi \circ \psi$ is also an isometry.*

We denote the set of all the isometries of $\mathbb{R}^2$ by $\mathrm{Iso}(\mathbb{R}^2)$.[1]

In addition to these isometries, what other ones are there? The answer is somewhat surprising. Every isometry is a reflection or a composition of reflections. We will illustrate this fact in the coming sections.

Exercises

1.1. Decide whether each of the following maps is an isometry:

(a) $\phi(x, y) = \left(2x, \frac{y}{2}\right)$ (b) $\phi(x, y) = (x, |y|)$

(c) $\phi(x, y) = \frac{1}{\sqrt{2}}(x - y, x + y + 1)$ (d) $\phi(x, y) = \frac{1}{5}(3x + 4y, 4x - 3y)$

1.2. Let L be a line determined by the equation

$$ax + by + c = 0,$$

[1] Theorem 1.3 implies that $\mathrm{Iso}(\mathbb{R}^2)$ forms a "group" together with the composition operation.

where a, b, and c are real numbers with $(a, b) \neq (0, 0)$. Show that there exist two distinct points p_1 and p_2 such that

$$L = \{p \in \mathbb{R}^2 \mid d(p_1, p) = d(p_2, p)\}.$$

1.3. Show that an isometry maps a circle to a circle of the same radius.

1.4. Suppose that a map $\phi : \mathbb{R}^2 \to \mathbb{R}^2$ preserves distance. Show that this map is injective.

1.5. Show that an isometry ϕ is a rotation about the origin if and only if

$$\phi(x, y) = (ax - by, bx + ay)$$

for some real numbers a and b, with $a^2 + b^2 = 1$.

1.6. Explicitly express $r_{p,\theta}(x, y)$ for $p = (a, b)$.

1.7. Prove Theorem 1.3.

1.8. Show the following:

(a) $\bar{r}^{-1} = \bar{r}$,
(b) $t_\alpha \circ t_\beta = t_{\alpha+\beta}$, $\quad t_\alpha^{-1} = t_{-\alpha}$ for $\alpha, \beta \in \mathbb{R}^2$,
(c) $r_\theta \circ r_{\theta'} = r_{\theta+\theta'}$, $\quad r_\theta^{-1} = r_{-\theta}$ for $\theta, \theta' \in \mathbb{R}$.

1.9. A point α in $\mathbb{R}^2$ is said to be *fixed by* a map $f : \mathbb{R}^2 \to \mathbb{R}^2$ if $f(\alpha) = \alpha$. Suppose that two distinct points p and q are fixed points of an isometry ϕ. Show that every point on the line through p, q is a fixed point of ϕ.

1.2 Three Reflections Theorem

An isometry is determined by how it maps three non-collinear points.

Theorem 1.4 (Three points theorem). *Let ϕ and ψ be isometries of $\mathbb{R}^2$. If*

$$\phi(p_1) = \psi(p_1), \phi(p_2) = \psi(p_2) \text{ and } \phi(p_3) = \psi(p_3)$$

for some set of non-collinear points p_1, p_2, and p_3, then $\phi = \psi$.

Proof. The isometry ϕ maps the triangle $\triangle p_1 p_2 p_3$ to a congruent triangle

$$\triangle \phi(p_1)\phi(p_2)\phi(p_3),$$

which implies that the points $\phi(p_1)$, $\phi(p_2)$, and $\phi(p_3)$ are also non-collinear (Figure 1.4).

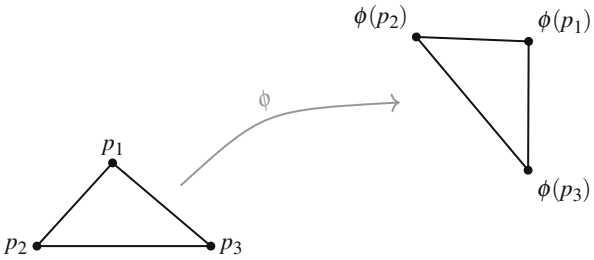


Fig. 1.4 The isometry ϕ maps the triangle $\triangle p_1 p_2 p_3$ to a congruent triangle $\triangle\phi(p_1)\phi(p_2)\phi(p_3)$

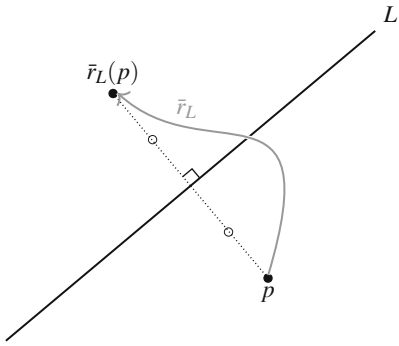


Fig. 1.5 Reflection $\bar{r}_L$ in a line L

Suppose that $\phi \neq \psi$. Then, there exists a point p such that

$$\phi(p) \neq \psi(p),$$

and we can define a line $L = L_{\phi(p),\psi(p)}$. Note that

$$\begin{aligned} d(\phi(p), \phi(p_1)) &= d(p, p_1) && (\because\ \phi \text{ is an isometry}) \\ &= d(\psi(p), \psi(p_1)) && (\because\ \psi \text{ is an isometry}) \\ &= d(\psi(p), \phi(p_1)), \end{aligned}$$

i.e., $d(\phi(p), \phi(p_1)) = d(\psi(p), \phi(p_1))$. Therefore, $\phi(p_1) \in L$. Similarly, we have the following:

$$\phi(p_2) \in L \text{ and } \phi(p_3) \in L.$$

Then, the points $\phi(p_1)$, $\phi(p_2)$, and $\phi(p_3)$ are collinear, which is a contradiction.

□

For a line L, let us denote the reflection in the line L by $\bar{r}_L$. It is then easy to see that $\bar{r}_L$ is an isometry (Figure 1.5).

Remark 1.5. From Figures 1.1 and 1.5, the following is evident:

(a)

$$\bar{r}_{L_{p_1,p_2}}(p_1) = p_2,\ \bar{r}_{L_{p_1,p_2}}(p_2) = p_1.$$

(b)

$$\bar{r}_{L_{p_1,p_2}}(p) = p \text{ for each point } p \in L_{p_1,p_2}.$$

(c)

$$\bar{r}_{L_{p_1,p_2}} \circ \bar{r}_{L_{p_1,p_2}} = \mathrm{id}_{\mathbb{R}^2},\ \text{i.e., } \bar{r}^{-1}_{L_{p_1,p_2}} = \bar{r}_{L_{p_1,p_2}}.$$

Now we can show that every isometry can be expressed as a composition of reflections.

Theorem 1.6 (Three reflections theorem). *An isometry of $\mathbb{R}^2$ is a composition of at most three reflections.*

Proof. Let $\phi : \mathbb{R}^2 \to \mathbb{R}^2$ be an isometry and p_1, p_2, and p_3 be non-collinear points. We divide the situation into four cases.

Case 1. Assume that

$$\phi(p_1) = p_1,\ \phi(p_2) = p_2,\ \phi(p_3) = p_3;$$

then, we have $\phi = \mathrm{id}_{\mathbb{R}^2}$, letting $\psi = \mathrm{id}_{\mathbb{R}^2}$ in Theorem 1.4. Note that $\mathrm{id}_{\mathbb{R}^2} = \bar{r}_L \circ \bar{r}_L$ for every line L.

Case 2. If only two of p_1, p_2, and p_3 coincide with their images under ϕ, say

$$\phi(p_1) = p_1, \phi(p_2) = p_2 \text{ but } \phi(p_3) \neq p_3,$$

then

$$d(p_3, p_1) = d(\phi(p_3), \phi(p_1)) = d(\phi(p_3), p_1).$$

Therefore, letting $L = L_{p_3,\phi(p_3)}$, we have $p_1 \in L$. Similarly, we have $p_2 \in L$. Let $\psi = \bar{r}_L \circ \phi$.

$$\begin{aligned} \psi(p_1) &= \bar{r}_L(\phi(p_1)) \\ &= \bar{r}_L(p_1) \\ &= p_1 \qquad\qquad (\because\ p_1 \in L). \end{aligned}$$

Similarly, we have $\psi(p_2) = p_2$. Note also that

$$\psi(p_3) = \bar{r}_L(\phi(p_3)) = p_3 \qquad (\because\ L = L_{p_3,\phi(p_3)}).$$

Consequently, $\psi = \mathrm{id}_{\mathbb{R}^2}$ by Theorem 1.4 again, i.e., $\bar{r}_L \circ \phi = \mathrm{id}_{\mathbb{R}^2}$. Therefore,

$$\phi = \bar{r}_L^{-1} = \bar{r}_L,$$

and ϕ is a reflection.

Case 3. If only one of p_1, p_2, and p_3 coincides with its image under ϕ, say

$$\phi(p_1) = p_1 \text{ but } \phi(p_2) \neq p_2, \phi(p_3) \neq p_3,$$

then

$$d(p_2, p_1) = d(\phi(p_2), \phi(p_1)) = d(\phi(p_2), p_1).$$

Therefore, letting $M = L_{p_2,\phi(p_2)}$, we have $p_1 \in M$. If $\phi' = \bar{r}_M \circ \phi$, then

$$\phi'(p_1) = \bar{r}_M(\phi(p_1)) = \bar{r}_M(p_1) = p_1$$

and

$$\phi'(p_2) = \bar{r}_M(\phi(p_2)) = p_2.$$

Therefore, we return to Case 1 or Case 2. Therefore, we have $\phi' = \mathrm{id}_{\mathbb{R}^2}$ or $\phi' = \bar{r}_L$ for some line L, i.e., $\phi = \bar{r}_M$ or $\phi = \bar{r}_M \circ \bar{r}_L$.

Case 4. Assume, finally, that

$$\phi(p_1) \neq p_1, \phi(p_2) \neq p_2, \phi(p_3) \neq p_3. \tag{1.1}$$

Let $N = L_{p_1,\phi(p_1)}$ and $\phi'' = \bar{r}_N \circ \phi$. Note that

$$\phi''(p_1) = \bar{r}_N(\phi(p_1)) = p_1.$$

Therefore, we return to Case 1, Case 2, or Case 3 and we have

$$\phi'' = \mathrm{id}_{\mathbb{R}^2}, \phi'' = \bar{r}_M \text{ or } \phi'' = \bar{r}_M \circ \bar{r}_L$$

for some lines L and M, i.e.,

$$\phi = \bar{r}_N, \phi = \bar{r}_N \circ \bar{r}_M \text{ or } \phi = \bar{r}_N \circ \bar{r}_M \circ \bar{r}_L.$$

□

Theorem 1.4 and Theorem 1.6 are the main theorems of this chapter, and very similar theorems will appear for the other types of surfaces, such as the sphere,

the hyperbolic plane, and the Lorentz–Minkowski plane. The proofs are also very similar.

Exercises

1.10. Let L_1 and L_2 be lines. Suppose that $\bar{r}_{L_1} \circ \bar{r}_{L_2} = \mathrm{id}_{\mathbb{R}^2}$. Show that $L_1 = L_2$.

1.11. Suppose that a map $\phi : \mathbb{R}^2 \to \mathbb{R}^2$ preserves distance. Show that this map is bijective.

1.12. Given isometries ϕ and ψ, the *conjugation* of ψ by ϕ is the isometry

$$^{\phi}\psi = \phi \circ \psi \circ \phi^{-1}.$$

(a) For isometries ϕ, ψ, and ξ, show that

$$^{(\phi\circ\psi)}\xi = {}^{\phi}\left({}^{\psi}\xi\right)$$

and

$$^{\phi}(\psi \circ \xi) = \left({}^{\phi}\psi\right) \circ \left({}^{\phi}\xi\right).$$

(b) For reflections $\bar{r}_L$ and $\bar{r}_M$, show that

$$^{\bar{r}_M}\bar{r}_L = \bar{r}_{L'},$$

where $L' = \bar{r}_M(L)$.

(c) For an isometry ϕ, show that

$$^{\phi}\bar{r}_L = \bar{r}_{L'},$$

where $L' = \phi(L)$.

1.13. Two isometries ϕ, ϕ' are said to be *conjugate* if there is an isometry ψ such that $\phi = {}^{\psi}\phi'$. Show that any two reflections are conjugate.

1.14. For two points $p_1 = (x_1, y_1)$, $p_2 = (x_2, y_2) \in \mathbb{R}^2$, recall that the inner product between them is defined as follows:

$$p_1 \cdot p_2 = x_1x_2 + y_1y_2.$$

Then, the norm is $\|p\| = \sqrt{p \cdot p}$.

For a line $L = L_{p_1,p_2}$, show that

$$\bar{r}_L(p) = p - \frac{2(p-u)\cdot v}{\|v\|^2} v$$

for every point $p \in \mathbb{R}^2$, where $u = \frac{1}{2}(p_1 + p_2)$ and $v = \frac{1}{2}(p_1 - p_2)$.

1.3 Rotations and Translations

In the previous section, we proved that every isometry $\mathbb{R}^2$ is a composition of at most three reflections. Therefore, a rotation and a translation should be compositions of reflections. We investigate how they come to be compositions of reflections. Let L and M be two lines, and consider the composition $\bar{r}_M \circ \bar{r}_L$. The situation can be divided into three cases.

First, when $L = M$, trivially, we have $\bar{r}_M \circ \bar{r}_L = \mathrm{id}_{\mathbb{R}^2}$.

Second, when L meets M at a single point p with angle θ, as in Figure 1.6, we claim that

$$\bar{r}_M \circ \bar{r}_L = r_{p,2\theta}.$$

We prove this in the following. Clearly, $\bar{r}_M \circ \bar{r}_L(p) = p = r_{p,2\theta}(p)$. Choose a point p_1 on L and a point p_2 on M that are different from p. From Figure 1.6, we clearly have

$$(\bar{r}_M \circ \bar{r}_L)(p) = p = r_{p,2\theta}(p),$$

$$(\bar{r}_M \circ \bar{r}_L)(p_1) = \bar{r}_M(p_1) = r_{p,2\theta}(p_1)$$

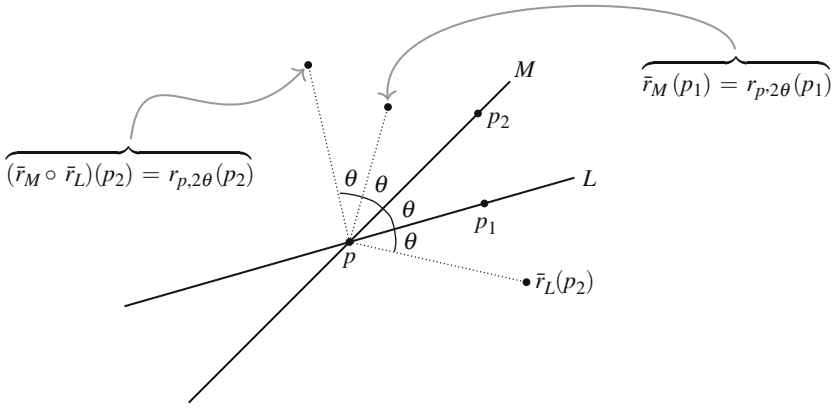


Fig. 1.6 A composition of two reflections in non-disjoint lines is a rotation

and

$$(\bar{r}_M \circ \bar{r}_L)(p_2) = \bar{r}_M(\bar{r}_L(p_2)) = r_{p,2\theta}(p_2).$$

By Theorem 1.4 and noting that the three points p, p_1, and p_2 are not collinear, we conclude that

$$\bar{r}_M \circ \bar{r}_L = r_{p,2\theta}.$$

We note that $M = r_\theta(L)$.

Conversely, it is also clear from Figure 1.6 that every rotation is the composition of two reflections in non-disjoint lines.

Example 1.7. Consider two lines

$$\begin{cases} L: \ x - y = 1 \\ M: \ x + y = 1. \end{cases}$$

Since they meet at $p = (1, 0)$ with angle $\frac{\pi}{2}$, we have

$$\bar{r}_M \circ \bar{r}_L = r_{p,\pi}.$$

Finally, the third case is when L does not meet M (i.e., they are parallel). In this case, we claim that

$$\bar{r}_M \circ \bar{r}_L = t_{2a},$$

where the vector a $(= p_3 - p_2)$ is as shown in Figure 1.7.

Choose two distinct points p_1 and p_2 from the line L and p_3 from the line M such that the line through points p_2 and p_3 meets the lines L and M orthogonally (Figure 1.7). Noting that $\bar{r}_L(p_1) = p_1$, we have

$$(\bar{r}_M \circ \bar{r}_L)(p_1) = \bar{r}_M(p_1) = t_{2a}(p_1),$$

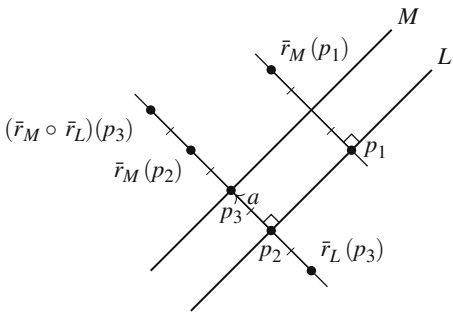
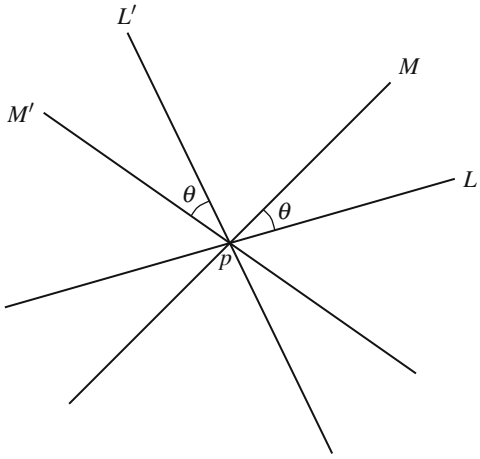


Fig. 1.7 A composition of two reflections in parallel lines is a translation

Fig. 1.8
$\bar{r}_M \circ \bar{r}_L = \bar{r}_{M'} \circ \bar{r}_{L'} = r_{p,2\theta}$



i.e., $(\bar{r}_M \circ \bar{r}_L)(p_1) = t_{2a}(p_1)$. Similarly, we have

$$(\bar{r}_M \circ \bar{r}_L)(p_2) = t_{2a}(p_2).$$

It is not hard to see that

$$(\bar{r}_M \circ \bar{r}_L)(p_3) = t_{2a}(p_3).$$

Since p_1, p_2, and p_3 are not collinear, we conclude that

$$\bar{r}_M \circ \bar{r}_L = t_{2a}.$$

We note also that $M = t_a(L)$.

Conversely, Figure 1.8 implies that every translation is the composition of two reflections in parallel lines.

In summary, we have proved the following theorem.

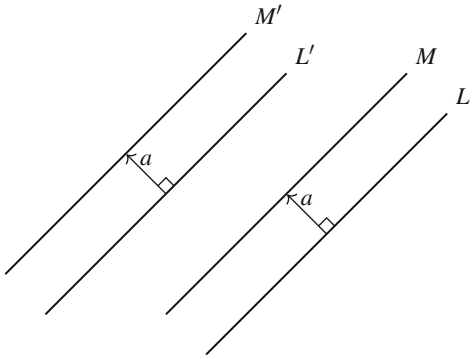
Theorem 1.7.

1. a. *An isometry is a rotation about a point p if and only if it is a composition of two reflections in lines through the point p.*
   b. *Let L, M, L', and M' be lines through a point p. If $r_{p,\theta}(L) = M$ and $r_{p,\theta}(L') = M'$ for some angle θ (Figure 1.8), then*

$$\bar{r}_M \circ \bar{r}_L = \bar{r}_{M'} \circ \bar{r}_{L'} = r_{p,2\theta}.$$

2. a. *An isometry is a translation if and only if it is a composition of two reflections in parallel lines.*

Fig. 1.9
$\bar{r}_M \circ \bar{r}_L = \bar{r}_{M'} \circ \bar{r}_{L'} = t_{2a}$



b. Let L, M, L', and M' be parallel lines such that both the displacement vector from L to M and that from L' to M' are the same (the displacement vector is a) (Figure 1.9). Then,

$$\bar{r}_M \circ \bar{r}_L = \bar{r}_{M'} \circ \bar{r}_{L'} = t_{2a}.$$

Example 1.8. Consider two lines

$$\begin{cases} L: \ x + y = -1 \\ M: \ x + y = 0. \end{cases}$$

The displacement vector from L to M is $a = (\frac{1}{2}, \frac{1}{2})$. Therefore, we have

$$\bar{r}_M \circ \bar{r}_L = t_{2a} = t_{(1,1)}.$$

A composition of translations is again a translation. What about a composition of rotations or a mixture of rotations and translations? The answer is very simple, as stated by the following theorem.

Theorem 1.8. *The set of translations and rotations is closed under composition.*

Proof. We can divide the situation into several cases:

(a) T ∘ T (b) T ∘ R

(c) R ∘ T (d) R ∘ R,

where "T" means translation and "R" means rotation.

The proofs for all the cases are very similar. We will give the proof of (b) as an example and leave the proofs for the other cases as easy exercises.

Consider a composition of a translation and a rotation, e.g., $t_q \circ r_{p,\theta}$, where p and q are some points and θ is an angle. Note that there are some non-parallel lines H and L such that $r_{p,\theta} = \bar{r}_L \circ \bar{r}_H$. Also by Theorem 1.7, there are some parallel

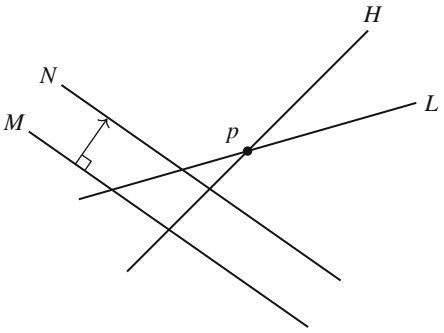


Fig. 1.10 Parallel lines M and N such that $t_q = \bar{r}_N \circ \bar{r}_M$

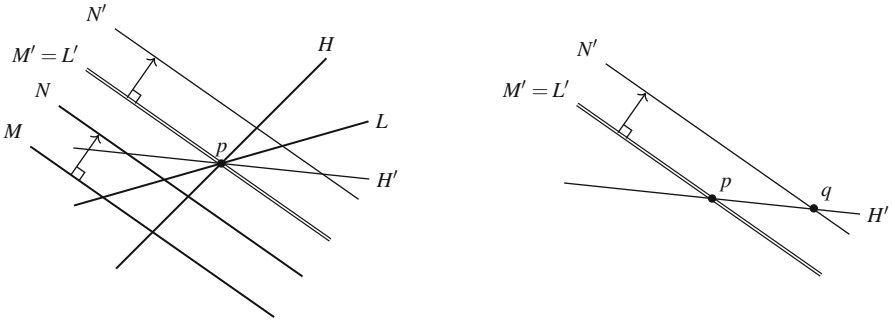


Fig. 1.11 Lines H, H', L, L', M, M', N, and N'

lines M and N such that $t_q = \bar{r}_N \circ \bar{r}_M$ (Figure 1.10). Translate the lines M and N to lines M' and N' together so that M' goes through the point p $(= H \cap L)$ (the left diagram in Figure 1.11). Then,

$$t_q = \bar{r}_N \circ \bar{r}_M = \bar{r}_{N'} \circ \bar{r}_{M'}$$

by Theorem 1.7. Rotate the lines H and L about p together to lines H' and L', respectively, so that L' coincides with M' (the right diagram in Figure 1.11). Then,

$$r_{p,\theta} = \bar{r}_L \circ \bar{r}_H = \bar{r}_{L'} \circ \bar{r}_{H'}.$$

See the right diagram in Figure 1.11. Thus,

$$\begin{aligned} t_q \circ r_{p,\theta} &= \bar{r}_{N'} \circ \bar{r}_{M'} \circ \bar{r}_{L'} \circ \bar{r}_{H'} \\ &= \bar{r}_{N'} \circ \bar{r}_{H'} && (\because\ M' = L') \\ &= r_{q,\theta}, \end{aligned}$$

where $\{q\} = N' \cap H'$. Thus, $t_q \circ r_{p,\theta}$ is a rotation.

□

Exercises

1.15. Prove or disprove the following:

(a) $\bar{r}_M \circ \bar{r}_L = \bar{r}_L \circ \bar{r}_M$,
(b) $t_{p_2} \circ t_{p_1} = t_{p_1} \circ t_{p_2}$,
(c) $r_{\theta_2} \circ r_{\theta_1} = r_{\theta_1} \circ r_{\theta_2}$,
(d) $r_{p_2,\theta_2} \circ r_{p_1,\theta_1} = r_{p_1,\theta_1} \circ r_{p_2,\theta_2}$,
(e) $t_\alpha \circ \bar{r}_L = \bar{r}_L \circ t_\alpha$,
(f) $r_{\alpha,\theta} \circ \bar{r}_L = \bar{r}_L \circ r_{\alpha,\theta}$, where L and M are lines; p_1, p_2, and α are points; and θ_1, θ_2, and θ are angles.

1.16. Suppose that two reflections $\bar{r}_{L_1}$ and $\bar{r}_{L_2}$ satisfy $\bar{r}_{L_1} \circ \bar{r}_{L_2} = \bar{r}_{L_2} \circ \bar{r}_{L_1}$. Show that $L_1 = L_2$ or that L_1 and L_2 are orthogonal to each other.

1.17. (a) For a reflection $\bar{r}_L$ in a line L through a point p and a rotation $r_{p,\theta}$, show that

$$^{(\bar{r}_L)}\left(r_{p,\theta}\right) = r_{p,-\theta}.$$

(b) For an isometry ϕ that fixes a point p, show that

$$^{\phi}r_{p,\theta} = r_{p,\theta} \text{ or } r_{p,-\theta}.$$

1.18. Let ϕ and ψ be isometries of $\mathbb{R}^2$. Suppose that

$$\phi(p_1) = \psi(p_1) \text{ and } \phi(p_2) = \psi(p_2)$$

for some set of two distinct points p_1 and p_2. Show that

$$\phi = \psi \text{ or } \phi \circ \bar{r}_L = \psi,$$

where L is the line going through points p_1 and p_2.

1.19. A *halfturn* σ_p for a point $p \in \mathbb{R}^2$ is a rotation by the angle π about the point p. Prove the following:

(a) The composition of two halfturns is a translation.

(b) Every translation is a composition of two halfturns.

(c) If p_2 is the midpoint of points p_1 and p_3, then

$$\sigma_{p_2} \circ \sigma_{p_1} = \sigma_{p_3} \circ \sigma_{p_2}.$$

(d) A composition of three halfturns is a halfturn. In particular, if three points p_1, p_2, and p_3 are non-collinear, then $\sigma_{p_3} \circ \sigma_{p_2} \circ \sigma_{p_1} = \sigma_p$, where $\square p_1 p_2 p_3 p$ is a parallelogram.

(e) $\sigma_{p_3} \circ \sigma_{p_2} \circ \sigma_{p_1} = \sigma_{p_1} \circ \sigma_{p_2} \circ \sigma_{p_3}$ for any three points p_1, p_2, and p_3.

1.4 Glide Reflections and Orientation

We have seen that a composition of two reflections is either a rotation or a translation. Let us now consider a composition of three reflections.

Example 1.9 (Glide reflection). Let L be a line and α be a vector parallel to L. Then, we consider an isometry

$$\bar{g}_{L,\alpha} = t_\alpha \circ \bar{r}_L,$$

which we call a *glide reflection*. A glide reflection is a reflection followed by a translation parallel to the reflection line (Figure 1.12).

If $\alpha = \mathbf{0}$, then $\bar{g}_{L,\alpha}$ is a reflection. Therefore, every reflection is also a glide reflection.

Since t_α can factor into the composition $\bar{r}_N \circ \bar{r}_M$ of two reflections $\bar{r}_N$ and $\bar{r}_M$, a glide reflection is the composition of three reflections,

$$\bar{g}_{L,\alpha} = \bar{r}_N \circ \bar{r}_M \circ \bar{r}_L,$$

where the lines L, M, and N are as shown in Figure 1.12. The converse also holds, as stated in the following theorem.

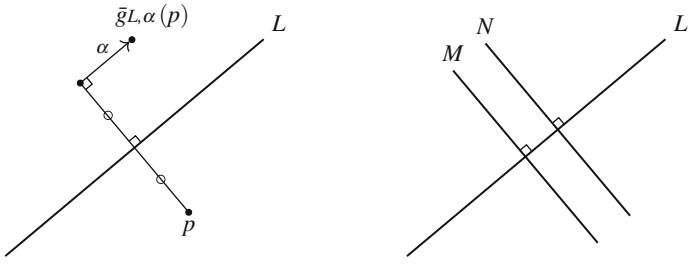


Fig. 1.12 $\bar{g}_{L,\alpha} = \bar{r}_N \circ \bar{r}_M \circ \bar{r}_L$

Theorem 1.9. *The composition of three reflections is a glide reflection.*

Proof. Let ϕ be the composition of reflections in lines L, M, and N:

$$\phi = \bar{r}_N \circ \bar{r}_M \circ \bar{r}_L.$$

For the proof, we must consider several cases, determining whether the lines L, M, and N are parallel to another.

Case 1. Assume that the lines L, M are orthogonal to each other. Let p be the point where the lines L, M meet and M' be a line through the point p that is parallel to the line N. Then there is another line L' through the point p such that

$$\bar{r}_{M'} \circ \bar{r}_{L'} = \bar{r}_M \circ \bar{r}_L.$$

Note that the line L' is orthogonal to M'. Now

$$\phi = \bar{r}_N \circ \bar{r}_M \circ \bar{r}_L = \bar{r}_N \circ \bar{r}_{M'} \circ \bar{r}_{L'},$$

which is a glide reflection along the line L.

Case 2. Assume that both the lines M and N pass through a point p. One can choose a line M' through the point p that is orthogonal to the line L. Then, there is another line N' through the point p such that

$$\bar{r}_{N'} \circ \bar{r}_{M'} = \bar{r}_N \circ \bar{r}_M.$$

See Figure 1.13.

$$\phi = \bar{r}_N \circ \bar{r}_M \circ \bar{r}_L = \bar{r}_{N'} \circ \bar{r}_{M'} \circ \bar{r}_L$$

Now, we have a configuration of lines L, M', N' that belongs to Case 1 and we conclude that ϕ is a glide reflection.

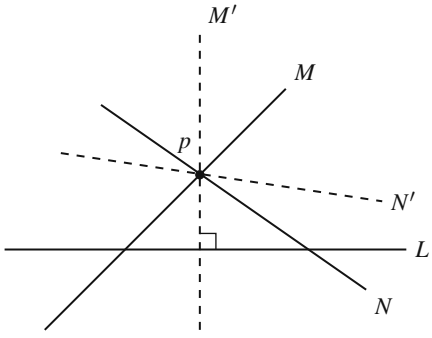


Fig. 1.13 Non-parallel lines M and N, intersecting at the point p

Case 3. Assume that the lines M and N are parallel to each other. If L is also parallel to them, we can choose another parallel line N' such that

$$\bar{r}_N \circ \bar{r}_M = \bar{r}_{N'} \circ \bar{r}_L.$$

Then,

$$\phi = \bar{r}_N \circ \bar{r}_M \circ \bar{r}_L = \bar{r}_{N'} \circ \bar{r}_L \circ \bar{r}_L = \bar{r}_{N'},$$

which is a reflection.

If L is not parallel to them, the lines L, M meet at a point p. Rotate both the lines L, M about the point p to get L', M', respectively, so that

$$\bar{r}_{M'} \circ \bar{r}_{L'} = \bar{r}_M \circ \bar{r}_L$$

and the line N meets with M' at a point.

$$\phi = \bar{r}_N \circ \bar{r}_M \circ \bar{r}_L = \bar{r}_N \circ \bar{r}_{M'} \circ \bar{r}_{L'}.$$

Now, we have a configuration of lines L', M', N that belongs to Case 2 and we conclude that ϕ is a glide reflection. □

By using Theorem 1.6, together with Theorem 1.7 and Theorem 1.9, we obtain the following classification theorem.

Theorem 1.10 (Classification of isometries of $\mathbb{R}^2$). *An isometry of the Euclidean plane is a rotation, a translation, or a glide reflection.*

Proof. By Theorem 1.6, an isometry is a composition of at most three reflections. By Theorem 1.7 and Theorem 1.9, it is a rotation, a translation, or a glide reflection. □

Theorem 1.6 gives us much information about an isometry. For example, it is now very clear that an isometry maps a geometric figure to a congruent geometric figure because a reflection does. Intuitively speaking, two geometric figures are regarded as congruent if they have the same shape and size or if one has the same shape and size as the mirror image of the other. Now we give a more formal but more precise definition of congruence.

Definition 1.11. Two sets of points in the Euclidean plane are said to be *congruent* if and only if one is the image of the other under some isometry of the Euclidean plane.

Let γ be a curve with length $l(\gamma)$. For a reflection $\bar{r}_L$, it is intuitively clear that

$$l(\bar{r}_L(\gamma)) = l(\gamma),$$

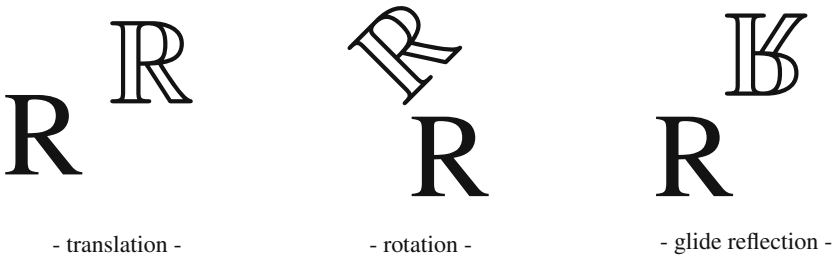


Fig. 1.14 Transformations of the figure of the letter "R" by isometries

i.e., a reflection preserves the length of a curve. Again by Theorem 1.6, an isometry is a composition of reflections; therefore, it also preserves the length of every curve.

A reflection maps a figure to its mirror image. Figure 1.14 shows how the figure of the letter "R" is transformed by various isometries. Different from translation and rotation, the letter transformed by a glide reflection is not exactly the letter "R" even if the reader rotates or moves this book. The reason is that a glide reflection is the composition of an *odd* number of reflections.

Definition 1.12.

a) An isometry that is a composition of an even number of reflections is said to be *orientation-preserving*.
b) An isometry that is a composition of an odd number of reflections is said to be *orientation-reversing*.

Let $\mathrm{Iso}^+(\mathbb{R}^2)$ and $\mathrm{Iso}^-(\mathbb{R}^2)$ be the sets of orientation-preserving isometries and orientation-reversing isometries, respectively.

Theorem 1.13. $\mathrm{Iso}^+(\mathbb{R}^2)$ *consists of translations and rotations, and* $\mathrm{Iso}^-(\mathbb{R}^2)$ *consists of glide reflections.*

Proof. Translations and rotations are orientation-preserving since they are compositions of two reflections. Conversely, consider an orientation-preserving isometry,

$$\phi = \bar{r}_{L_1} \circ \bar{r}_{L_2} \circ \cdots \circ \bar{r}_{L_{2n}}.$$

Let $\phi_i = \bar{r}_{L_{2i-1}} \circ \bar{r}_{L_{2i}}$, which is either a translation or a rotation. Note that

$$\phi = \phi_1 \circ \phi_2 \circ \cdots \circ \phi_n.$$

By Theorem 1.8, it is either a rotation or a translation. This proves the first statement.

Glide reflections are orientation-reversing since they are compositions of *three* reflections. Conversely, consider an orientation-reversing isometry

$$\phi = \bar{r}_{L_1} \circ \bar{r}_{L_2} \circ \cdots \circ \bar{r}_{L_{2n+1}}.$$

Let $\phi_i = \bar{r}_{L_{2i-1}} \circ \bar{r}_{L_{2i}}$, which is either a translation or a rotation. Let

$$\phi' = \bar{r}_{L_1} \circ \bar{r}_{L_2} \circ \cdots \circ \bar{r}_{L_{2n}} = \phi_1 \circ \phi_2 \circ \cdots \circ \phi_n.$$

By Theorem 1.8, ϕ' is either a rotation or a translation, which is a composition of two reflections. We thus have $\phi' = \bar{r}_M \circ \bar{r}_L$ for some two lines L and M. Therefore,

$$\phi = \left(\bar{r}_{L_1} \circ \bar{r}_{L_2} \circ \cdots \circ \bar{r}_{L_{2n}}\right) \circ \bar{r}_{L_{2n+1}} = \phi' \circ \bar{r}_{L_{2n+1}} = \bar{r}_M \circ \bar{r}_L \circ \bar{r}_{L_{2n+1}},$$

which is a glide reflection by Theorem 1.9. □

An orientation-reversing isometry, which is a glide reflection, cannot be realized as actual motions of physical objects (if the objects are forced to remain in a plane). By contrast, we frequently experience rotations and translations of objects. For this reason, an orientation-preserving isometry is also called a *rigid motion*.

Theorem 1.14. *The sets* $\mathrm{Iso}^+(\mathbb{R}^2)$ *and* $\mathrm{Iso}^-(\mathbb{R}^2)$ *are disjoint.*

Proof. For the proof, it is very useful to examine the fixed points of isometries. The set of fixed points of a non-trivial rotation (i.e., not the identity) is composed of a single point. The set of fixed points of a non-trivial translation is empty. The set of fixed points of a glide reflection is empty or is a line if it is a reflection. Therefore, if there exists an isometry ϕ in $\mathrm{Iso}^+(\mathbb{R}^2) \cap \mathrm{Iso}^-(\mathbb{R}^2)$, its set of fixed points is empty. Therefore, ϕ would be a translation, a composition of two reflections

$$\phi = \bar{r}_{L_1} \circ \bar{r}_{L_2},$$

and a glide reflection, a composition of three reflections

$$\phi = \bar{r}_{M_1} \circ \bar{r}_{M_2} \circ \bar{r}_{M_3}.$$

Therefore, we have

$$\phi = \bar{r}_{L_1} \circ \bar{r}_{L_2} = \bar{r}_{M_1} \circ \bar{r}_{M_2} \circ \bar{r}_{M_3},$$

and so

$$\bar{r}_{L_2} \circ \bar{r}_{L_1} \circ \bar{r}_{M_1} \circ \bar{r}_{M_2} \circ \bar{r}_{M_3} = \mathrm{id}_{\mathbb{R}^2}.$$

The right-hand-side identity map fixes every point; however, the left-hand-side isometry is a glide reflection whose set of fixed points is a line or empty, resulting in a contradiction. Therefore, $\mathrm{Iso}^+(\mathbb{R}^2)$ and $\mathrm{Iso}^-(\mathbb{R}^2)$ are disjoint. □

Exercises

1.20. Show that an isometry with exactly one fixed point is a rotation.

1.21. Show that an isometry is a glide reflection if and only if it is conjugate to $t_{(a,0)} \circ \bar{r}$ for some $a \in \mathbb{R}$.

1.22. Show that the inverse of every glide reflection is also a glide reflection.

1.23. Let ϕ be a glide reflection. Show that ϕ^2 is a translation.

1.24. Let ϕ be an isometry. A line L is said to be *invariant under* ϕ if $\phi(L) = L$. Show the following:

(a) ϕ has no invariant lines if and only if it is a rotation by $a\pi$ for some $a \notin \mathbb{Z}$.

(b) ϕ has a single invariant line if and only if it is a glide reflection with a non-zero shift.

(c) If ϕ has multiple invariant lines, it has infinitely many ones.

1.25.

(a) Classify all the isometries such that $\phi^2 = \phi \circ \phi = \mathrm{id}_{\mathbb{R}^2}$.

(b) Classify all the isometries such that $\phi^6 = \mathrm{id}_{\mathbb{R}^2}$.

1.26. Prove that there exists an isometry that cannot be expressed as a composition of one or two reflections.

1.27. For $x, y \in \mathbb{R}$, we define the distance in the usual way:

$$d(x, y) = |x - y|.$$

We define the isometries of $\mathbb{R}$ accordingly. Then, a reflection $\bar{r}_a$ in $a \in \mathbb{R}$ is given by

$$\bar{r}_a = 2a - x.$$

Show that every isometry of $\mathbb{R}$ is a composition of at most two reflections.

Chapter 2
Sphere

"The description of right lines and circles, upon which geometry is founded, belongs to mechanics. Geometry does not teach us to draw these lines, but requires them to be drawn."

Isaac Newton (1642–1727)

"Geometry is the art of correct reasoning from incorrectly drawn figures."

Henri Poincaré (1854–1912)

Spherical geometry is almost as old as Euclidean geometry. In fact, the word geometry means 'measurement of the Earth', and the Earth is (more or less) a sphere. The ancient Greek geometers knew that the Earth was spherical. Navigation motivated the study of spherical geometry because, even 2000 years ago, the fact that the earth is curved had a noticeable effect on cartography. In spherical geometry, the 'points' are points on the surface of the sphere. We are not concerned with the 'inside' of the sphere.

2.1 The Sphere $\mathbb{S}^2$ in $\mathbb{R}^3$

The (unit) sphere is

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

N.-H. Lee, *Geometry: from Isometries to Special Relativity*, Undergraduate Texts in Mathematics, https://doi.org/10.1007/978-3-030-42101-4_2

A *great circle* is a circle on the sphere that divides the sphere into two equal hemispheres. A great circle can also be defined as the intersection of the sphere and a plane that goes through the origin $\mathbf{0}$. Two points that are diametrically opposite on the sphere are called *antipodal points*. In spherical geometry, two points determine a great circle unless they are antipodal points, in which case there are infinitely many great circles joining them.

For two points $p_1 = (x_1, y_1, z_1)$ and $p_2 = (x_2, y_2, z_2)$ in $\mathbb{S}^2$, the distance between them in $\mathbb{R}^3$ is

$$d(p_1, p_2) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2}.$$

The *spherical distance* $d_{\mathbb{S}^2}(p_1, p_2)$ is the arc length of the shortest path on $\mathbb{S}^2$ from p_1 to p_2, which is a segment of the great circle through the points p_1 and p_2 (Figure 2.1).

Note that

$$d_{\mathbb{S}^2}(p_1, p_2) = \angle p_1 \mathbf{0} p_2 = 2\theta = 2 \arcsin\left(\frac{1}{2} d(p_1, p_2)\right), \tag{2.1}$$

where $\theta = \frac{1}{2}\angle p_1 \mathbf{0} p_2$ (Figure 2.2).

A plane in $\mathbb{R}^3$ is a set of points defined by a linear equation,

$$ax + by + cz = d,$$

where a, b, c, and d are constants with $(a, b, c) \neq (0, 0, 0)$. Similar to a line in the Euclidean plane, a plane in $\mathbb{R}^3$ can be regarded as the set of points equidistant from two distinct points p_1 and p_2 in $\mathbb{R}^3$:

$$P_{p_1, p_2} = \{p \in \mathbb{R}^3 \mid d(p_1, p) = d(p_2, p)\}.$$

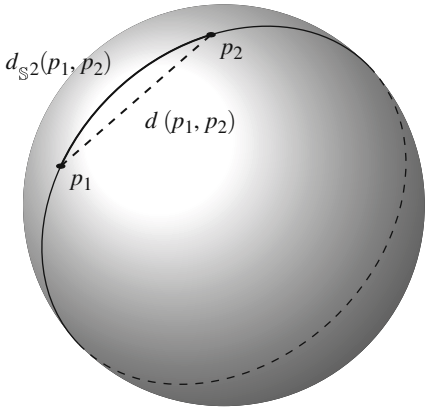


Fig. 2.1 Spherical distance $d_{\mathbb{S}^2}(p_1, p_2)$ between the two points p_1 and p_2 on $\mathbb{S}^2$

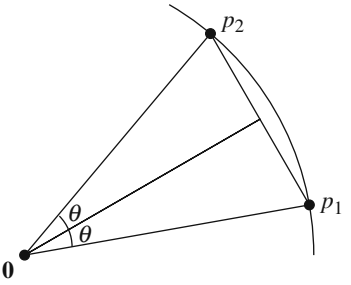


Fig. 2.2 $d_{\mathbb{S}^2}(p_1, p_2) = \angle p_1 \mathbf{0} p_2 = 2\theta = 2 \arcsin\left(\frac{1}{2} d(p_1, p_2)\right)$

Let

$$p_1 = (a_1, b_1, c_1) \text{ and } p_2 = (a_2, b_2, c_2).$$

A point $p = (x, y, z)$ belongs to P_{p_1, p_2} if and only if $d(p_1, p) = d(p_2, p)$, i.e.,

$$\sqrt{(a_1 - x)^2 + (b_1 - y)^2 + (c_1 - z)^2} = \sqrt{(a_2 - x)^2 + (b_2 - y)^2 + (c_2 - z)^2}$$

or

$$(a_1 - a_2)x + (b_1 - b_2)y + (c_1 - c_2)z - d = 0,$$

which is a linear equation, where $d = \frac{1}{2}((a_1^2 + b_1^2 + c_1^2) - (a_2^2 + b_2^2 + c_2^2))$. For a point $p = (x, y, z)$ in $\mathbb{R}^3$, the norm of p is defined by

$$\|p\| = \sqrt{x^2 + y^2 + z^2}.$$

Then, we obviously have $d(p, q) = \|p - q\|$.

Now let us define the notion of maps that preserve the geometric properties of figures on the sphere.

Definition 2.1. A bijective map $\phi : \mathbb{S}^2 \to \mathbb{S}^2$ is called an *isometry* of $\mathbb{S}^2$ if it preserves the spherical distance, i.e.,

$$d_{\mathbb{S}^2}(\phi(p_1), \phi(p_2)) = d_{\mathbb{S}^2}(p_1, p_2)$$

for any two points $p_1, p_2 \in \mathbb{S}^2$.

The main purpose of this chapter is to classify the isometries of the sphere as we did for the Euclidean plane in the previous chapter. In doing so, we will be able to gain considerable knowledge about the spherical geometry.

A bijective map $\phi : \mathbb{R}^3 \to \mathbb{R}^3$ is called an *isometry* of $\mathbb{R}^3$ if the distance is preserved, i.e.,

$$d(\phi(p_1), \phi(p_2)) = d(p_1, p_2)$$

for any two points $p_1, p_2 \in \mathbb{R}^3$.

According to (2.1),

$$d_{\mathbb{S}^2}(\phi(p_1), \phi(p_2)) = d_{\mathbb{S}^2}(p_1, p_2)$$

if and only if

$$2 \arcsin\left(\frac{1}{2} d(\phi(p_1), \phi(p_2))\right) = 2 \arcsin\left(\frac{1}{2} d(p_1, p_2)\right),$$

i.e.,

$$d(\phi(p_1), \phi(p_2)) = d(p_1, p_2).$$

In summary,

$$d_{\mathbb{S}^2}(\phi(p_1), \phi(p_2)) = d_{\mathbb{S}^2}(p_1, p_2) \quad \Leftrightarrow \quad d(\phi(p_1), \phi(p_2)) = d(p_1, p_2). \tag{2.2}$$

Therefore, isometries of the sphere and isometries of $\mathbb{R}^3$ are closely related.

Proposition 2.2. *If an isometry of $\mathbb{R}^3$ fixes the origin $\mathbf{0}$, then it induces an isometry of $\mathbb{S}^2$. In other words, if an isometry ϕ of $\mathbb{R}^3$ satisfies*

$$\phi(\mathbf{0}) = \mathbf{0},$$

then the map $\psi : \mathbb{S}^2 \to \mathbb{S}^2$, given by $\psi(p) = \phi(p)$, is well defined and an isometry of the sphere.

Proof. First, we must verify that $\psi(p)$ belongs to the sphere if the point p does. This is shown by the following:

$$d(\psi(p), \mathbf{0}) = d(\phi(p), \mathbf{0}) = d(\phi(p), \phi(\mathbf{0})) = d(p, \mathbf{0}) = 1.$$

Since ϕ^{-1} is also an isometry of $\mathbb{R}^3$, the map $\psi' : \mathbb{S}^2 \to \mathbb{S}^2$ by $\psi'(p) = \phi^{-1}(p)$ is also well defined. Then, ψ' is the inverse map of ψ, and so ψ is bijective. Finally, by (2.2), ψ preserves the spherical distance because ϕ preserves the distance. □

We denote the set of all the isometries of $\mathbb{S}^2$ by $\mathrm{Iso}(\mathbb{S}^2)$. In Chapter 1, we saw that reflections in lines play important roles when studying isometries. We define a *spherical line* for two distinct points p_1 and p_2 in $\mathbb{S}^2$ (Figure 2.3) as follows:

$$G_{p_1,p_2} = \{p \in \mathbb{S}^2 \mid d_{\mathbb{S}^2}(p_1, p) = d_{\mathbb{S}^2}(p_2, p)\}.$$

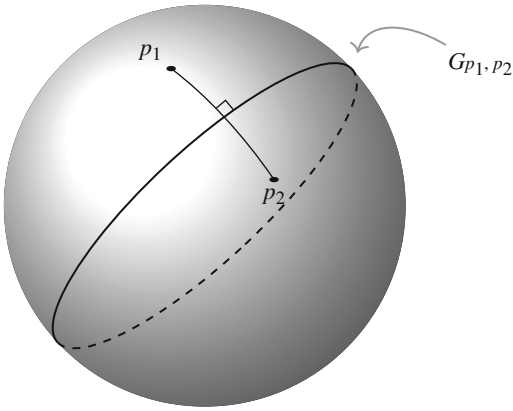


Fig. 2.3 Spherical line G_{p_1,p_2} with respect to points p_1 and p_2 on $\mathbb{S}^2$

Proposition 2.3. *A spherical line is a great circle.*

Proof. For two distinct points p_1 and p_2 on $\mathbb{S}^2$, consider a spherical line

$$G_{p_1,p_2} = \{p \in \mathbb{S}^2 \mid d_{\mathbb{S}^2}(p_1, p) = d_{\mathbb{S}^2}(p_2, p)\}.$$

By (2.2),

$$\begin{aligned} G_{p_1,p_2} &= \{p \in \mathbb{S}^2 \mid d(p_1, p) = d(p_2, p)\} \\ &= \mathbb{S}^2 \cap \{p \in \mathbb{R}^3 \mid d(p_1, p) = d(p_2, p)\} \\ &= \mathbb{S}^2 \cap P_{p_1,p_2}, \end{aligned}$$

which is a intersection of the sphere and a plane P_{p_1,p_2}. The plane P_{p_1,p_2} passes through the origin because

$$d(p_1, \mathbf{0}) = 1 = d(p_2, \mathbf{0}),$$

i.e., $d(p_1, \mathbf{0}) = d(p_2, \mathbf{0})$. □

Points on the sphere are said to be *collinear* if they lie on a great circle.

Lemma 2.4. *An isometry of the sphere maps non-collinear points p_1, p_2, and p_3 to non-collinear points.*

Proof. Let $\phi : \mathbb{S}^2 \to \mathbb{S}^2$ be an isometry. Suppose that the points $\phi(p_1)$, $\phi(p_2)$, and $\phi(p_3)$ are collinear, then they lie on a great circle $G_{p,q}$. Thus, we have

$$\begin{aligned} d_{\mathbb{S}^2}(p, \phi(p_1)) &= d_{\mathbb{S}^2}(q, \phi(p_1)), \\ d_{\mathbb{S}^2}(p, \phi(p_2)) &= d_{\mathbb{S}^2}(q, \phi(p_2)) \end{aligned}$$

and

$$d_{\mathbb{S}^2}(p, \phi(p_3)) = d_{\mathbb{S}^2}(q, \phi(p_3)).$$

However, then,

$$d_{\mathbb{S}^2}(\phi^{-1}(p), p_1) = d_{\mathbb{S}^2}(\phi^{-1}(q), p_1),$$

$$d_{\mathbb{S}^2}(\phi^{-1}(p), p_2) = d_{\mathbb{S}^2}(\phi^{-1}(q), p_2)$$

and

$$d_{\mathbb{S}^2}(\phi^{-1}(p), p_3) = d_{\mathbb{S}^2}(\phi^{-1}(q), p_3),$$

which imply that the points p_1, p_2, and p_3 lie on the great circle $G_{\phi^{-1}(p),\phi^{-1}(q)}$. Because this is a contradiction, the points $\phi(p_1)$, $\phi(p_2)$, and $\phi(p_3)$ are non-collinear. □

In Theorem 1.4, we saw that an isometry of $\mathbb{R}^2$ is determined by how it maps three non-collinear points. The following theorem is a version of Theorem 1.4 for the spherical geometry.

Theorem 2.5 (Three points theorem for the sphere). *Let ϕ and ψ be isometries of the sphere. If*

$$\phi(p_1) = \psi(p_1), \phi(p_2) = \psi(p_2) \text{ and } \phi(p_3) = \psi(p_3)$$

for some set of non-collinear points p_1, p_2, and p_3, then $\phi = \psi$.

Proof. According to Lemma 2.4, the three points $\phi(p_1)$, $\phi(p_2)$, and (p_3) are also non-collinear.

Suppose that $\phi \neq \psi$. Then, there exists a point p such that

$$\phi(p) \neq \psi(p),$$

and we can define a great circle $G = G_{\phi(p),\psi(p)}$. Note that

$$\begin{aligned} d_{\mathbb{S}^2}(\phi(p), \phi(p_1)) &= d_{\mathbb{S}^2}(p, p_1) && (\because \ \phi \text{ is an isometry}) \\ &= d_{\mathbb{S}^2}(\psi(p), \psi(p_1)) && (\because \ \psi \text{ is an isometry}) \\ &= d_{\mathbb{S}^2}(\psi(p), \phi(p_1)), \end{aligned}$$

i.e., $d_{\mathbb{S}^2}(\phi(p), \phi(p_1)) = d_{\mathbb{S}^2}(\psi(p), \phi(p_1))$. So $\phi(p_1) \in G$. Similarly,

$$\phi(p_2) \in G \text{ and } \phi(p_3) \in G.$$

However, the three points $\phi(p_1)$, $\phi(p_2)$, and (p_3) are then collinear, which is a contradiction. □

Exercises

2.1. Let $p_1 = \frac{1}{\sqrt{2}}(1, 1, 0)$ and $p_2 = \frac{1}{\sqrt{2}}(1, 0, 1)$. Calculate the following:

(a) $d(p_1, p_2)$ (b) $d_{\mathbb{S}^2}(p_1, p_2)$.

2.2. (a) Show that there is no map $f : \mathbb{R}^2 \to \mathbb{S}^2$ such that

$$d(p, q) = d_{\mathbb{S}^2}(f(p), f(q))$$

for all points p and q in $\mathbb{R}^2$.

(b) Show that there is no map $g : \mathbb{S}^2 \to \mathbb{R}^2$ such that

$$d_{\mathbb{S}^2}(p, q) = d(g(p), g(q))$$

for all points p and q in $\mathbb{S}^2$.

2.3. A spherical circle C of radius ρ with its center $p \in \mathbb{S}^2$ is defined by

$$C = \{x \in \mathbb{S}^2 \mid d_{\mathbb{S}^2}(p, x) = \rho\}.$$

Determine its circumference $\Pi(\rho)$. Show that $\Pi(\rho) < 2\pi\rho$ and

$$\lim_{\rho \to 0^+} \frac{\Pi(\rho)}{\rho} = 2\pi.$$

2.2 Isometries of the Sphere $\mathbb{S}^2$

The reflection $\bar{r}_{P_{p_1,p_2}}$ in the plane P_{p_1,p_2} is an isometry of $\mathbb{R}^3$ such that the line segment from the given point p in $\mathbb{R}^3$ to the point $\bar{r}_{P_{p_1,p_2}}(p)$ intersects the plane P_{p_1,p_2} orthogonally at its midpoint. For example, if P is the yz-plane, then

$$\bar{r}_P(x, y, z) = (-x, y, z).$$

Similar to a reflection in a line, a reflection in a plane satisfies the following:

(a)

$$\bar{r}_{P_{p_1,p_2}}(p_1) = p_2, \bar{r}_{P_{p_1,p_2}}(p_2) = p_1.$$

(b)

$$\bar{r}_{P_{p_1,p_2}}(p) = p \text{ for each point } p \in P_{p_1,p_2}.$$

(c)

$$\bar{r}_{P_{p_1,p_2}} \circ \bar{r}_{P_{p_1,p_2}} = \mathrm{id}_{\mathbb{R}^3}, \text{ i.e., } \bar{r}^{-1}_{P_{p_1,p_2}} = \bar{r}_{P_{p_1,p_2}}.$$

If a plane P contains the z-axis, its intersection L with the xy-plane is a line, and the line L goes through the origin. Then, we can consider a reflection $\bar{r}_L : \mathbb{R}^2 \to \mathbb{R}^2$ of the xy-plane in this line. It is easy to verify that

$$\bar{r}_P(x, y, z) = (\bar{r}_L(x, y), z). \tag{2.3}$$

If the points p_1 and p_2 are on the sphere, then the plane P_{p_1,p_2} goes through the origin and $P_{p_1,p_2} \cap \mathbb{S}^2 = G_{p_1,p_2}$, which is a great circle. Since $\bar{r}_{P_{p_1,p_2}}$ fixes the origin, it induces an isometry $\bar{r}_{G_{p_1,p_2}}$ of the sphere by Proposition 2.2.

Remark 2.6. Similar to the reflection $\bar{r}_{P_{p_1,p_2}}$, a reflection in a great circle also satisfies (cf. Remark 1.5):

(a)

$$\bar{r}_{G_{p_1,p_2}}(p_1) = p_2, \bar{r}_{G_{p_1,p_2}}(p_2) = p_1.$$

(b)

$$\bar{r}_{G_{p_1,p_2}}(p) = p \text{ for each point } p \in G_{p_1,p_2}.$$

(c)

$$\bar{r}_{G_{p_1,p_2}} \circ \bar{r}_{G_{p_1,p_2}} = \mathrm{id}_{\mathbb{S}^2}, \text{ i.e., } \bar{r}^{-1}_{G_{p_1,p_2}} = \bar{r}_{G_{p_1,p_2}}.$$

As in the case of the Euclidean plane, every isometry is a composition of reflections. Note that the proof of the following theorem is very similar to that of Theorem 1.6.

Theorem 2.7 (Three reflections theorem for $\mathbb{S}^2$). *An isometry of $\mathbb{S}^2$ is a composition of at most three reflections.*

Proof. Let $\phi : \mathbb{S}^2 \to \mathbb{S}^2$ be an isometry and p_1, p_2, and p_3 be non-collinear points on $\mathbb{S}^2$. We consider four cases.

Case 1. If

$$\phi(p_1) = p_1, \phi(p_2) = p_2 \text{ and } \phi(p_3) = p_3,$$

then $\phi = \mathrm{id}_{\mathbb{S}^2}$ if we let $\psi = \mathrm{id}_{\mathbb{S}^2}$ in Theorem 2.5.

Case 2. If only two of the three points p_1, p_2, and p_3 coincide with their images under ϕ, e.g.,

$$\phi(p_1) = p_1, \phi(p_2) = p_2 \text{ but } \phi(p_3) \neq p_3,$$

then

$$d_{\mathbb{S}^2}(p_3, p_1) = d_{\mathbb{S}^2}(\phi(p_3), \phi(p_1)) = d_{\mathbb{S}^2}(\phi(p_3), p_1).$$

Therefore, if $G = G_{p_3,\phi(p_3)}$, we have $p_1 \in G$. Similarly we have $p_2 \in G$. Let $\psi = \bar{r}_G \circ \phi$; then,

$$\begin{aligned} \psi(p_1) &= \bar{r}_G(\phi(p_1)) \\ &= \bar{r}_G(p_1) \\ &= p_1 \qquad\qquad (\because\ p_1 \in G). \end{aligned}$$

Similarly, $\psi(p_2) = p_2$. Note also that

$$\psi(p_3) = \bar{r}_G(\phi(p_3)) = p_3 \qquad (\because\ G = G_{p_3,\phi(p_3)}).$$

Again, $\psi = \mathrm{id}_{\mathbb{S}^2}$ from Theorem 2.5, i.e., $\bar{r}_G \circ \phi = \mathrm{id}_{\mathbb{S}^2}$. Therefore, $\phi = \bar{r}_G^{-1} = \bar{r}_G$, and ϕ is a reflection.

Case 3. If only one of the three points p_1, p_2, and p_3 coincides with its image under ϕ, e.g.,

$$\phi(p_1) = p_1 \text{ but } \phi(p_2) \neq p_2, \phi(p_3) \neq p_3,$$

then

$$d_{\mathbb{S}^2}(p_2, p_1) = d_{\mathbb{S}^2}(\phi(p_2), \phi(p_1)) = d_{\mathbb{S}^2}(\phi(p_2), p_1).$$

Therefore, letting $M = G_{p_2,\phi(p_2)}$, we have $p_1 \in M$. Let $\phi' = \bar{r}_M \circ \phi$. We then have

$$\phi'(p_1) = \bar{r}_M(\phi(p_1)) = \bar{r}_M(p_1) = p_1$$

and

$$\phi'(p_2) = \bar{r}_M(\phi(p_2)) = p_2,$$

which leads us back to Case 1 or Case 2. Therefore, $\phi' = \mathrm{id}_{\mathbb{S}^2}$ or $\phi' = \bar{r}_G$ for some great circle G, i.e., $\phi = \bar{r}_M$ or $\phi = \bar{r}_M \circ \bar{r}_G$.

Case 4. Finally, assume that

$$\phi(p_1) \neq p_1, \phi(p_2) \neq p_2, \phi(p_3) \neq p_3.$$

If $N = G_{p_1,\phi(p_1)}$ and we let $\phi'' = \bar{r}_N \circ \phi$, then

$$\phi''(p_1) = \bar{r}_N(\phi(p_1)) = p_1.$$

This leads us again back to Case 1, Case 2, or Case 3. Therefore,

$$\phi'' = \mathrm{id}_{\mathbb{S}^2}, \phi'' = \bar{r}_M \text{ or } \phi'' = \bar{r}_M \circ \bar{r}_G$$

for some great circles G and M, i.e.,

$$\phi = \bar{r}_N, \phi = \bar{r}_N \circ \bar{r}_M \text{ or } \phi = \bar{r}_N \circ \bar{r}_M \circ \bar{r}_G.$$

□

Consider the rotation ϕ of $\mathbb{R}^3$ by angle θ about the z-axis,

$$\phi(x, y, z) = (x\cos\theta - y\sin\theta, x\sin\theta + y\cos\theta, z).$$

Note that

$$\phi(x, y, z) = (r_\theta(x, y), z). \tag{2.4}$$

Since this rotation fixes the origin, it induces a rotation $r_{z,\theta}$ on $\mathbb{S}^2$. For each line l through the origin, similarly, one can define a rotation $r_{l,\theta}$ about the line.

Consider a composition of two reflections in the great circles G_1 and G_2. There exist two planes P_1 and P_2 through the origin $\mathbf{0}$ such that $G_1 = P_1 \cap \mathbb{S}^2$ and $G_2 = P_2 \cap \mathbb{S}^2$. Note that the intersection of these two planes is a line through the origin. We set the coordinate system such that this line coincides with the z-axis. Let L_1 and L_2 be the lines on the xy-plane that are cut by P_1 and P_2, respectively, then L_1 and L_2 are lines on the xy-plane that pass through the origin. By (2.3),

$$\bar{r}_{P_1}(x, y, z) = (\bar{r}_{L_1}(x, y), z)$$

$$\bar{r}_{P_2}(x, y, z) = (\bar{r}_{L_2}(x, y), z)$$

$$\bar{r}_{P_2}\left(\bar{r}_{P_1}(x, y, z)\right) = (\bar{r}_{L_2}\left(\bar{r}_{L_1}(x, y)\right), z)$$

for each point $(x, y, z) \in \mathbb{R}^3$. Note that $\bar{r}_{L_2} \circ \bar{r}_{L_1}$ is a rotation of the xy-plane about the origin, i.e.,

$$\bar{r}_{L_2} \circ \bar{r}_{L_1} = r_\theta$$

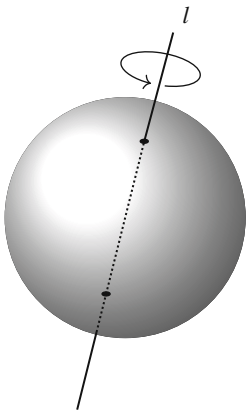


Fig. 2.4 Rotation $r_{l,\theta}$ of $\mathbb{S}^2$ by an angle θ about a line l

for some angle θ.

If we restrict this rotation to the xy-plane, by (2.4), we have

$$\left(\bar{r}_{G_2} \circ \bar{r}_{G_1}\right)(x, y, z) = (r_\theta(x, y), z) = r_{z,\theta}(x, y, z)$$

for each point $(x, y, z) \in \mathbb{S}^2$. In summary, a composition of two reflections in a great circle is a rotation. By $r_{l,\theta}$, we mean a rotation by angle θ about a line l through the origin (Figure 2.4).

Let us show that a rotation can be also expressed as a composition of two reflections in great circles. Consider a rotation $r_{l,\theta}$. We choose a coordinate system such that the line l coincides with the z-axis. Therefore,

$$r_{l,\theta}(x, y, z) = r_{z,\theta}(x, y, z) = (r_\theta(x, y), z).$$

Recall that a rotation of the Euclidean plane is a composition of two reflections in lines:

$$r_\theta = \bar{r}_{L_2} \circ \bar{r}_{L_1},$$

where L_1 and L_2 are lines on the xy-plane that pass through the origin. There exist planes P_1 and P_2 that contain the lines L_1 and L_2, respectively, and the z-axis. Then, by (2.3),

$$\begin{aligned} r_{l,\theta}(x, y, z) &= r_{z,\theta}(x, y, z) = (r_\theta(x, y), z) = (\bar{r}_{L_2}(\bar{r}_{L_1}(x, y)), z) \\ &= \bar{r}_{P_2}(\bar{r}_{L_1}(x, y), z) = \bar{r}_{P_2}(\bar{r}_{P_1}(x, y, z)) = (\bar{r}_{P_2} \circ \bar{r}_{P_1})(x, y, z) \\ &= (\bar{r}_{G_2} \circ \bar{r}_{G_1})(x, y, z), \end{aligned}$$

where $G_1 = P_1 \cap \mathbb{S}^2$ and $G_2 = P_2 \cap \mathbb{S}^2$. Therefore,

$$r_{l,\theta} = \bar{r}_{G_2} \circ \bar{r}_{G_1}.$$

Furthermore, for a given great circle G_1' that intersects with line l, it is not difficult to see that there exists a great circle G_2' such that

$$r_{l,\theta} = \bar{r}_{G_2'} \circ \bar{r}_{G_1'}.$$

Theorem 2.8. *A composition of two rotations is a rotation.*

Proof. Consider two rotations $r_{l_1,\theta_1}, r_{l_2,\theta_2}$ and their composition

$$\phi = r_{l_2,\theta_2} \circ r_{l_1,\theta_2}.$$

If $l_1 = l_2$, then $\phi = r_{l_1,\theta_1+\theta_2}$, which is a rotation. Assume that $l_1 \neq l_2$. Let P be the plane that contains both the lines l_1, l_2 and $G = P \cap \mathbb{S}^2$. Note that G is a great circle which meets with l_1, l_2 at two points, respectively. Thus there are great circles G_1, G_2 such that

$$r_{l_1,\theta_1} = \bar{r}_G \circ \bar{r}_{G_1}$$

and

$$r_{l_2,\theta_2} = \bar{r}_{G_2} \circ \bar{r}_G.$$

Then

$$\phi = r_{l_2,\theta_2} \circ r_{l_1,\theta_2} = \bar{r}_{G_2} \circ \bar{r}_G \circ \bar{r}_G \circ \bar{r}_{G_1} = \bar{r}_{G_2} \circ \bar{r}_{G_1},$$

which is a rotation. □

Let $\mathrm{Iso}^+(\mathbb{S}^2)$ be the set of all compositions of even numbers of reflections of the sphere and $\mathrm{Iso}^-(\mathbb{S}^2)$ be the set of all compositions of odd numbers of reflections. An isometry in $\mathrm{Iso}^+(\mathbb{S}^2)$ is said to be *orientation-preserving*, and an isometry in $\mathrm{Iso}^-(\mathbb{S}^2)$ is said to be *orientation-reversing*. By Theorem 2.8, an isometry is orientation-preserving if and only if it is a rotation. This theorem is known as Euler's rotation theorem.

Lemma 2.9. *Let $\bar{r}_1, \bar{r}_2, \cdots, \bar{r}_n$ be reflections in great circles. If the composition $\bar{r}_1 \circ \bar{r}_2 \circ \cdots \circ \bar{r}_n$ is the identity map, then n is even.*

Proof. Suppose that n is odd; then, $n = 2k + 1$ for some k. Note that

$$\bar{r}_1 \circ \bar{r}_2 \circ \cdots \circ \bar{r}_n = r_1 \circ r_2 \circ \cdots \circ r_k \circ \bar{r}_n,$$

where $r_i = \bar{r}_{2i-1} \circ \bar{r}_{2i}$, which is a rotation. Applying Theorem 2.8, we have $r_1 \circ r_2 \circ \cdots \circ r_k \circ \bar{r}_n = r \circ \bar{r}_n$, where r is a rotation. Hence,

$$r \circ \bar{r}_n = \mathrm{id}_{\mathbb{S}^2},$$

and finally, $r = \bar{r}_n$. Note that the set of fixed points of $\bar{r}_n$ is a great circle. However, the set of fixed points of a rotation r is composed of two points or the set $\mathbb{S}^2$. Therefore, we have a contradiction, and n is even. □

Theorem 2.10. *The sets* $\mathrm{Iso}^+(\mathbb{S}^2)$ *and* $\mathrm{Iso}^-(\mathbb{S}^2)$ *are disjoint.*

Proof. Suppose that $\mathrm{Iso}^+(\mathbb{S}^2) \cap \mathrm{Iso}^-(\mathbb{S}^2)$ is not empty; then, it contains an isometry ϕ. Since $\phi \in \mathrm{Iso}^+(\mathbb{S}^2)$,

$$\phi = \bar{r}_1 \circ \bar{r}_2 \circ \cdots \circ \bar{r}_n,$$

where each $\bar{r}_i$ is a reflection in a great circle and n is even. Moreover, because $\phi \in \mathrm{Iso}^-(\mathbb{S}^2)$,

$$\phi = \bar{r}_{n+1} \circ \bar{r}_{n+2} \circ \cdots \circ \bar{r}_{n+m},$$

where m is odd. Hence,

$$\bar{r}_1 \circ \bar{r}_2 \circ \cdots \circ \bar{r}_n = \bar{r}_{n+1} \circ \bar{r}_{n+2} \circ \cdots \circ \bar{r}_{n+m}$$

and

$$\bar{r}_n \circ \bar{r}_{n-1} \circ \cdots \circ \bar{r}_1 \circ \bar{r}_{n+1} \circ \bar{r}_{n+2} \circ \cdots \circ \bar{r}_{n+m} = \mathrm{id}_{\mathbb{S}^2}.$$

By Lemma 2.9, $n + m$ is even, which is contradictory to the fact that n is even and m is odd, thus proving the theorem. □

Exercises

2.4. Recall that the conjugation of ψ by ϕ is the isometry

$$^{\phi}\psi = \phi \circ \psi \circ \phi^{-1}$$

for the given isometries ϕ and ψ.

(a) For reflections $\bar{r}_G$ and $\bar{r}_H$, show that

$$^{\bar{r}_H}\bar{r}_G = \bar{r}_{G'},$$

where $G' = \bar{r}_H(G)$.

(b) For an isometry ϕ, show that

$$^{\phi}\bar{r}_G = \bar{r}_{G'},$$

where $G' = \phi(G)$.

2.5. The antipodal map $\hat{a} : \mathbb{S}^2 \to \mathbb{S}^2$, defined by $p \mapsto -p$, is an isometry.

(a) Show that the antipodal map is orientation-reversing.

(b) Let ϕ be an isometry of the sphere. Show that it is orientation-reversing if and only if $\phi = \hat{a} \circ r_{l,\theta}$ for some rotation $r_{l,\theta}$.

2.6. Let ϕ be an isometry of the sphere. Show the following:

(a) It has at least two fixed points if it is not fixed-point-free.
(b) ϕ is a rotation if it has exactly two fixed points.

2.7. Suppose that an isometry ϕ of the sphere is fixed-point-free and $\phi^2 = \mathrm{id}_{\mathbb{S}^2}$. Prove that $\phi = \hat{a}$, the antipodal map.

2.8. Classify all the isometries of the sphere such that $\phi^2 = \mathrm{id}_{\mathbb{S}^2}$.

2.9. Show that $\hat{a} \circ \phi = \phi \circ \hat{a}$ for every isometry ϕ of the sphere.

2.10. Show that there is no isometry ϕ of the sphere with the property $\phi \circ \phi = \hat{a}$.

2.11. Complete the following multiplication table of isometries of $\mathbb{S}^2$.
Legend: Re = reflection, Ro = Rotation, N = neither reflection nor rotation.

	Re	Ro	N
Re	(i)	(ii)	(iii)
Ro	(iv)	Ro	(vii)
N	(v)	(vi)	Ro

(i) = Re $\circ$ Re, (ii) = Re $\circ$ Ro, and so on. For example, Re and N are possible solutions for (ii).

2.12. Prove that there exists an isometry of $\mathbb{S}^2$ that cannot be expressed as a composition of one or two reflections.

2.3 Area of a Spherical Triangle

In the Euclidean plane, Heron's formula gives the area of a triangle when the length of its sides are known:

$$\text{Area} = \sqrt{s(s-a)(s-b)(s-c)},$$

where a, b, and c are the lengths of the sides, and $s = \frac{1}{2}(a+b+c)$ is the semiperimeter of the triangle. A region on the sphere bounded by three distinct great circles is called a *spherical triangle* (Figure 2.6). There is a remarkably simple

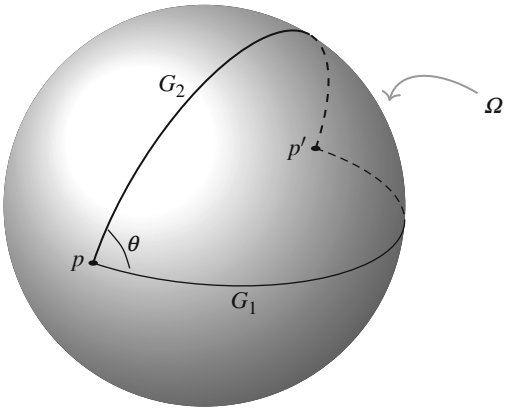


Fig. 2.5 Spherical lune of interior angle θ

formula for the area of a spherical triangle. It is expressed solely by the interior angles.

Let G_1 and G_2 be half great circles from a point p to its antipodal point p' and θ be the interior angle between G_1 and G_2 (Figure 2.5). Then,

$$\text{Area}(\mathbb{S}^2) : \text{Area}(\Omega) = 2\pi : \theta,$$

where Ω is a spherical lune, i.e., the region bounded by the circles G_1 and G_2, as in Figure 2.5. Recall that

$$\text{Area}(\mathbb{S}^2) = 4\pi.$$

Hence,

$$\text{Area}(\Omega) = 4\pi \cdot \frac{\theta}{2\pi} = 2\theta. \tag{2.5}$$

Theorem 2.11. *The area of a spherical triangle with interior angles α, β, and γ ($\triangle pqr$ in Figure 2.6) is*

$$\alpha + \beta + \gamma - \pi.$$

Proof. Let p', q', and r' be the antipodal points of p, q, and r, respectively, as shown in Figure 2.6, and

$$\begin{aligned} A &= \text{Area}(\triangle pqr), & A' &= \text{Area}(\triangle p'q'r'),\\ P &= \text{Area}(\triangle p'qr), & P' &= \text{Area}(\triangle pq'r'),\\ Q &= \text{Area}(\triangle pq'r), & Q' &= \text{Area}(\triangle p'qr'),\\ R &= \text{Area}(\triangle pqr'), & R' &= \text{Area}(\triangle p'q'r). \end{aligned}$$

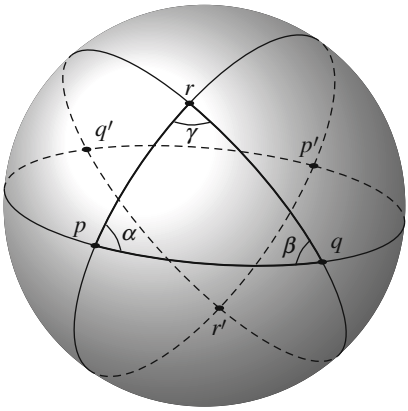


Fig. 2.6 Spherical triangle with interior angles α, β, and γ

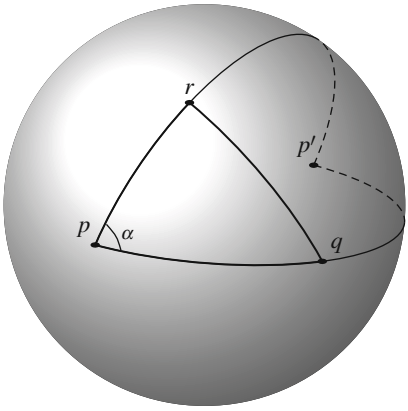


Fig. 2.7 A spherical triangle in a spherical lune

Since these 8 spherical triangles completely cover the sphere,

$$A + P + Q + R + A' + P' + Q' + R' = \text{Area}(\mathbb{S}^2) = 4\pi.$$

By (2.5), $A + P = 2\alpha$ (see Figure 2.7 and compare with Figure 2.5). Similarly,

$$\begin{aligned} A + Q &= 2\beta, \\ A + R &= 2\gamma, \\ A' + P' &= 2\alpha, \\ A' + Q' &= 2\beta, \\ A' + R' &= 2\gamma. \end{aligned}$$

Adding all six equations,

$$2(A + A') + A + P + Q + R + A' + P' + Q' + R' = 4(\alpha + \beta + \gamma).$$

Note that the two spherical triangles $\triangle pqr$ and $\triangle p'q'r'$ are antipodal; therefore, $A = A'$. Finally,

$$2(A + A) + 4\pi = 4(\alpha + \beta + \gamma),$$

and so

$$A = \alpha + \beta + \gamma - \pi.$$

□

Since the area of a spherical triangle is positive, the sum of its interior angles is *greater* than π. We can calculate the area of a spherical polygon by dividing it into several spherical triangles. For example, consider a spherical pentagon whose five vertices are $p_1, \cdots, p_5$ with interior angles $\theta_1, \cdots, \theta_5$ (Figure 2.8). The area of this pentagon is then

$$\text{Area}(\triangle p_1 p_4 p_5) + \text{Area}(\triangle p_1 p_3 p_4) + \text{Area}(\triangle p_1 p_2 p_3).$$

$$\begin{aligned}
\text{Area}(\triangle p_1 p_4 p_5) &= \angle p_5 p_1 p_4 + \angle p_1 p_4 p_5 + \angle p_4 p_5 p_1 - \pi, \\
\text{Area}(\triangle p_1 p_3 p_4) &= \angle p_4 p_1 p_3 + \angle p_1 p_3 p_4 + \angle p_3 p_4 p_1 - \pi, \\
\text{Area}(\triangle p_1 p_2 p_3) &= \angle p_3 p_1 p_2 + \angle p_1 p_2 p_3 + \angle p_2 p_3 p_1 - \pi.
\end{aligned}$$

Note that $\theta_1 = \angle p_5 p_1 p_4 + \angle p_4 p_1 p_3 + \angle p_3 p_1 p_2$, etc. Adding all three equations, the area of the pentagon is as follows:

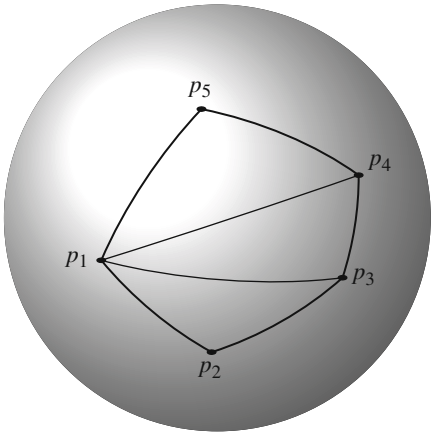


Fig. 2.8 Spherical pentagon and its subdivision into spherical triangles

$$\theta_1 + \theta_2 + \theta_3 + \theta_4 + \theta_5 - 3\pi.$$

Generalizing this argument, one can similarly prove the following corollary.

Corollary 2.12. *The area of an n-gon on* $\mathbb{S}^2$ *with interior angles* $\theta_1, \cdots, \theta_n$ *is*

$$\theta_1 + \cdots + \theta_n - (n-2)\pi. \tag{2.6}$$

Theorem 2.13 (Euler's Theorem). *Let* v *, e, and* f *denote the number of vertices, edges, and faces, respectively, of a convex polyhedron. Then,*

$$v - e + f = 2.$$

Proof. Let us place the polyhedron in $\mathbb{R}^3$ so that the origin $\mathbf{0}$ coincides with the center of the polyhedron. Project the faces of the polyhedron from the origin $\mathbf{0}$ onto the sphere via the map

$$p \mapsto \frac{p}{\|p\|}.$$

Now the sphere is covered by f spherical polygons $P_1, \cdots, P_f$ that correspond to the faces of the polyhedron (see Figure 2.9 for the case of a cube). Therefore, the sum of the areas of these polygons is 4π, i.e.,

$$\text{Area}(P_1) + \cdots + \text{Area}(P_f) = 4\pi.$$

Let n_i be the number of edges of P_i and α_{ij}, for $j = 1, \cdots, n_i$, be its interior angles. By Corollary 2.12,

$$\text{Area}(P_i) = \sum_{j=1}^{n_i} \alpha_{ij} - (n_i - 2)\pi,$$

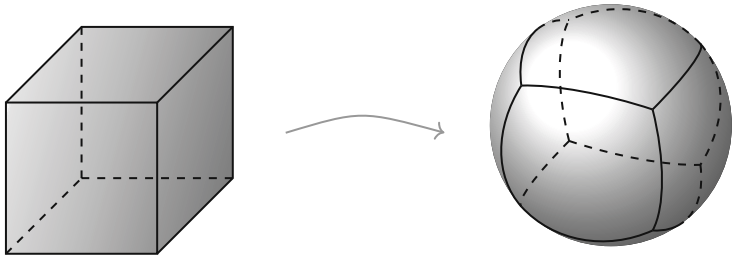


Fig. 2.9 Projection of the faces of a cube from the origin onto the sphere

and thus,

$$4\pi = \sum_{i=1}^{f} \text{Area}(P_i) = \sum_{i=1}^{f} \left(\sum_{j=1}^{n_i} \alpha_{ij} - (n_i - 2)\pi \right).$$

Since every edge is shared by two polygons,

$$\sum_{i=1}^{f} n_i = 2e.$$

Since the sum of the angles at every vertex is 2π,

$$\sum_{i=1}^{f} \left(\sum_{j=1}^{n_i} \alpha_{ij} \right) = 2\pi v.$$

Hence,

$$\begin{aligned} 4\pi &= \sum_{i=1}^{f} \text{Area}(P_i) \\ &= \sum_{i=1}^{f} \left(\sum_{j=1}^{n_i} \alpha_{ij} - (n_i - 2)\pi \right) \\ &= \sum_{i=1}^{f} \left(\sum_{j=1}^{n_i} \alpha_{ij} \right) - \pi \sum_{i=1}^{f} n_i + \sum_{i=1}^{f} 2\pi \\ &= 2\pi v - \pi \cdot 2e + 2\pi f, \end{aligned}$$

and so

$$v - e + f = 2.$$

□

A Platonic solid is a polyhedron whose vertices all have the same degree and whose faces are all congruent to the same regular polygon.

Theorem 2.14. *There are exactly* five *Platonic solids: tetrahedrons, octahedrons, icosahedrons, cubes, and dodecahedrons (see Figure 2.10).*

Proof. Let P be a Platonic solid for which the degree of each of vertex is a, and let each of its faces be a regular polygon with b sides. Then, $2e = av$, and $2e = bf$. Note that $a, b \geq 3$.

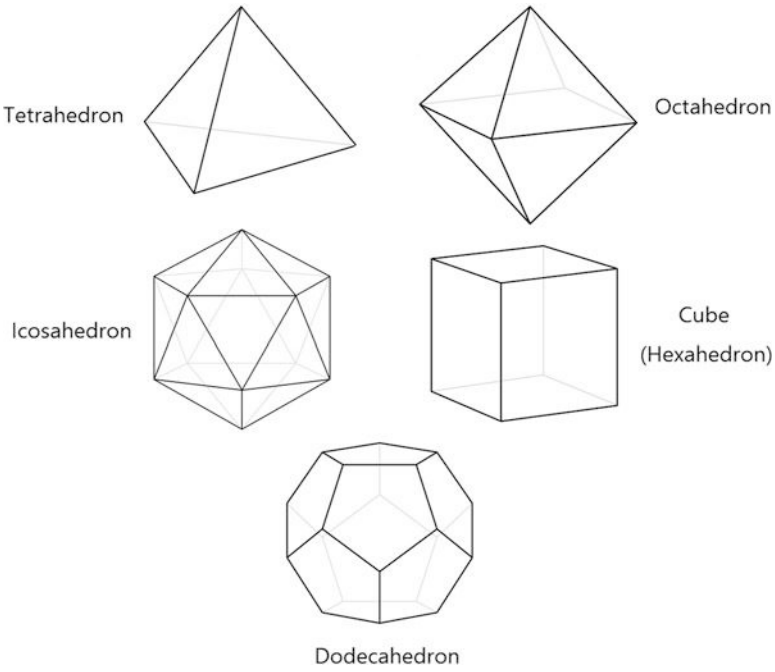


Fig. 2.10 Platonic solids

By Euler's Theorem, $v - e + f = 2$; hence,

$$\frac{2e}{a} - e + \frac{2e}{b} = 2.$$

Therefore,

$$\frac{1}{a} + \frac{1}{b} = \frac{1}{2} + \frac{1}{e} > \frac{1}{2}.$$

If $a \geq 6$ or $b \geq 6$, then

$$\frac{1}{a} + \frac{1}{b} \leq \frac{1}{3} + \frac{1}{6} = \frac{1}{2},$$

which is a contradiction. Hence, $a < 6$ and $b < 6$, which gives us a finite number of cases to check, as shown in the following table.

□

a	b	e	v	f	Solid
3	3	6	4	4	Tetrahedron
3	4	12	8	6	Cube
3	5	30	20	12	Dodecahedron
4	3	12	6	8	Octahedron
4	4				X
4	5				X
5	3	30	12	20	Icosahedron
5	4				X
5	5				X

Exercises

2.13. Suppose that an equilateral spherical triangle has sides of length a and interior angles θ. Show

$$\cos\frac{a}{2}\sin\frac{\theta}{2}=\frac{1}{2}.$$

From this, show that $\theta > \frac{\pi}{3}$ directly (so that the sum of the interior angles of an equilateral spherical triangle exceeds π).

2.14. Let T be a right-angled isosceles triangle on $\mathbb{S}^2$ with like sides of spherical length a (i.e., one angle is $\frac{\pi}{2}$). Show that

$$\frac{\text{Area}(T)}{\frac{1}{2}a^2}$$

converges to one as a approaches zero.

2.15. Show that it is not possible to draw a "spherical rectangle," i.e., a quadrilateral with 4 right angles.

2.4 Orthogonal Transformations of Euclidean Spaces

Consider $\mathbb{R}^3$, the Euclidean space of dimension three. For two points $p_1 = (x_1, y_1, z_1)$, $p_2 = (x_2, y_2, z_2) \in \mathbb{R}^3$, recall that the distance between p_1 and p_2 is given by

$$d(p_1, p_2) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2}.$$

Recall that a bijective map $\phi : \mathbb{R}^3 \to \mathbb{R}^3$ is an isometry of $\mathbb{R}^3$ if it preserves the distance, i.e.,

$$d(\phi(p_1), \phi(p_2)) = d(p_1, p_2)$$

for any two points $p_1, p_2 \in \mathbb{R}^3$.

We have seen that isometries of $\mathbb{S}^2$ have a close relation with some particular types of isometries of $\mathbb{R}^3$.

Definition 2.15. An isometry ϕ of $\mathbb{R}^3$ is called an *orthogonal transformation* of $\mathbb{R}^3$ if $\phi(\mathbf{0}) = \mathbf{0}$.

The set of all orthogonal transformations of $\mathbb{R}^3$ is denoted $\mathrm{O}(3)$.

Proposition 2.16. *Suppose that ϕ is an orthogonal transformation of $\mathbb{R}^3$. Then, we have $\phi(\mathbb{S}^2) = \mathbb{S}^2$.*

Proof. Let $p \in \mathbb{S}^2$. Since $\|p\| = 1$,

$$\|\phi(p)\| = d_{\mathbb{R}^3}(\phi(p), \mathbf{0}) = d_{\mathbb{R}^3}(\phi(p), \phi(\mathbf{0})) = d_{\mathbb{R}^3}(p, \mathbf{0}) = \|p\| = 1.$$

Thus, $\phi(\mathbb{S}^2) \subset \mathbb{S}^2$.

Conversely, since ϕ is surjective, there exists some $q \in \mathbb{R}^3$ such that $\phi(q) = p$. Note that

$$\|q\| = d_{\mathbb{R}^3}(q, \mathbf{0}) = d_{\mathbb{R}^3}(\phi(q), \phi(\mathbf{0})) = d_{\mathbb{R}^3}(p, \mathbf{0}) = \|p\| = 1.$$

Thus, $q \in \mathbb{S}^2$, and $\phi(q) = p$. We then have $\mathbb{S}^2 \subset \phi(\mathbb{S}^2)$. □

Hence, the restriction of ϕ to $\mathbb{S}^2$ is a surjective map

$$\phi|_{\mathbb{S}^2} : \mathbb{S}^2 \to \mathbb{S}^2.$$

Proposition 2.17. *Suppose that ϕ is an orthogonal transformation of $\mathbb{R}^3$. The restriction $\phi|_{\mathbb{S}^2} : \mathbb{S}^2 \to \mathbb{S}^2$ of ϕ is an isometry of $\mathbb{S}^2$.*

Proof. We need to show that $\phi|_{\mathbb{S}^2}$ preserves the spherical distance. For $p_1, p_2 \in \mathbb{S}^2$,

$$\begin{aligned} d_{\mathbb{S}^2}(\phi(p_1), \phi(p_2)) &= 2 \arcsin\left(\frac{1}{2} d(\phi(p_1), \phi(p_2))\right) \\ &= 2 \arcsin\left(\frac{1}{2} d(p_1, p_2)\right) = d_{\mathbb{S}^2}(p_1, p_2). \end{aligned}$$

□

Now we have a function

$$\Psi : \mathrm{O}(3) \to \mathrm{Iso}(\mathbb{S}^2)$$

defined by $\phi \mapsto \phi|_{\mathbb{S}^2}$.

Theorem 2.18. *The map $\Psi : \mathrm{O}(3) \to \mathrm{Iso}(\mathbb{S}^2)$ is bijective.*

1. *A reflection in a hyperplane through the origin maps to a reflection in a great circle, and*
2. *$\phi_1 \circ \phi_2 \mapsto \phi_1|_{\mathbb{S}^2} \circ \phi_2|_{\mathbb{S}^2}$.*

Proof. For each $\phi \in \mathrm{Iso}(\mathbb{S}^2)$, define $\tilde{\phi} : \mathbb{R}^3 \to \mathbb{R}^3$ by

$$\tilde{\phi}(p) = \|p\| \cdot \phi\left(\frac{p}{\|p\|}\right)$$

for $p \neq \mathbf{0}$ and $\tilde{\phi}(\mathbf{0}) = \mathbf{0}$. It can be verified that $\tilde{\phi} \in \mathrm{O}(3)$. Thus, we have a function

$$\Phi : \mathrm{Iso}(\mathbb{S}^2) \to \mathrm{O}(3).$$

It is also straightforward to verify that

$$\Phi \circ \Psi = \mathrm{id}_{\mathrm{O}(3)}, \Psi \circ \Phi = \mathrm{id}_{\mathrm{Iso}(\mathbb{S}^2)};$$

hence, Ψ is bijective.

Properties 1 and 2 are obvious. □

With this theorem and Theorem 2.7, we have the following two corollaries:

Corollary 2.19. *Every orthogonal transformation is a composition of at most three reflections in a hyperplane through the origin.*

Corollary 2.20. *Each isometry of* $\mathbb{R}^3$ *is a composition of at most four reflections in hyperplanes.*

Proof. Let $\phi : \mathbb{R}^3 \to \mathbb{R}^3$ be an isometry. Note that there exists some reflection $\bar{r}$ in a hyperplane of $\mathbb{R}^3$ such that $\bar{r}(\mathbf{0}) = \phi(\mathbf{0})$. Let $\psi = \bar{r} \circ \phi$. Then, ψ is an orthogonal transformation and can thus be expressed as a composition of at most three reflections. Since $\phi = \bar{r} \circ \psi$, the proof is complete. □

Let

$$\mathbb{S}^{n-1} = \{p \in \mathbb{R}^n \mid \|p\| = 1\}$$

and $\mathrm{Iso}(\mathbb{S}^{n-1})$ be the set of all bijective maps from $\mathbb{S}^{n-1}$ to itself that preserve distance. Further, define $\mathrm{O}(n)$ as the set of all isometries of $\mathbb{R}^n$ that map the origin to the origin.

Using the same arguments, we can show that there is a one-to-one correspondence between $\mathrm{O}(n)$ and $\mathrm{Iso}(\mathbb{S}^{n-1})$ in general. For higher dimensions ($n > 3$), we do not have good geometric intuition; thus, it is not easy to investigate $\mathrm{Iso}(\mathbb{S}^{n-1})$ geometrically, as we have done for $\mathbb{S}^2$. However, we can investigate $\mathrm{O}(n)$ using *linear algebra*, for which geometric intuition is not indispensable. Since there is a one-to-one correspondence between $\mathrm{O}(n)$ and $\mathrm{Iso}(\mathbb{S}^{n-1})$, we can see the structure of $\mathrm{Iso}(\mathbb{S}^{n-1})$ by looking at $\mathrm{O}(n)$. This is one of the reasons why we need to study various fields of mathematics.

Chapter 3
Stereographic Projection and Inversions

"Everything has beauty, but not everyone sees it."

Confucius (551–479 BC)

"Inspiration is needed in geometry, just as much as in poetry."

Alexander Pushkin (1799–1837)

It is impossible to map figures on the sphere onto *congruent* ones on the plane. Because the Earth is spherical, any flat representation of it causes distortions such that areas and shapes cannot both be conserved simultaneously, i.e., the distance cannot be preserved. The mapmaker must choose a projection method suitable for the region to be mapped and the purpose of the map. Stereographic projection is one method of making maps that preserves angles.

3.1 Stereographic Projection

The point $N = (0, 0, 1)$ on the sphere is called the north pole. Let $\mathbb{S}^{2^*} = \mathbb{S}^2 - \{N\}$, and define a map $\Phi : \mathbb{S}^{2^*} \to \mathbb{R}^2$ as follows.

Definition 3.1. For $p \in \mathbb{S}^{2^*}$, there exists a unique point $q = (u, v)$ on $\mathbb{R}^2$ such that the line from N to $(u, v, 0)$ passes through p (Figure 3.1). The map

$$\Phi : \mathbb{S}^{2^*} \to \mathbb{R}^2,$$

defined by $\Phi(p) = q$, is called the *stereographic projection*.

N.-H. Lee, *Geometry: from Isometries to Special Relativity*, Undergraduate Texts in Mathematics, https://doi.org/10.1007/978-3-030-42101-4_3

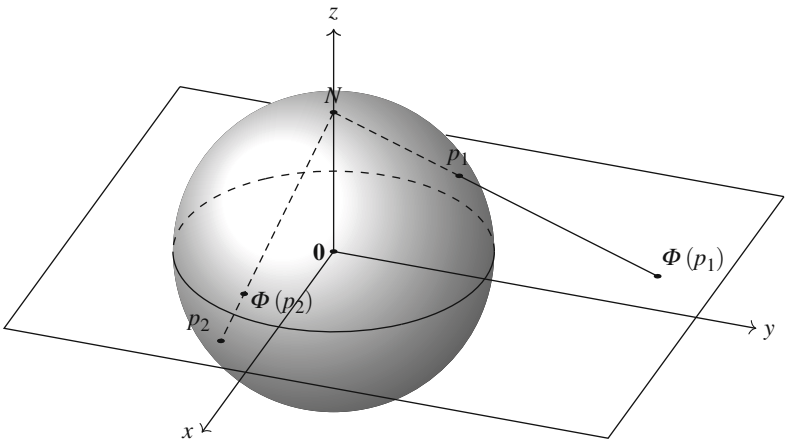


Fig. 3.1 Stereographic projection Φ

Proposition 3.2. *We have*

$$\Phi(x, y, z) = \left(\frac{x}{1-z}, \frac{y}{1-z}\right)$$

for $(x, y, z) \in \mathbb{S}^{2*}$, *and*

$$\Phi^{-1}(u, v) = \left(\frac{2u}{u^2+v^2+1}, \frac{2v}{u^2+v^2+1}, \frac{u^2+v^2-1}{u^2+v^2+1}\right)$$

for $(u, v) \in \mathbb{R}^2$.

Proof. Let $(\alpha, \beta) = \Phi(x, y, z)$. The line passing through $(\alpha, \beta, 0)$ and (x, y, z) is

$$\{t(x, y, z) + (1-t)(\alpha, \beta, 0) \mid t \in \mathbb{R}\}.$$

Hence,

$$(0, 0, 1) = N = t(x, y, z) + (1-t)(\alpha, \beta, 0)$$

for some $t \in \mathbb{R}$. Thus, $t = \frac{1}{z}$, and

$$\alpha = \frac{x}{1-z}, \qquad \beta = \frac{y}{1-z}.$$

It is clear that Φ is a bijection. It is then trivial to check that

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

Therefore,

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

□

In many cases, it is convenient to define Φ on the whole sphere. Hence, we add a point (denoted by "∞") at infinity to $\mathbb{R}^2$. The point ∞ is considered to be near to points with very large norms, just as the origin $\mathbf{0}$ is near to points with very small norms.

Definition 3.3. The set

$$\mathbb{R}^2_\infty := \mathbb{R}^2 \cup \{\infty\}$$

is called the *extended plane*.

Now we can define the stereographic projection on the whole sphere by setting

$$\Phi(N) = \infty.$$

Then, the stereographic projection becomes a bijective map.

Theorem 3.4. *The stereographic projection maps circles of the unit sphere that contain the north pole to Euclidean straight lines in the plane, and it maps circles of the unit sphere that do not contain the north pole to circles in the plane (Figures 3.2 and 3.3).*

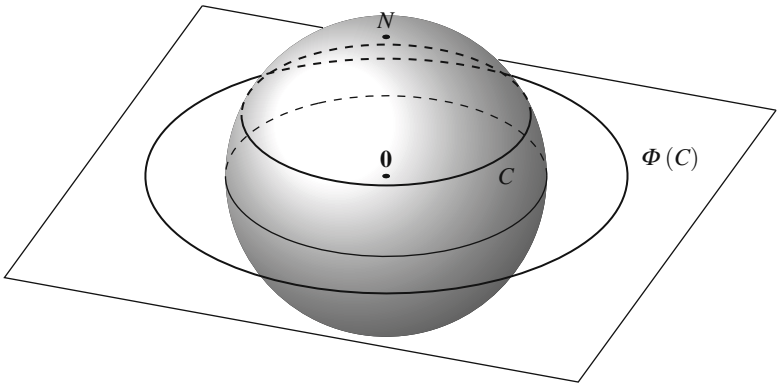


Fig. 3.2 The image $\Phi(C)$ of a circle C on $\mathbb{S}^2$, centered at N

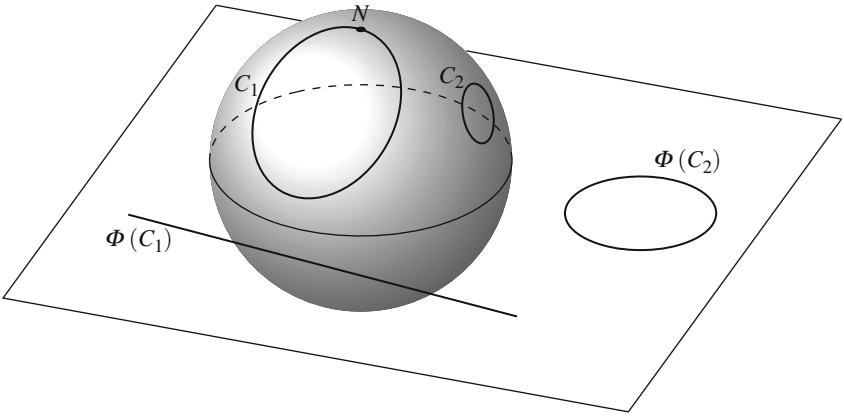


Fig. 3.3 The images of circles on $\mathbb{S}^2$ obtained by Φ are circles or lines

Conversely, the inverse images of circles or lines on $\mathbb{R}^2$ *obtained by the stereographic projection are circles on* $\mathbb{S}^2$.

Proof. Let C be a circle on the unit sphere. The circle C is the set of all points (x, y, z) on $\mathbb{S}^2$ that lie on some slicing plane E. The plane E is defined by

$$ax + by + cz + d = 0,$$

where the real numbers a, b, c, and d satisfy $a^2 + b^2 + c^2 \neq 0$. Then, the projection points $(u, v) \in \mathbb{R}^2$ of the circle C satisfy

$$a\left(\frac{2u}{1+u^2+v^2}\right) + b\left(\frac{2v}{1+u^2+v^2}\right) + c\left(\frac{u^2+v^2-1}{1+u^2+v^2}\right) + d = 0,$$

i.e.,

$$2au + 2bv + (u^2 + v^2)(c + d) = c - d.$$

If the circle C contains the north pole $N = (0, 0, 1)$, then we have $c + d = 0$. Thus, in this case, the equation indicates that the projection of C is a Euclidean straight line. If, however, the circle C does not contain the north pole, then $c + d \neq 0$. In this case, we obtain

$$\left(u + \frac{a}{c+d}\right)^2 + \left(v + \frac{b}{c+d}\right)^2 = \frac{a^2+b^2+c^2-d^2}{(c+d)^2}.$$

Note that $\frac{a^2+b^2+c^2-d^2}{(c+d)^2} > 0$ (why?). Hence, we have the equation for a Euclidean circle in the plane.

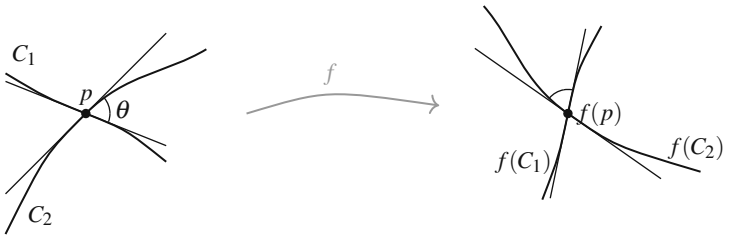


Fig. 3.4 Angle preservation at point p

Conversely, let Γ be either a circle or a line on $\mathbb{R}^2$. Choose three distinct points q_1, q_2, and q_3 from Γ, and let p_1, p_2, and p_3 be the inverse images of these points on $\mathbb{S}^2$. Then, there exists a unique circle C on $\mathbb{S}^2$ that passes through p_1, p_2, and p_3, and $\Phi(C)$ is a circle or a line, as shown already. Let us denote it by Γ'. Note that Γ' also contains q_1, q_2, and q_3. Hence, $\Gamma = \Gamma'$, which completes the proof. □

Given two smooth curves C_1 and C_2 that intersect at a point p, we mean by the angle between C_1 and C_2 at p the angle between their tangent lines at p, denoted by

$$\angle_p(C_1, C_2)$$

(Figure 3.4). Consider a map f. Then, the curves $f(C_1)$ and $f(C_2)$ intersect at point $f(p)$. If f satisfies

$$\angle_p(C_1, C_2) = \angle_{f(p)}(f(C_1), f(C_2))$$

for any smooth curves C_1 and C_2 through p, f is said to *preserve angles* at p. If f preserves angles at every point of its domain, it is simply said to preserve angles.

Example 3.1. Consider a map $\phi : \mathbb{R}^2 \to \mathbb{R}^2$, defined by $\phi(x, y) = (x, y + y^2)$, and curves

$$\gamma_1(t) = (t \cos\theta_1, t \sin\theta_1),$$

$$\gamma_2(t) = (t \cos\theta_2, t \sin\theta_2),$$

where θ_1 and θ_2 are some fixed angles. Then, γ_1 and γ_2 are lines that pass through the origin, with

$$\angle_{\mathbf{0}}(\gamma_1, \gamma_2) = \theta_2 - \theta_1.$$

Note that

$$\phi(\gamma_i(t)) = (t \cos\theta_i, t \sin\theta_i + t^2 \sin^2\theta_i) \quad (i = 1, 2)$$

which also passes through the origin. Then,

$$\frac{d}{dt}\phi(\gamma_i(t))\bigg|_{t=0} = (\cos\theta_i, \sin\theta_i)$$

is a tangent vector of it at the origin. Hence,

$$\angle_{\mathbf{0}}(\phi(\gamma_1), \phi(\gamma_2)) = \theta_2 - \theta_1.$$

Therefore, ϕ preserves angles at the origin.

Consider another point $p = (1, 1)$. Note that $\phi(p) = (1, 2)$. The lines

$$\gamma_1(t) = (t\cos\theta_1 + 1, t\sin\theta_1 + 1),$$

$$\gamma_2(t) = (t\cos\theta_2 + 1, t\sin\theta_2 + 1)$$

pass through the point p, with

$$\angle_p(\gamma_1, \gamma_2) = \theta_2 - \theta_1.$$

Note that

$$\phi(\gamma_i(t)) = (t\cos\theta_i + 1, t^2\sin^2\theta_i + 3t\sin\theta_i + 2).$$

Similarly,

$$\frac{d}{dt}\phi(\gamma_i(t))\bigg|_{t=0} = (\cos\theta_i, 3\sin\theta_i)$$

is a tangent vector of $\phi(\gamma_i)$ at point $\phi(p)$. Consider some specific angles, for example, $\theta_1 = 0$ and $\theta_2 = \frac{\pi}{4}$. Then, the tangent vectors for $\phi(\gamma_1)$ and $\phi(\gamma_2)$ are

$$(1, 0) and \left(\frac{1}{\sqrt{2}}, \frac{3}{\sqrt{2}}\right),$$

respectively. Hence, the angle between $\phi(\gamma_1)$ and $\phi(\gamma_2)$ is

$$\cos^{-1}\left(\frac{1}{\sqrt{10}}\right) \neq \frac{\pi}{4}.$$

Therefore, ϕ does not preserve angles at p.

Example 3.2. Consider a map $\phi : \mathbb{R}^2 \to \mathbb{R}^2$, defined by $\phi(x, y) = (x - y, x + y)$. We will show that ϕ preserves angles. For a given point $p = (a, b) \in \mathbb{R}^2$ and an angle θ, consider two curves $\gamma_i : [-1, 1] \to \mathbb{R}^2$, with

$$\gamma_1(0) = p,\ \ \gamma_2(0) = p,$$

$$\gamma_1'(0) = (\cos\theta_0, \sin\theta_0),$$

and

$$\gamma_2'(0) = (\cos(\theta_0 + \theta), \sin(\theta_0 + \theta)).$$

Then,

$$\angle_p(\gamma_1, \gamma_2) = \theta.$$

Let $\gamma_i(t) = (x_i(t), y_i(t))$; then, $\gamma_i'(0) = (x_i'(0), y_i'(0))$. Note that

$$\phi(\gamma_i(t)) = (x_i(t) - y_i(t), x_i(t) + y_i(t)),$$

and thus,

$$\frac{d\phi(\gamma_i(t))}{dt} = (x_i'(t) - y_i'(t), x_i'(t) + y_i'(t)).$$

Hence,

$$\left.\frac{d\phi(\gamma_1(t))}{dt}\right|_{t=0} = (\cos\theta_0 - \sin\theta_0, \cos\theta_0 + \sin\theta_0) = \sqrt{2}\left(\cos\left(\theta_0 + \frac{\pi}{4}\right), \sin\left(\theta_0 + \frac{\pi}{4}\right)\right),$$

and similarly,

$$\left.\frac{d\phi(\gamma_2(t))}{dt}\right|_{t=0} = \sqrt{2}\left(\cos\left(\theta_0 + \theta + \frac{\pi}{4}\right), \sin\left(\theta_0 + \theta + \frac{\pi}{4}\right)\right).$$

Therefore,

$$\angle_{\phi(p)}(\phi(\gamma_1), \phi(\gamma_2)) = \theta = \angle_p(\gamma_1, \gamma_2),$$

and thus, ϕ preserves angles.

It is easy to see that a composition of angle-preserving maps is also angle-preserving. Since a reflection of the Euclidean plane clearly preserves angles and every isometry of the Euclidean plane is a composition of reflections, an isometry of the Euclidean plane preserves angles. Similarly, one can conclude that an isometry of the sphere also preserves angles.

Theorem 3.5. *The stereographic projection Φ preserves angles.*

Proof. Let p be a point on the sphere and $\mathbf{t}_1$ and $\mathbf{t}_2$ be two tangent vectors of $\mathbb{S}^2$ at p separated by angle θ. We can find circles C_1 and C_2 on $\mathbb{S}^2$ that pass through p and

the north pole N such that $\mathbf{t}_1$ and $\mathbf{t}_2$ are tangent vectors of C_1 and C_2, respectively, at p (Figure 3.5). Hence,

$$\angle_p(C_1, C_2) = \theta.$$

Note that

$$\angle_N(C_1, C_2) = \angle_p(C_1, C_2) = \theta.$$

Then, there exist planes P_1 and P_2 such that

$$C_1 = \mathbb{S}^2 \cap P_1$$

and

$$C_2 = \mathbb{S}^2 \cap P_2.$$

Note that these planes pass through N. Therefore,

$$\Phi(C_1) = P_1 \cap P_{xy}$$

and

$$\Phi(C_2) = P_2 \cap P_{xy},$$

where P_{xy} is the xy-plane. Let $L_1 = \Phi(C_1)$ and $L_2 = \Phi(C_2)$. We need to show that

$$\angle_{\Phi(p)}(L_1, L_2) = \theta.$$

Let

$$P' = \{(x, y, z) \in \mathbb{R}^3 \mid z = 1\},$$

$$L_1' = P_1 \cap P'$$

and

$$L_2' = P_2 \cap P'.$$

Note that P' is the tangent plane of $\mathbb{S}^2$ at N. Hence, L_1' and L_2' are tangent lines of C_1 and C_2, respectively, at N. Therefore,

$$\angle_N(L_1', L_2') = \angle_N(C_1, C_2) = \theta.$$

Since P' is parallel to P_{xy}, L_1' and L_2' are parallel to L_1 and L_2, respectively. Therefore,

$$\angle_{\Phi(p)}(L_1, L_2) = \angle_N(L_1', L_2') = \theta,$$

and the proof is complete (Figure 3.5).

Now let γ_1 and γ_2 be curves that pass through the point p, with

$$\angle_p(\gamma_1, \gamma_2) = \theta.$$

Then, there are circles C_1 and C_2 that pass through N and are tangent to the curves γ_1 and γ_2 at point p. Hence,

$$\angle_p(C_1, C_2) = \theta.$$

As shown previously,

$$\angle_{\Phi(p)}(\Phi(C_1), \Phi(C_2)) = \theta,$$

and the curves $\Phi(\gamma_i)$, $\Phi(C_i)$ are still tangent at $\Phi(p)$ for $i = 1, 2$. Therefore,

$$\angle_{\Phi(p)}(\Phi(\gamma_1), \Phi(\gamma_2)) = \angle_{\Phi(p)}(\Phi(C_1), \Phi(C_2)) = \theta,$$

and the proof is finished.

□

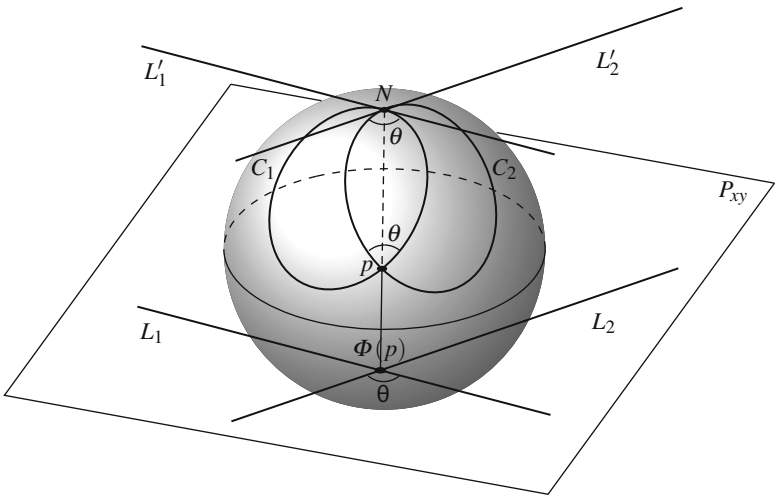


Fig. 3.5 The stereographic projection preserves angles

Exercises

3.1. Show that the stereographic projection does not preserve distances, i.e., there exist some points $p_1, p_2 \in \mathbb{S}^2$ such that

$$d_{\mathbb{S}^2}(p_1, p_2) \neq d\left(\Phi(p_1), \Phi(p_2)\right).$$

3.2. Show that the stereographic projection does not preserve areas, i.e., there exists some region $\Omega \subset \mathbb{S}^2$ such that

$$\text{Area}(\Omega) \neq \text{Area}\left(\Phi(\Omega)\right).$$

3.3. Note that the formula in Proposition 3.2 for Φ can be extended to any point $p = (x, y, z) \in \mathbb{R}^3$ with $z \neq 1$. Consider a circle C on $\mathbb{S}^2$ that is not a great circle. Assume that C does not go through the north pole. Then, $\Phi(C)$ is a circle on $\mathbb{R}^2$. Note that there exists a cone in $\mathbb{R}^3$ tangent to $\mathbb{S}^2$ along C. Let q be its apex.

Show that $\Phi(q)$ is the center of the circle $\Phi(C)$.

3.4. Determine whether each of the following preserves angles:

(a) $\phi : \mathbb{R}^2 \to \mathbb{R}^2$, defined by $\phi(x, y) = (x, 2y)$.
(b) $\phi : \mathbb{R}^2 \to \mathbb{R}^2$, defined by $\phi(x, y) = (2x, 2y)$.

3.5. Prove that the map

$$\phi : \mathbb{R}^2 \to \mathbb{R}^2,$$

defined by

$$\phi(x, y) = (x^2 - y^2, 2xy),$$

preserves angles.

3.2 Inversions on the Extended Plane

Definition 3.6. Let $f : \mathbb{R}^2_\infty \to \mathbb{R}^2_\infty$ and $g : \mathbb{S}^2 \to \mathbb{S}^2$ be maps such that

$$g = \Phi^{-1} \circ f \circ \Phi$$

or, equivalently,

$$f = \Phi \circ g \circ \Phi^{-1}.$$

Then, f is said to be *induced by* g, and g is said to be *induced by* f.

We are interested in the maps of the extended plane induced by isometries of the sphere. The following observation is our starting point.

Proposition 3.7. *Let $I : \mathbb{R}^2_\infty \to \mathbb{R}^2_\infty$ be the map induced by the reflection $\bar{r}_G$ of the sphere in the great circle $G = \{(x, y, z) \in \mathbb{S}^2 \mid z = 0\}$; then,*

$$I(u, v) = \frac{1}{u^2 + v^2}(u, v)$$

for $(u, v) \neq (0, 0)$, $I(\mathbf{0}) = \infty$ and $I(\infty) = \mathbf{0}$.

Proof. The last two are obvious. Note that $\bar{r}_G(x, y, z) = (x, y, -z)$. Therefore, for $(u, v) \neq (0, 0)$,

$$\begin{aligned}
I(u, v) &= \left(\Phi \circ \bar{r}_G \circ \Phi^{-1}\right)(u, v) \\
&= \Phi\left(\bar{r}_G\left(\Phi^{-1}(u, v)\right)\right) \\
&= \Phi\left(\bar{r}_G\left(\frac{2u}{u^2 + v^2 + 1}, \frac{2v}{u^2 + v^2 + 1}, \frac{u^2 + v^2 - 1}{u^2 + v^2 + 1}\right)\right) \\
&= \Phi\left(\frac{2u}{u^2 + v^2 + 1}, \frac{2v}{u^2 + v^2 + 1}, -\frac{u^2 + v^2 - 1}{u^2 + v^2 + 1}\right) \\
&= \left(\frac{2u}{2(u^2 + v^2)}, \frac{2v}{2(u^2 + v^2)}\right) \\
&= \frac{1}{u^2 + v^2}(u, v).
\end{aligned}$$

□

For each point p ($\neq \mathbf{0}$),

$$\overline{\mathbf{0}p} \cdot \overline{\mathbf{0}I(p)} = 1$$

and

$$p - \mathbf{0} = a\,(I(p) - \mathbf{0})$$

for some $a > 0$. Before generalizing the map I, we introduce a term that includes circles and lines.

Definition 3.8. A *circline* is a circle or

$$L \cup \{\infty\}$$

in $\mathbb{R}^2_\infty$, where L is a line in $\mathbb{R}^2$.

For any given distinct three points (including ∞) in the extended plane, there exists a unique circline that goes through these three points.

Definition 3.9. Let C be a circline. The *inversion* in C, denoted by I_C, is a map

$$I_C : \mathbb{R}^2_\infty \to \mathbb{R}^2_\infty$$

defined as follows.

a. If $C = L \cup \{\infty\}$ for a line L, then $I_C(p) = \bar{r}_L(p)$ for $p \neq \infty$ and $I_C(\infty) = \infty$.
b. If C is a circle of radius r centered at q, then for p ($\neq q, \infty$), $I_C(p)$ is a unique point such that

$$\overline{qp} \cdot \overline{q\ I_C(p)} = r^2$$

and

$$p - q = a\,(I_C(p) - q)$$

for some $a > 0$. Explicitly,

$$I_C(p) = \frac{r^2}{\|p - q\|^2}(p - q) + q$$

for $p \neq q$. Moreover, $I_C(q) = \infty$ and $I_C(\infty) = q$ (Figure 3.6).

Note that the previously defined map I is the inversion in the unit circle centered at the origin. It is trivial to verify that an inversion is bijective. It satisfies properties very similar to those satisfied by a reflection in a line.

a. $I_C(p) = p$ for every point $p \in C$.
b. $I_C^{-1} = I_C$.
c. All points outside of the circle C are mapped to the inside of C, and with the exception of the circle's center, vice versa.

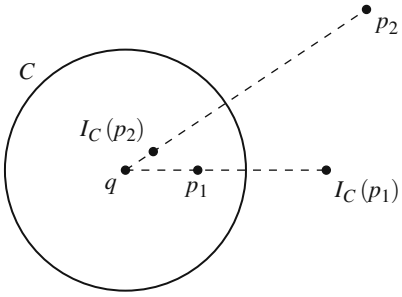


Fig. 3.6 Inversion I_C in a circle C

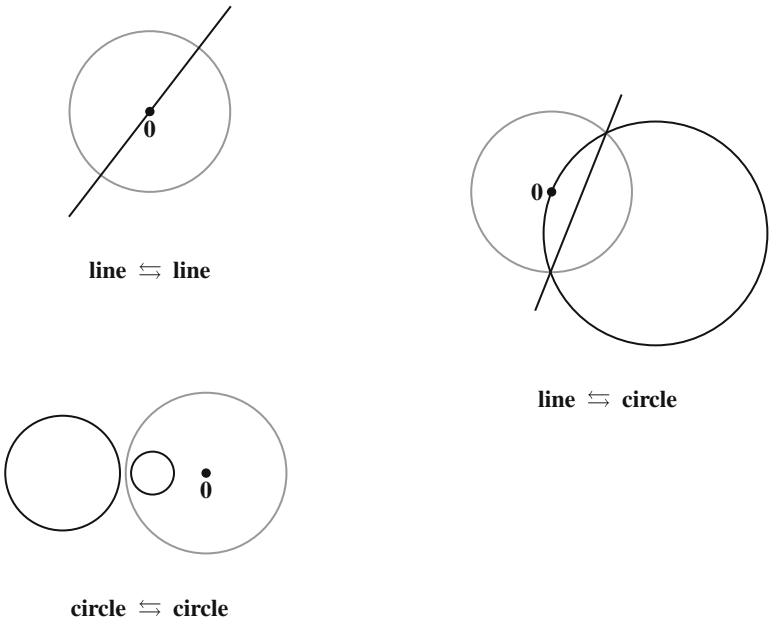


Fig. 3.7 The inversion I maps a circline to a circline

For $p \in \mathbb{R}^2$ and $r > 0$, let $C_{p,r}$ be a circle with center p and radius r. It is not difficult to verify that

1. $I_{C_{\mathbf{0},r}} = d_r \circ I \circ d_{\frac{1}{r}}$ and
2. $I_{C_{p,r}} = t_p \circ I_{C_{\mathbf{0},r}} \circ t_{-p}$,

where $d_r : \mathbb{R}^2_\infty \to \mathbb{R}^2_\infty$ is defined by $d_r(q) = rq$ (Exercise 3.6).

Proposition 3.10. *The inversion I maps a circline to a circline (Figure 3.7).*

Proof. Let C be a circline. By Proposition 3.7, we have

$$I = \Phi \circ \bar{r}_G \circ \Phi^{-1},$$

where G is the equator on $\mathbb{S}^2$. $\Phi^{-1}(C)$ is a circle on $\mathbb{S}^2$, and $\bar{r}_G(\Phi^{-1}(C))$ is a circle on $\mathbb{S}^2$; thus,

$$I(C) = \Phi(\bar{r}_G(\Phi^{-1}(C)))$$

is a circline, where we used Theorem 3.4. □

Proposition 3.11. *The inversion I preserves angles.*

Proof. According to Theorem 3.5, the stereographic projection Φ preserves angles. Hence, Φ^{-1} also preserves angles. Note that I is the composition

$$I = \Phi \circ \bar{r}_G \circ \Phi^{-1}$$

of the angle-preserving maps Φ, $\bar{r}_G$, and Φ^{-1}, where $\bar{r}_G$ is a reflection of the sphere. Therefore, I preserves angles.

□

Corollary 3.12. *For a circline C,*

a. I_C maps a circline to a circline, and
b. I_C preserves angles.

Proof. If C is a line, then I_C is a reflection, so the above obviously holds. Otherwise, $C = C_{p,r}$ is a circle with center p and radius r. Note (Exercise 3.6) that

$$I_{C_{p,r}} = t_p \circ d_r \circ I \circ d_{\frac{1}{r}} \circ t_{-p}.$$

Since the maps t_p, d_r, I, $d_{\frac{1}{r}}$, and t_{-p} send circlines to other circlines and preserve angles, $I_{C_{p,r}}$ does the same. □

For a plane P in $\mathbb{R}^3$, let

$$P_\infty = P \cup \{\infty\}.$$

The circlines on P_∞ and inversions in them can be similarly defined. One can understand the stereographic projection via those inversions. Let p be a point on $\mathbb{S}^2$ that is different from the north pole $N = (0, 0, 1)$. Then the three points p, N, and the origin $\mathbf{0}$ altogether determine a plane P that contains all of them. Then it is obvious that the plane also contains the points $(\Phi(p), 0)$ (simply denoted by $\Phi(p)$) on the xy-plane. Let

$$\Gamma = P \cap \mathbb{S}^2$$

and L be the intersection of P with the xy-plane, then Γ is a circle on P whose center is $\mathbf{0}$ and its radius is 1. Draw another circle Σ on the plane P with the center N such that it passes through the two points where the circle Γ and the line L intersect. See Figure 3.8.

Consider the inversion

$$I_\Sigma : P_\infty \to P_\infty$$

in Σ on P_∞. Draw a line l through N and p, then

$$\{\Phi(p), \infty\} = L \cap l.$$

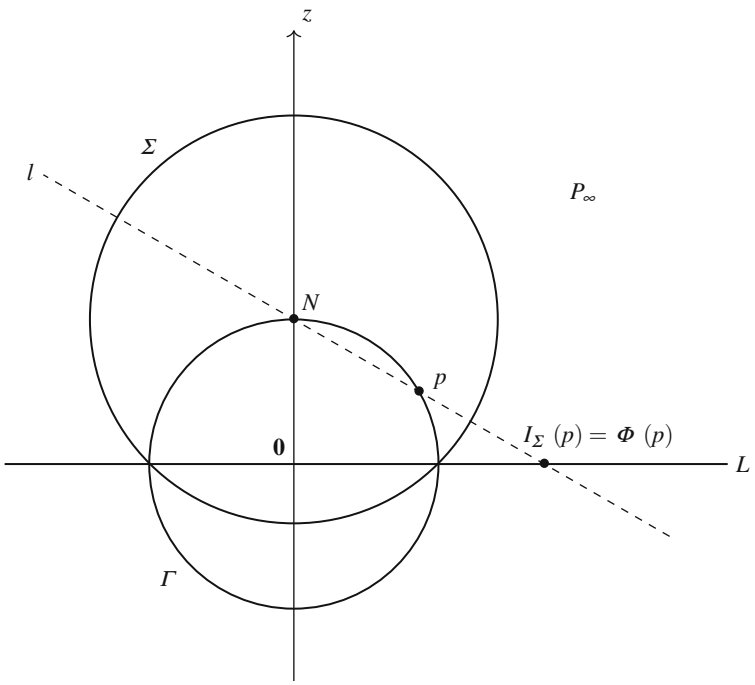


Fig. 3.8 $I_\Sigma(p) = \Phi(p)$

It is easy to see that

$$I_\Sigma(\Gamma) = L, \;\; I_\Sigma(l) = l.$$

Since

$$\Gamma \cap l = \{N, p\},$$

$$\{I_\Sigma(p), I_\Sigma(N)\} = I_\Sigma(\Gamma \cap l) = I_\Sigma(\Gamma) \cap I_\Sigma(l) = L \cap l = \{\Phi(p), \infty\}.$$

Because $I_\Sigma(N) = \infty$, we conclude that

$$I_\Sigma(p) = \Phi(p). \tag{3.1}$$

This viewpoint will be extended more when we introduce inversions in spheres in Section 3.3.

A converse of Corollary 3.12 also holds, as stated in the following lemma.

Lemma 3.13. *Let C be a circline on $\mathbb{R}^2_\infty$. Then, the set $\mathbb{R}^2_\infty - C$ is composed of two connected regions R_1 and R_2. For a bijective map $f : \mathbb{R}^2_\infty \to \mathbb{R}^2_\infty$, if*

1. *f fixes each point on C,*
2. *$f(R_1) = R_2$,*
3. *f maps a circline to a circline, and*
4. *f preserves angles,*

then

$$f = I_C.$$

Proof. If $p \in C$, then $f(p) = p = I_C(p)$. Assume that $p \notin C$. We will show that $f(p) = I_C(p)$.

Case 1. If

$$C = L \cup \{\infty\},$$

where L is a line, we have that

$$p \neq \infty \text{ and } f(\infty) = \infty$$

because the point ∞ belongs to C. Since f is bijective, $f(p) \neq \infty$. Draw a circle Γ that has the points p and $f(p)$ as its antipodal points and a line M that goes through p and $f(p)$. Then, M and Γ meet each other orthogonally at the points p and $f(p)$. Let

$$L \cap \Gamma = \{p_1, p_2\},$$

then the circline $f(\Gamma)$ goes through three distinct points $f(p)$, p_1, and p_2 that lie on Γ. Hence, $f(\Gamma) = \Gamma$. The lines L and M meet at a single point q. The circline

$$f(M \cup \{\infty\})$$

goes through two distinct points $f(p)$ and q that lie on M. Since the map f preserves angles, $f(M \cup \{\infty\})$ meets $f(\Gamma) = \Gamma$ orthogonally with Γ at $f(p)$, where Γ meets orthogonally with M. Hence,

$$f(M \cup \{\infty\}) = M \cup \{\infty\}.$$

Since $f(\infty) = \infty$, we conclude that $f(M) = M$. Note that $I_C(M) = M$ and $I_C(\Gamma) = \Gamma$. Now it is obvious that

$$f(p) = \bar{r}_L(p) = I_C(p).$$

See Figure 3.9.

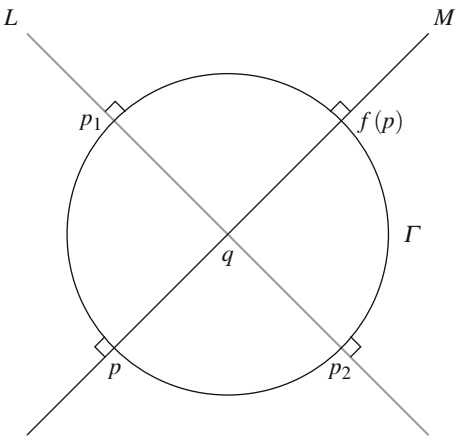


Fig. 3.9 The map f is an inversion ($C = L \cup \{\infty\}$)

Case 2. It remains to consider the case that C is a circle. For $p \notin C$, first, assume that neither p nor $f(p)$ is ∞. Draw a circle Γ that has the points p and $f(p)$ as its antipodal points and a line M that goes through p and $f(p)$. Then, M and Γ meet each other orthogonally at the points p and $f(p)$. Let

$$C \cap \Gamma = \{p_1, p_2\},$$

then the circline $f(\Gamma)$ goes through three distinct points $f(p)$, p_1, and p_2 that lie on Γ. Hence, $f(\Gamma) = \Gamma$. Since the map f preserves angles, Γ and C meet each other orthogonally. Note that C and M meet at two points q_1 and q_2. The circline

$$f(M \cup \{\infty\})$$

goes through three distinct collinear points q_1, q_2, and $f(p)$ that lie on M. Hence,

$$f(M \cup \{\infty\}) = f(M \cup \{\infty\}).$$

See Figure 3.10.

Note that $I_C(M) = M$ and $I_C(\Gamma) = \Gamma$. Hence,

$$\{p, f(p)\} = M \cap \Gamma = I_C(M) \cap I_C(\Gamma) = I_C(M \cap \Gamma) = \{I_C(p), I_C(f(p)\}.$$

If $p \in R_1$, then $f(p) \in R_2$ and $I_C(p) \in R_2$. Therefore, $f(p) = I_C(p)$.

Now, consider the situation that p or $f(p)$ is ∞. Let α be the center of the circle α. Suppose that $f(\alpha) \neq \infty$, then the previous argument applies and we conclude that

$$f(\alpha) = I_C(\alpha) = \infty,$$

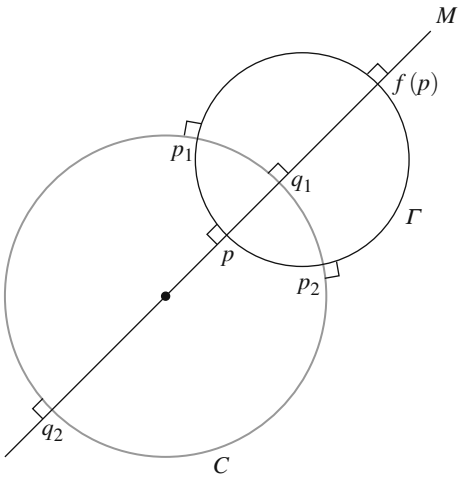


Fig. 3.10 The map f is an inversion (C is a circle)

which is a contradiction. Hence $f(\alpha) = \infty$. We showed that

$$f(p) = I_C(p)$$

for $p \in \mathbb{R}^2$. Note that

$$I_C(\mathbb{R}^2) = \mathbb{R}^2_\infty - \{\alpha\}.$$

Hence, the bijectivity of f implies that

$$f(\infty) = \alpha = I_C(\infty).$$

□

Now we can prove the following theorem.

Theorem 3.14. *A reflection of the sphere in a great circle induces an inversion of the extended plane.*

Proof. Let $\phi = \Phi \circ \bar{r}_G \circ \Phi^{-1}$ be the map on $\mathbb{R}^2_\infty$ induced by a reflection $\bar{r}_G$ of the sphere in the great circle G. Let $C = \Phi(G)$, which is a circline. We will show that $\phi = I_C$. First, we show that ϕ satisfies conditions 1 and 2 in Lemma 3.13. For $p \in C$, $\Phi^{-1}(p) \in G$. Hence,

$$\phi(p) = \Phi(\bar{r}_G(\Phi^{-1}(p))) = \Phi(\Phi^{-1}(p)) = p.$$

This is the first condition. Let Q_1 and Q_2 be the connected regions of $\mathbb{S}^2 - G$ and $R_i = \Phi(Q_i)$ for $i = 1, 2$. Then R_1 and R_2 are the connected regions of $\mathbb{R}^2_\infty - C$. For $p \in R_1$,

$$\Phi^{-1}(p) \in Q_1 \Rightarrow \bar{r}_G(\Phi^{-1}(p)) \in Q_2 \Rightarrow \Phi(\bar{r}_G(\Phi^{-1}(p))) \in R_2.$$

Since $\Phi(\bar{r}_G(\Phi^{-1}(p))) = \phi(p)$, ϕ satisfies the second condition.

For a circline Γ, $\Phi^{-1}(\Gamma)$ and $\bar{r}_G(\Phi^{-1}(\Gamma))$ are circles on $\mathbb{S}^2$, and thus,

$$\phi(\Gamma) = \Phi(\bar{r}_G(\Phi^{-1}(\Gamma)))$$

is a circline, which is the third condition in Lemma 3.13 for ϕ. Since the maps Φ, $\bar{r}_G$, and Φ^{-1} all preserve angles, $\phi = \Phi \circ \bar{r}_G \circ \Phi^{-1}$ also preserves angles. This is the final condition in Lemma 3.13. Thus, $\phi = I_C$. □

The following will be useful in proving several theorems.

Lemma 3.15. *For two circlines C and Γ on $\mathbb{R}^2_\infty$,*

$$I_C \circ I_\Gamma \circ I_C = I_{\Gamma'},$$

where $\Gamma' = I_C(\Gamma)$.

Proof. We will use the result of Lemma 3.13. Let

$$f = I_C \circ I_\Gamma \circ I_C.$$

We will check the conditions for f in Lemma 3.13 to show that

$$f = I_{\Gamma'}.$$

1. Let $p \in \Gamma'$, then $p = I_C(q)$ for some $q \in \Gamma$.

$$\begin{aligned} f(p) &= f(I_C(q)) \\ &= (I_C \circ I_\Gamma \circ I_C)(I_C(q)) \\ &= (I_C \circ I_\Gamma)(q) \\ &= I_C(q) \\ &= p \end{aligned}$$

2. The set $\mathbb{R}^2_\infty - \Gamma$ is composed of two connected regions R_1 and R_2. We need to show that

$$f(R_1) = R_2 \text{ and } f(R_2) = R_1.$$

Since $f(\Gamma) = \Gamma$ as shown in the above and f sends a connected region to a connected region, $f(R_1) = R_1$ or R_2. Suppose that $f(R_1) = R_1$. For any points

$q \in R_1$, there are two distinct circlines C_1, C_2, passing through q, which intersect orthogonally with Γ. Let

$$c_i = C_i \cap (\Gamma \cup R_1).$$

Then

$$c_1 \cap c_2 = \{q\}.$$

The end points of c_i lie on Γ and so they are fixed by the f, $f(c_i)$ is a part of a circline and it meets orthogonally with $f(\Gamma) = \Gamma$. Hence, $f(c_i) = c_i$.

$$\begin{aligned} \{f(q)\} &= f(\{q\}) \\ &= f(\{c_1 \cap c_2\}) \\ &= f(c_1) \cap f(c_2) \\ &= c_1 \cap c_2 \\ &= \{q\} \end{aligned}$$

and so $f(q) = q$, i.e.,

$$(I_C \circ I_\Gamma \circ I_C)(q) = q,$$

which implies

$$I_\Gamma(I_C(q)) = I_C(q).$$

We conclude that $I_C(q)$ belongs to the circline Γ. Since q can be any points in R_1,

$$I_C(R_1) \subset \Gamma,$$

which is impossible. Hence, $f(R_1) \neq R_1$ and so $f(R_1) = R_2$.

3. Each of the maps I_C, I_Γ sends a circline to a circline. Hence, the map

$$f = I_C \circ I_\Gamma \circ I_C$$

sends a circline to a circline.

4. Each of the maps I_C, I_Γ preserves angles. Hence, the map f preserves angles.

□

Exercises

3.6. Show the following:

(a) $I_{C_{\mathbf{0},r}} = d_r \circ I \circ d_{\frac{1}{r}}$.
(b) $I_{C_{p,r}} = t_p \circ I_{C_{\mathbf{0},r}} \circ t_{-p}$.

3.7. Let

$$\phi : \mathbb{R}^2_\infty \to \mathbb{R}^2_\infty$$

be the map induced by the antipodal map $\hat{a}$. Show that $\phi = -I$.

3.8. Show the following:

(a)

$$d\left(I(p), I(q)\right) = \frac{1}{\|p\|\|q\|} d(p, q)$$

for $p, q \in \mathbb{R}^2$,

(b)

$$d\left(I_C(p), I_C(q)\right) = \frac{r^2}{\|p - \alpha\|\|q - \alpha\|} d(p, q)$$

for $p, q \in \mathbb{R}^2$, where $C = C_{\alpha,r}$.

3.9. A Steiner chain is a collection of finitely many circles on $\mathbb{R}^2$, all of which are tangent to two given non-intersecting circles α and β (the center circle and the outer circle in Figure 3.11) such that each circle in the chain is tangent to the previous and subsequent circles in the chain. Note that the first and last circles are also tangent to each other. A Steiner chain of 17 circles is shown in Figure 3.11.

Prove the following:

"If at least one Steiner chain of n circles exists for two given circles α and β, then there is an infinite number of Steiner chains of n circles for the circles α and β"

(*Hint.* Consider an inversion that maps α and β to concentric circles.)

3.10. In Figure 3.12, starting with the circle P_1 tangent to the three semicircles forming the arbelos, construct a chain of tangent circles P_i all tangent to one of the two small interior circles and to the large exterior one. This chain is called the *Pappus chain*. Let r_n be the radius of P_n and h_n be the distance to the line AB from the center of P_n. Show that

$$h_n = 2nr_n.$$

(*Hint.* Consider an inversion that produces Figure 3.13.)

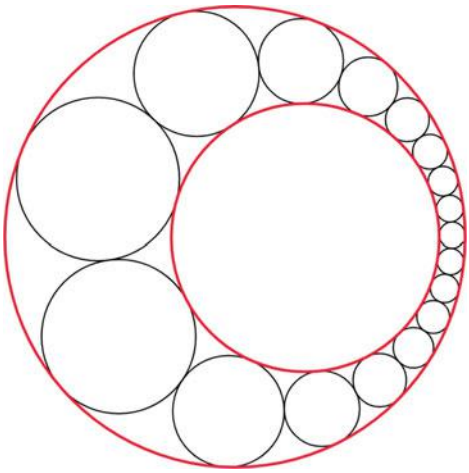


Fig. 3.11 A Steiner chain of 17 circles

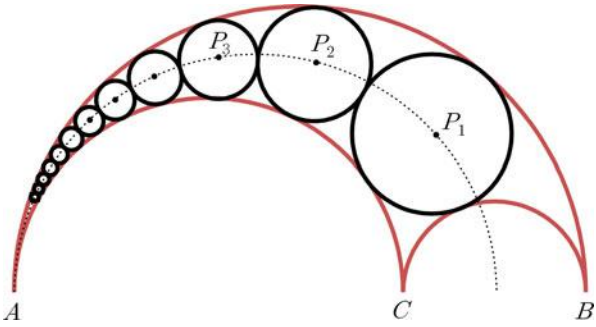


Fig. 3.12 A Pappus chain

3.3 Inversions on the Sphere $\mathbb{S}^2$

We showed that a reflection on $\mathbb{S}^2$ induces an inversion on $\mathbb{R}^2_\infty$. Then what kind of maps on $\mathbb{S}^2$ is induced by an inversion on $\mathbb{R}^2_\infty$? Answering this question will be our next task.

Let us define inversions in circles on the sphere.

Definition 3.16. Let C be a circle on the sphere. The *inversion in the circle* C is a map

$$I_C : \mathbb{S}^2 \to \mathbb{S}^2$$

defined as follows:

(a) If C is a great circle, then I_C is defined as the reflection $\bar{r}_C$ in C.
(b) Otherwise, let q be the apex of the cone in $\mathbb{R}^3$ tangent to $\mathbb{S}^2$ along C. For each point p on $\mathbb{S}^2$, if p lies on C, then $I_C(p) = p$. Otherwise, the point $I_C(p)$

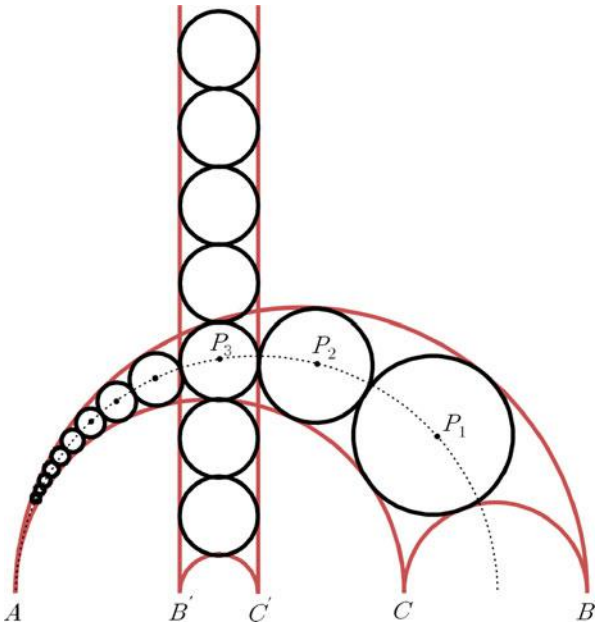
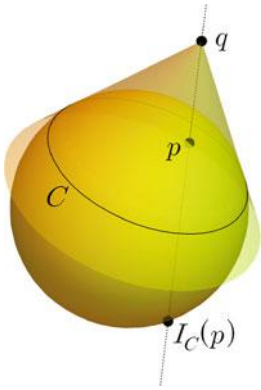


Fig. 3.13 An inversion of a Pappus chain

Fig. 3.14 Inversion I_C in a circle C on $\mathbb{S}^2$

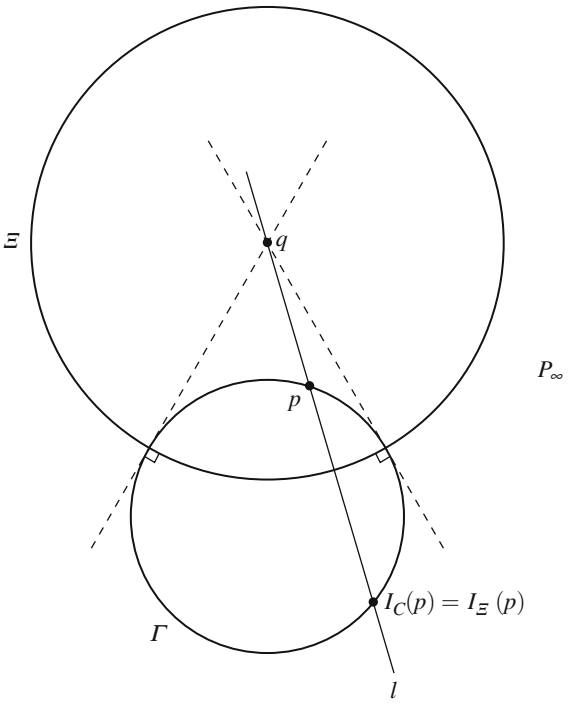


on $\mathbb{S}^2$ is defined as the second intersection of the straight line through q and p (Figure 3.14).

Clearly, an inversion on the sphere is bijective. We will show that inversions on the sphere exhibit properties similar to those of the inversions on the extended plane.

One can understand the inversions on the sphere via inversions in planes when C is not a great circle. For a given point p on $\mathbb{S}^2$, choose a plane P in $\mathbb{R}^3$ which contains p, q and the center on $\mathbb{S}^2$ of the circle C. Then it is obvious that the plane also contains the points $I_C(p)$. Let

Fig. 3.15 $I_C(p) = I_\Xi(p)$



$$\Gamma = P \cap \mathbb{S}^2,$$

then Γ is a circle on P whose center is $\mathbf{0}$ and its radius is 1. Draw another circle Ξ on the plane P with the center q such that it passes through the two points where the circle C and the plane P intersect. See Figure 3.15.

Note that the circle Γ intersects orthogonally with the circle Ξ. Consider the inversion

$$I_\Xi : P_\infty \to P_\infty$$

in Ξ on P_∞. Draw a line l through q and p, then

$$\{p, I_C(p)\} = \Gamma \cap l.$$

It is easy to see that

$$I_\Xi(\Gamma) = \Gamma, \ \ I_\Xi(l) = l.$$

Since

$$\Gamma \cap l = \{N, p\},$$

$$\{I_\Xi(p), I_\Xi(I_C(p))\} = I_\Xi(\Gamma \cap l) = I_\Xi(\Gamma) \cap I_\Xi(l) = \Gamma \cap l = \{p, I_C(p)\}.$$

Because $I_\Xi(p) \neq p$, we conclude that

$$I_\Xi(p) = I_C(p). \tag{3.2}$$

Proposition 3.17. *An inversion on the sphere preserves angles and maps a circle to a circle.*

Proof. Let C be a circle on the sphere. If C is a great circle, then I_C is the reflection in C, which clearly satisfies the required properties. □

Assume that C is not a great circle. Let q be the apex of the cone in $\mathbb{R}^3$ tangent to $\mathbb{S}^2$ along C. By choosing a suitable coordinate system (without changing the origin), we can assume that q lies on the z-axis with positive z-coordinate.

First, we prove the following claim:

Claim. For each point p on $\mathbb{S}^2$,

$$I_C(p) = (\Phi^{-1} \circ I_{C'} \circ \Phi)(p),$$

where $C' = \Phi(C)$.

Proof of Claim. Choose a plane P in $\mathbb{R}^3$ which contains p, q and the center on $\mathbb{S}^2$ of the circle C. Then it is obvious that the plane also contains the points $I_C(p)$. Let

$$\Gamma = P \cap \mathbb{S}^2$$

and L be the intersection of P with the xy-plane, then Γ is a circle on P whose center is $\mathbf{0}$ and its radius is 1. Draw a circle Σ on the plane P with the center N such that it passes through the two points where the circle Γ and the line L intersect (Figure 3.16).

For the inversion $I_\Sigma : P_\infty \to P_\infty$, we have showed in (3.1):

$$I_\Sigma(\alpha) = \Phi(\alpha) \tag{3.3}$$

for each point $\alpha \in \Gamma$ and so

$$I_\Sigma^{-1}(\beta) = \Phi^{-1}(\beta) \tag{3.4}$$

for each point $\beta \in L$. Let Ξ be the circle on the plane P whose center is the point q such that it passes through the two points where the circle C and the plane P intersect. For the inversion $I_\Xi : P_\infty \to P_\infty$, we also showed in (3.2):

$$I_\Xi(\alpha) = I_C(\alpha) \tag{3.5}$$

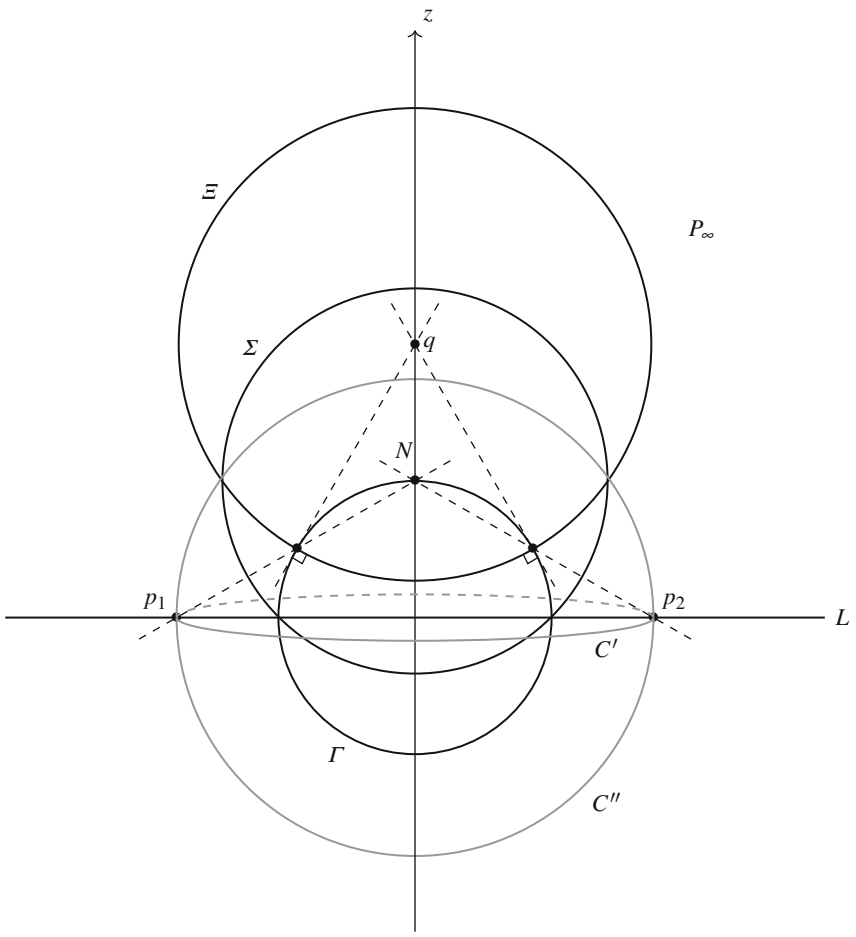


Fig. 3.16 $I_C(p) = (\Phi^{-1} \circ I_{C'} \circ \Phi)(p)$

for each point $\alpha \in \Gamma$. Note that the circles C', C'' and the line L intersect orthogonally with another at the points p_1, p_2. Hence,

$$I_{C'}(\alpha) = I_{C''}(\alpha) \tag{3.6}$$

for each point α on the line L. The circles Ξ and Γ intersect orthogonally with each other at two points whose images under the inversion I_Σ are p_1, p_2. Since $I_\Sigma(\Gamma) = L_\infty$ and C'' intersects orthogonally with L at p_1, p_2, we conclude that $I_\Sigma(\Xi) = C''$ and so

$$I_\Sigma(C'') = \Xi. \tag{3.7}$$

Finally, we have

$$\begin{aligned}
(\Phi^{-1} \circ I_{C'} \circ \Phi)(p) &= \Phi^{-1}\left(I_{C'}\left(\Phi(p)\right)\right) \\
&= \Phi^{-1}\left(I_{C'}\left(I_{\Sigma}(p)\right)\right) && (\because (3.3)) \\
&= \Phi^{-1}\left(I_{C''}\left(I_{\Sigma}(p)\right)\right) && (\because (3.6)) \\
&= I_{\Sigma}^{-1}\left(I_{C''}\left(I_{\Sigma}(p)\right)\right) && (\because (3.4)) \\
&= I_{\Sigma}\left(I_{C''}\left(I_{\Sigma}(p)\right)\right) \\
&= \left(I_{\Sigma} \circ I_{C''} \circ I_{\Sigma}\right)(p) \\
&= I_{I_{\Sigma}(C'')}(p) && (\because \text{Lemma } 3.15) \\
&= I_{\Xi}(p) && (\because (3.7)) \\
&= I_C(p), && (\because (3.5))
\end{aligned}$$

which completes the proof of Claim.

According to Claim,

$$I_C = \Phi^{-1} \circ I_{C'} \circ \Phi,$$

which is a composition of maps that preserve angles and map a circle to a circle. Hence, the map I_C also preserves angles and maps a circle to a circle. □

Theorem 3.18. *Every inversion on the sphere induces an inversion on the extended plane, and every inversion on the extended plane induces an inversion on the sphere.*

Proof. Let C be a circle on the sphere, and let

$$f = \Phi \circ I_C \circ \Phi^{-1}$$

be the map on the extended plane, induced by the inversion I_C on the sphere. Let $C' = \Phi(C)$, which is a circline on $\mathbb{R}^2_\infty$. Then, the set $\mathbb{R}^2_\infty - C'$ is composed of two connected regions R_1 and R_2. It is trivial to verify that f fixes each point on C' and $f(R_1) = R_2$. In Proposition 3.17, we showed that I_C preserves angles and maps a circle to a circle, which implies that f maps a circline to a circline and preserves angles. Now we have checked all the conditions in Lemma 3.13, so we can conclude that f is the inversion in C'.

Conversely, let Γ be a circline on the extended plane and $C = \Phi^{-1}(\Gamma)$. Note that C is a circle on the sphere. We also note that $\Phi(C) = \Gamma$. We showed that the inversion I_C on the sphere induced the inversion I_Γ on the extended plane:

$$\Phi \circ I_C \circ \Phi^{-1} = I_\Gamma.$$

Hence,

$$I_C = \Phi^{-1} \circ I_\Gamma \circ \Phi,$$

i.e., I_Γ induces the inversion I_C on the sphere. □

One can define an inversion in a sphere on $\mathbb{R}^3$ in a similar way. Let S be a sphere of radius r, centered at q in $\mathbb{R}^3$. The inversion

$$I_S : \mathbb{R}^3 - \{q\} \to \mathbb{R}^3$$

in S is defined to satisfy the relations

$$\overline{qp} \cdot \overline{qI_S(p)} = r^2$$

and

$$p - q = a\,(I_S(p) - q)$$

for some $a > 0$ and any point $p \in \mathbb{R}^3 - \{q\}$. Concretely,

$$I_S(p) = \frac{r^2}{\|p - q\|^2}(p - q) + q.$$

One can show that an inversion in a sphere satisfies properties similar to those satisfied by an inversion in a circle on the extended plane.

The stereographic projection can be understood by using an inversion in a sphere. It is not difficult to verify that

$$\Phi(p) = I_S(p)$$

for each $p \in \mathbb{S}^2 - \{N\}$, where S is a sphere of radius $\sqrt{2}$ centered at the north pole N.

Exercises

3.11. Let q be a point $\mathbb{R}^3$, with $\|q\| < 1$. For each point $p \in \mathbb{S}^2$, the line through p and q intersects $\mathbb{S}^2$ at another point p' (Figure 3.17).

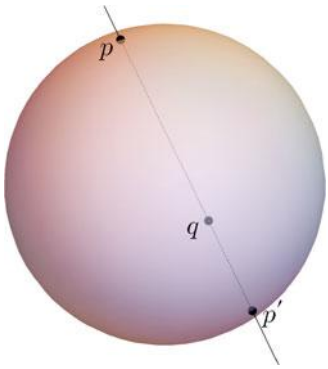
Define a map

$$\hat{a}_q : \mathbb{S}^2 \to \mathbb{S}^2$$

by $\hat{a}_q(p) = p'$. Note that $\hat{a}_\mathbf{0}$ is the antipodal map $\hat{a}$.

Show that $\hat{a}_q$ preserves angles and maps a circle to a circle.

Fig. 3.17 $\hat{a}_q(p) = p'$



(*Hint.* Show that $\hat{a}_q$ is a composition of a rotation by the angle π and an inversion.)

3.12. Show that an inversion in a circle (not a great circle) on $\mathbb{S}^2$ is a restriction to $\mathbb{S}^2$ of an inversion in a sphere. In concrete words, for an inversion I_C of $\mathbb{S}^2$ in a circle C (not a great circle) on $\mathbb{S}^2$, show that there is a sphere S in $\mathbb{R}^3$ such that

$$I_C(p) = I_S(p)$$

for each point p on $\mathbb{S}^2$.

3.13.

- The set $\mathbb{R}_\infty = \mathbb{R} \cup \{\infty\}$ is called the *projectively extended real line*. In addition to the standard operations on the subset $\mathbb{R}$ of $\mathbb{R}_\infty$, the following operations are defined for $x \in \mathbb{R}_\infty$, with exceptions as indicated:

$$\begin{aligned} -(\infty) = \infty,\ x + \infty = \infty + x = \infty \ &\text{if } x \neq \infty, \\ x \cdot \infty = \infty \cdot x = \infty \ &\text{if } x \neq 0, \\ x/\infty = 0 \ &\text{if } x \neq \infty, \\ x/0 = \infty \ &\text{if } x \neq 0. \end{aligned}$$

- A linear fractional transformation (also called Möbius transformation) is a map $f : \mathbb{R}_\infty \to \mathbb{R}_\infty$ that has the form

$$f(x) = \frac{ax + b}{cx + d}$$

for some fixed real numbers a, b, c, d with $ad - bc \neq 0$ and $f(\infty) = \frac{a}{c}$.

- For $a \in \mathbb{R}$, the reflection $\bar{r}_a$, defined in Exercise 1.27, is extended to $\mathbb{R}_\infty$ by setting $\bar{r}_a(\infty) = \infty$. For $\alpha \in \mathbb{R}$ and $r > 0$, we define a map $I_{\alpha,r} : \mathbb{R}_\infty \to \mathbb{R}_\infty$ by

$$I_{\alpha,r}(x) = \frac{r^2}{|x-\alpha|^2}(x-\alpha)+\alpha = \frac{r^2}{(\alpha - x)}+\alpha$$

for $x \neq \alpha$ and $I_{\alpha,r}(\alpha) = \infty$.

We call $\bar{r}_a$ and $I_{a,r}$ inversions of $\mathbb{R}_\infty$.

1. For a 2×2 matrix

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix},$$

with entries in $\mathbb{R}$ such that $\det(A) \neq 0$, define a linear fractional transformation f_A by

$$f(x) = \frac{ax+b}{cx+d}.$$

Show that

$$f_A \circ f_B = f_{AB}$$

and

$$f_A^{-1} = f_{A^{-1}}$$

for 2×2 matrices A, B with $\det(A) \neq 0, \det(B) \neq 0$.

2. For given three distinct elements x_2, x_3, x_4 in $\mathbb{R}^3$, show that there exists a unique linear fractional transformation f such that

$$f(x_2) = 1,\ f(x_3) = 0,\ f(x_4) = \infty.$$

For given x_1 in $\mathbb{R}_\infty$, the *cross ratio* $(x_1, x_2; x_3, x_4)$ is defined as the image of x_1 under the linear fractional transformation f. Show that

$$(x_1, x_2; x_3, x_4) = \frac{x_2 - x_4}{x_2 - x_3}\frac{x_1 - x_3}{x_1 - x_4}$$

when x_1, x_2, x_3, x_4 are distinct elements of $\mathbb{R}$.

3. An injective map $f : \mathbb{R}_\infty \to \mathbb{R}_\infty$ is said to preserve the cross ratio if

$$(f(x_1), f(x_2); f(x_3), f(x_4)) = (x_1, x_2; x_3, x_4)$$

for each $x_1 \in \mathbb{R}_\infty$ and any distinct elements x_2, x_3, x_4 of $\mathbb{R}_\infty$.

Show that every linear fractional transformation preserves the cross ratio.

4. For each inversion ψ of $\mathbb{R}_\infty$, show that there exists a unique inversion ϕ of $\mathbb{R}^2_\infty$ such that

$$\phi(x, 0) = (\psi(x), 0)$$

for any $x \in \mathbb{R}_\infty$, where we regard $(\infty, 0)$ (for $\infty \in \mathbb{R}_\infty$) as ∞ (for $\infty \in \mathbb{R}^2_\infty$).

5. For an injective map $f : \mathbb{R}_\infty \to \mathbb{R}_\infty$, show that the following are equivalent:
   a. f is a composition of inversions of $\mathbb{R}_\infty$.
   b. f is a linear fractional transformation of $\mathbb{R}_\infty$.
   c. f preserves the cross ratio.

3.4 Representation of the Sphere in the Extended Plane

In most cultures, ancient people believed that the Earth's shape was a plane or a disk. Imagine that there is a fantasy world, called Sphereland, which is in the shape of a sphere. There are two-dimensional beings livings in Sphereland. They are tiny (size of approximately 10^{-7}), similar to how human beings are very tiny compared with the Earth. Different from us, they have no conception of the three-dimensional space that might exist outside of their world. There lives a tiny but clever mathematician, named Kyuri, in Sphereland. Since she does not travel far enough, compared with the radius of the sphere, her entire world simply looks like the Euclidean plane. Thus, she may well develop Euclidean geometry there as Euclid did in ancient Greece 2000 years ago. Since she is a great mathematician, she succeeds in showing that the value of π (which she thinks is a constant) lies between $\frac{223}{71}$ (approximately 3.1408) and $\frac{22}{7}$ (approximately 3.1429), as did Archimedes. However, after she measures the circumference of some huge circles, she realizes that the number π is not a constant in her world. Finally, she correctly concludes that her world is not flat. Still, she has no conception of three-dimensional space, and she has no choice but to use $\mathbb{R}^2$ to study the geometry of her world. Thus, she maps Sphereland to $\mathbb{R}^2_\infty$ using the stereographic projection. This section is about how she studies her world using $\mathbb{R}^2_\infty$.

Definition 3.19. For $p, q \in \mathbb{R}^2_\infty$, the *stereographic distance* between p and q is defined as

$$d_\Phi(p, q) = d_{\mathbb{S}^2}\left(\Phi^{-1}(p), \Phi^{-1}(q)\right).$$

For a curve C on $\mathbb{R}^2_\infty$, its stereographic length is defined as

$$l_\Phi(C) = l\left(\Phi^{-1}(C)\right).$$

Let $\gamma : [a, b] \to \mathbb{R}^2$ be a smooth plane curve and $\gamma(t) = (x(t), y(t))$. The length of this curve is given by

$$l(\gamma) = \int_a^b \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2}\, dt.$$

Since every reflection on the plane preserves the length of a plane curve, an isometry of the plane preserves the length of the plane curve.

Similarly, for a smooth spherical curve $\gamma : [a, b] \to \mathbb{S}^2$ with $\gamma(t) = (x(t), y(t), z(t))$, its length is

$$l(\gamma) = \int_a^b \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2 + \left(\frac{dz}{dt}\right)^2}\, dt.$$

Since every reflection on the sphere preserves the length of the spherical curve, an isometry of the sphere preserves the length of the spherical curve.

Let $\gamma : [a, b] \to \mathbb{R}^2$ be a smooth curve, with $\gamma(t) = (u(t), v(t))$. Let $(x(t), y(t), z(t)) = \Phi^{-1}(u(t), v(t))$; then,

$$x = \frac{2u}{u^2 + v^2 + 1}, \quad y = \frac{2v}{u^2 + v^2 + 1}, \quad z = \frac{u^2 + v^2 - 1}{u^2 + v^2 + 1}.$$

$$\frac{dx}{dt} = \frac{\partial x}{\partial u}\frac{du}{dt} + \frac{\partial x}{\partial v}\frac{dv}{dt} = \frac{2 - 2u^2 + 2v^2}{(1 + u^2 + v^2)^2}\frac{du}{dt} + \frac{-4uv}{(1 + u^2 + v^2)^2}\frac{dv}{dt},$$

$$\frac{dy}{dt} = \frac{\partial y}{\partial u}\frac{du}{dt} + \frac{\partial y}{\partial v}\frac{dv}{dt} = \frac{-4uv}{(1 + u^2 + v^2)^2}\frac{du}{dt} + \frac{2 + 2u^2 - 2v^2}{(1 + u^2 + v^2)^2}\frac{dv}{dt}$$

and

$$\frac{dz}{dt} = \frac{\partial z}{\partial u}\frac{du}{dt} + \frac{\partial z}{\partial v}\frac{dv}{dt} = \frac{4u}{(1 + u^2 + v^2)^2}\frac{du}{dt} + \frac{4v}{(1 + u^2 + v^2)^2}\frac{dv}{dt}.$$

After routine but lengthy calculations, one can show that

$$\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2 + \left(\frac{dz}{dt}\right)^2 = \frac{4}{(1 + u^2 + v^2)^2}\left(\left(\frac{du}{dt}\right)^2 + \left(\frac{dv}{dt}\right)^2\right).$$

Hence,

$$\begin{aligned} l_\Phi(\gamma) &= l\left(\Phi^{-1}(\gamma)\right) \\ &= \int_a^b \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2 + \left(\frac{dz}{dt}\right)^2}\, dt \end{aligned}$$

$$= \int_a^b \frac{2}{1+u^2+v^2}\sqrt{\left(\frac{du}{dt}\right)^2+\left(\frac{dv}{dt}\right)^2}\,dt.$$

Hence, we have

$$l_\Phi(\gamma) = \int_a^b \frac{2}{1+u^2+v^2}\sqrt{\left(\frac{du}{dt}\right)^2+\left(\frac{dv}{dt}\right)^2}\,dt. \tag{3.8}$$

Definition 3.20. A *stereographic isometry* of $\mathbb{R}^2_\infty$ is a bijective map $f : \mathbb{R}^2_\infty \to \mathbb{R}^2_\infty$ such that

$$d_\Phi(f(p), f(q)) = d_\Phi(p, q)$$

for any $p, q \in \mathbb{R}^2_\infty$.

Then, the following two theorems immediately arise.

Theorem 3.21. *A map $f : \mathbb{R}^2_\infty \to \mathbb{R}^2_\infty$ is a stereographic isometry if and only if it induces an isometry of the sphere.*

Proof. ($\Rightarrow$) Let $f : \mathbb{R}^2_\infty \to \mathbb{R}^2_\infty$ be a stereographic isometry and $g = \Phi^{-1} \circ f \circ \Phi$. For any points $p, q \in \mathbb{S}^2$,

$$\begin{aligned} d_{\mathbb{S}^2}(g(p), g(q)) &= d_{\mathbb{S}^2}((\Phi^{-1} \circ f \circ \Phi)(p), (\Phi^{-1} \circ f \circ \Phi)(q)) \\ &= d_\Phi((f \circ \Phi)(p), (f \circ \Phi)(q)) \\ &= d_\Phi(\Phi(p), \Phi(q)) \\ &= d_{\mathbb{S}^2}(p, q). \end{aligned}$$

Hence, g is an isometry of $\mathbb{S}^2$.

($\Leftarrow$) For a map $f : \mathbb{R}^2_\infty \to \mathbb{R}^2_\infty$, let $g = \Phi^{-1} \circ f \circ \Phi$ be an isometry of $\mathbb{S}^2$. Then, $f = \Phi \circ g \circ \Phi^{-1}$. For any points $p, q \in \mathbb{R}^2_\infty$,

$$\begin{aligned} d_\Phi(f(p), f(q)) &= d_\Phi((\Phi \circ g \circ \Phi^{-1})(p), (\Phi \circ g \circ \Phi^{-1})(q)) \\ &= d_{\mathbb{S}^2}((g \circ \Phi^{-1})(p), (g \circ \Phi^{-1})(q)) \\ &= d_{\mathbb{S}^2}(\Phi^{-1}(p), \Phi^{-1}(q)) \\ &= d_\Phi(p, q). \end{aligned}$$

Hence, f is a stereographic isometry. □

Let us denote by $\mathrm{Iso}(\mathbb{R}^2_\infty)$ the set of all the stereographic isometries. Then Theorem 3.21 implies that there is a one-to-one correspondence between the sets $\mathrm{Iso}(\mathbb{S}^2)$ and $\mathrm{Iso}(\mathbb{R}^2_\infty)$.

Theorem 3.22. *A stereographic isometry is a composition of at most three inversions.*

Proof. Let ϕ be a stereographic isometry, then, by Theorem 3.21,

$$\Phi^{-1} \circ \phi \circ \Phi : \mathbb{S}^2 \to \mathbb{S}^2$$

is an isometry of $\mathbb{S}^2$. By Theorem 2.7, $\Phi^{-1} \circ \phi \circ \Phi$ is a composition of at most three reflections in great circles:

$$\Phi^{-1} \circ \phi \circ \Phi = \bar{r}_1 \circ \bar{r}_2 \circ \cdots \circ \bar{r}_n,$$

where $\bar{r}_1, \bar{r}_2, \cdots, \bar{r}_n$ are some reflections in great circles with $n \leq 3$. Then

$$\begin{aligned} \phi &= \Phi \circ \bar{r}_1 \circ \bar{r}_2 \circ \cdots \circ \bar{r}_n \circ \Phi^{-1} \\ &= \Phi \circ \bar{r}_1 \circ \Phi^{-1} \circ \Phi \circ \bar{r}_2 \circ \Phi^{-1} \circ \cdots \circ \Phi \circ \bar{r}_n \circ \Phi^{-1} \\ &= I_1 \circ I_2 \circ \cdots \circ I_n, \end{aligned}$$

where $I_i := \Phi \circ \bar{r}_i \circ \Phi^{-1}$ is an inversion on $\mathbb{R}^2_\infty$ by Theorem 3.14.

□

One can define a *stereographic line* on $\mathbb{R}^2_\infty$ as the set of points that have the same stereographic distance from two distinct points. One can show that an image of a great circle obtained by the stereographic projection is a stereographic line. It is not difficult to prove the following theorem.

Theorem 3.23. *A stereographic line is the unit circle centered at the origin or a circline that intersects with the unit circle at two antipodal points of the circle (Figure 3.18).*

Proof. Let Γ be a stereographic line. Then

$$\Gamma = \{p \in \mathbb{R}^2 \mid d_\Phi(p_1, p) = d_\Phi(p_2, p)\}$$

for some fixed distinct points p_1, p_2 in $\mathbb{R}^2_\infty$ and

$$\begin{aligned} p \in \Gamma &\Leftrightarrow d_\Phi(p_1, p) = d_\Phi(p_2, p) \\ &\Leftrightarrow d_{\mathbb{S}^2}(\Phi^{-1}(p_1), \Phi^{-1}(p)) = d_\Phi(\Phi^{-1}(p_2), \Phi^{-1}(p)) \\ &\Leftrightarrow \Phi^{-1}(p) \in G_{\Phi^{-1}(p_1), \Phi^{-1}(p_2)}. \end{aligned}$$

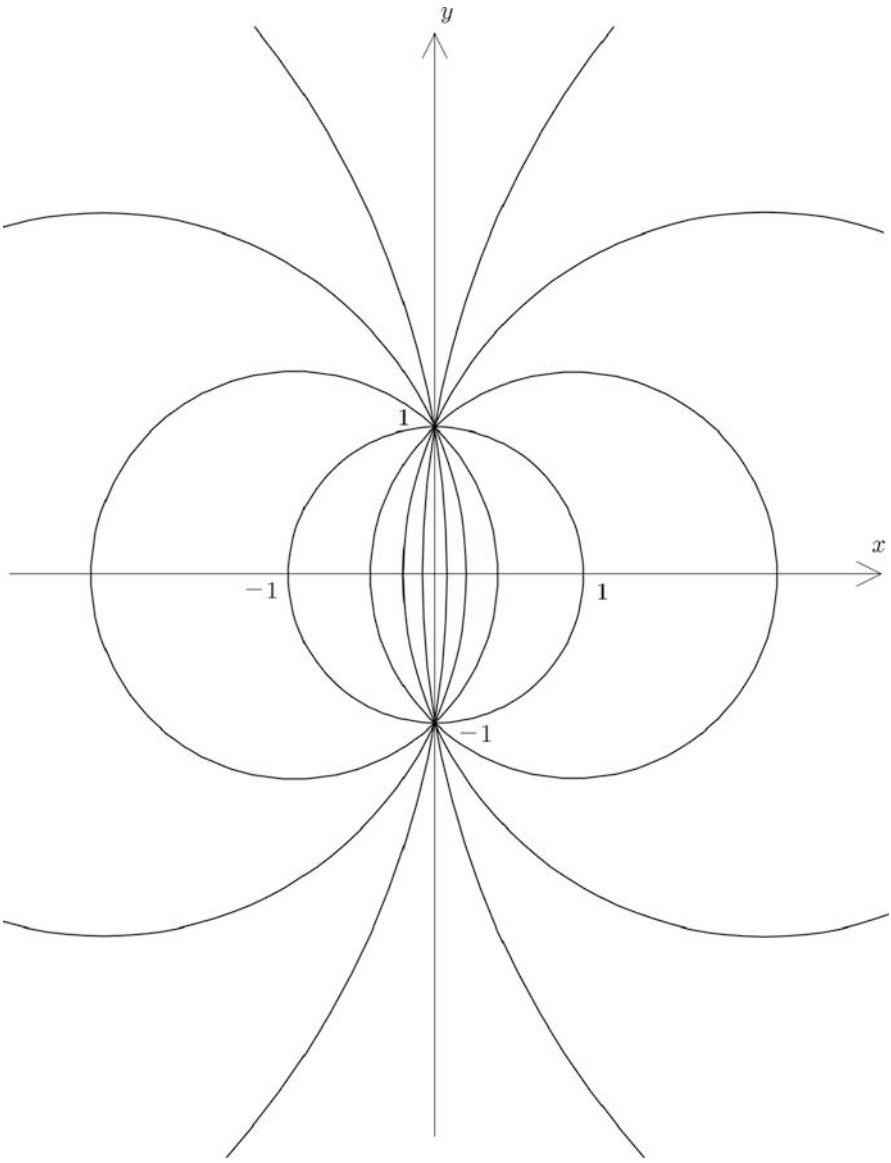


Fig. 3.18 Stereographic lines through points (0, 1) and (0, −1)

Hence,

$$\Phi^{-1}(\Gamma) = G_{\Phi^{-1}(p_1),\Phi^{-1}(p_2)}$$

is a spherical line, which is a great circle. If $\Phi^{-1}(\Gamma)$ is an equator of the sphere, then $\Gamma = \Phi(\Phi^{-1}(\Gamma))$ is the unit circle, centered at the origin. Otherwise, $\Phi^{-1}(\Gamma)$ meets the equator at two antipodal points. Then, $\Gamma = \Phi(\Phi^{-1}(\Gamma))$ is a circline that intersects with the unit circle at two antipodal points of the circle. On the other hand, let Γ be a circline that intersects with the unit circle at two antipodal points of the circle. Note that $\Phi^{-1}(\Gamma)$ is a circle on the sphere that meets the equator at two antipodal points. This means that $\Phi^{-1}(\Gamma)$ is a great circle. Hence, Γ is also a stereographic line in this case, or a circline that intersects with the unit circle at two antipodal points of the circle.

□

Definition 3.24. Let R be a region in $\mathbb{R}^2_\infty$. Then, its *stereographic area* is defined as

$$\text{Area}_\Phi(R) = \text{Area}\left(\Phi^{-1}(R)\right).$$

A stereographic triangle is a region on $\mathbb{R}^2_\infty$ bounded by three stereographic lines.

Theorem 3.25. *The area of a stereographic triangle with interior angles α, β, and γ is*

$$\alpha + \beta + \gamma - \pi.$$

Remark 3.26. Let R be a region in $\mathbb{R}^2_\infty$. Then, its stereographic area can be given as

$$\text{Area}_\Phi(R) = \iint_R \frac{4dudv}{u^2 + v^2 + 1}.$$

Thus far, we have considered a few spaces with distances: $(\mathbb{R}^2, d)$, $(\mathbb{S}^2, d_{\mathbb{S}^2})$ and $(\mathbb{R}^2_\infty, d_\Phi)$. One can regard distance as a function. For example, the distance d on the Euclidean plane $\mathbb{R}^2$ is the function

$$d : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}.$$

A set M with a distance function $d_M : M \times M \to \mathbb{R}$ is called a *(metric) space*; more conditions on d_M can be added, depending on the geometry under study. If there exists a bijective map ϕ from a space M_1 with distance d_{M_1} to another space M_2 with distance d_{M_2} that preserves the distances d_{M_1} and d_{M_1}, i.e.,

$$d_{M_2}(\phi(p), \phi(q)) = d_{M_1}(p, q)$$

for any $p, q \in M_1$, then the map ϕ is called an *isometry* from M_1 to M_2, and the spaces M_1 and M_2 are said to be *isometric*. If M_1 and M_2 are isometric, they share the same geometric properties. By the very definition of d_{Φ} on $\mathbb{R}^2_{\infty}$ (Definition 3.19), the spaces $(\mathbb{S}^2, d_{\mathbb{S}^2})$ and $(\mathbb{R}^2_{\infty}, d_{\Phi})$ are isometric. This is the reason why they satisfy basically the same geometric theorems with respect to lines, circles, and isometries. One can regard $(\mathbb{S}^2, d_{\mathbb{S}^2})$ and $(\mathbb{R}^2_{\infty}, d_{\Phi})$ as different models of the same space. Note that the space $(\mathbb{S}^2, d_{\mathbb{S}^2})$ sits in the three- dimensional Euclidean space. Hence, geometric intuition for three-dimensional space is very useful in studying this model. However, the beings on Sphereland do not have such intuition and so it may be very hard for them to understand this model.

On the other hand, the space $(\mathbb{R}^2_{\infty}, d_{\Phi})$ uses the set $\mathbb{R}^2$ as its base. This model looks awkward and unnatural to us because it represents the geometry of the sphere in $\mathbb{R}^3$ on the set $\mathbb{R}^2$. However, people on Sphereland may be much more comfortable with this model because they have no conception of the three-dimensional space.

We have good geometric intuition for the three-dimensional Euclidean space. It is possible that our universe is not isometric with $\mathbb{R}^3$ but some other space such as the following space in $\mathbb{R}^4$:

$$\{(x, y, z, u) \in \mathbb{R}^4 \mid x^2 + y^2 + z^2 + u^2 = R^2\},$$

where R is some huge number that represents the size of our universe. However, we would be more comfortable with a model that uses the set $\mathbb{R}^3$ as its base just as the beings on Sphereland are more comfortable with the model $(\mathbb{R}^2_{\infty}, d_{\Phi})$.

One can interpret the geometry of Sphereland somewhat differently (originally an idea of Poincaré). Consider a world with inhabitants that can be described by $\mathbb{R}^2$. One of its descriptors is temperature; the absolute temperature at (u, v) is

$$T(u, v) = \frac{1 + u^2 + v^2}{2}.$$

The lengths of objects, including living creatures, are proportional to the absolute temperature. How will a little flat creature endowed with reason living in this world describe the main physical laws of her world? The first question she may ask could be the following:

Is the world finite or infinite?

To answer this question, an expedition is organized; however, as the expedition moves away from the origin, the legs of the explorers become longer, and their steps become bigger, proportional to the temperature $T(u, v)$ (see Figure 3.19). Therefore, the length of their steps is

$$cT(u, v) = c\left(\frac{1 + u^2 + v^2}{2}\right),$$

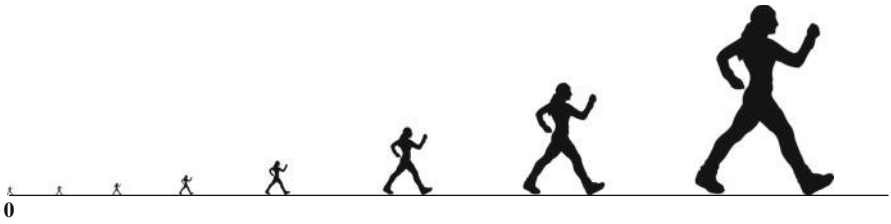


Fig. 3.19 As the legs of the explorers become longer, their steps become bigger

where c is some small positive constant, which depends on the size of their body. Therefore, the length of a curve in her world will be measured as

$$\int_a^b \frac{k}{T(u,v)}\sqrt{\left(\frac{du}{dt}\right)^2+\left(\frac{dv}{dt}\right)^2}\,dt$$

for a curve $\gamma : [a,b] \to \mathbb{R}^2$ with $\gamma(t) = (u(t), v(t))$, where k is some positive scaling constant. If one sets as $k = 1$, this is the formula given in (3.8).

Suppose that $\gamma(t) = (t, 0)$ is the path of the explorers and that they take one step per second. Then, the steps needed to reach the "end" of the world would be as follows:

$$\int_0^\infty \frac{2}{c(1+t^2+0^2)}\sqrt{\left(\frac{dt}{dt}\right)^2+\left(\frac{d0}{dt}\right)^2}\,dt = \int_0^\infty \frac{2}{c(1+t^2)}\,dt = \frac{\pi}{c},$$

which is finite! Hence, they will conclude that the world is finite.

The next question may be the following:

Can they realize that the temperature is varying in the world?

Having constructed a thermometer (based on different expansion coefficients of various materials), they carry it around their world and take measurements. However, since the lengths of all objects change simultaneously with temperature, the thermometer gives the same measurement all over the world. They conclude that the temperature is constant.

They might study straight lines, i.e., investigate the shortest path between two points. The ordinary Euclidean lines are not straight lines in their world. They will discover that the shortest path is a stereographic line, which is the image of a great circle on the stereographic map.

What do you think about our real universe? Do you think that it is really a Euclidean three-dimensional space? To draw a reasonable conclusion, one needs to make measurements on an astronomical scale.

Exercises

3.14. Find

$$d_{\Phi}(p_1, p_2) \text{ and } l_{\Phi}(C),$$

where $p_1 = (0, 0)$, $p_2 = (1, 0)$ and C is a circle having the line segment $\overline{p_1 p_2}$ as a diameter.

3.15. Let T be the stereographic triangle whose vertices are $(0, 0)$, $(0, 1)$, and $(1, 0)$. Find its stereographic area

$$\text{Area}_{\Phi}(T).$$

Chapter 4
Hyperbolic Plane

"One geometry cannot be more true than another; it can only be more convenient."

Henri Poincaré (1854–1912)

"Mathematics is the art of giving the same name to different things."

Henri Poincaré (1854–1912)

Hyperbolic geometry was created in the nineteenth century to better understand Euclid's axiomatic basis for geometry. However, hyperbolic geometry is similar to Euclidean geometry in many respects. It has the concepts of distance and angle, and there are many theorems common to both. However, there are also striking differences, e.g., the sum of the angles of a hyperbolic triangle is always less than π.

4.1 Poincaré Upper Half-Plane $\mathbb{H}^2$

The upper half-plane is the set

$$\mathbb{H}^2 = \{(x, y) \in \mathbb{R}^2 \mid y > 0\}.$$

In Section 3.4, we introduced the space $\mathbb{R}^2_\infty$, whose "lines" are the unit circle, centered at the origin, or circlines that intersect with the unit circle at two antipodal points of the unit circle (Theorem 3.23). This distortion of "lines" is caused because we described the geometry of the sphere, which is a curved surface, on a flat plane.

N.-H. Lee, *Geometry: from Isometries to Special Relativity*, Undergraduate Texts in Mathematics, https://doi.org/10.1007/978-3-030-42101-4_4

In this chapter, we introduce another interesting geometry of a certain curved surface. Different from the sphere, this surface cannot sit in $\mathbb{R}^3$; however, it is one of the very basic surfaces, along with the Euclidean plane and the sphere. We will describe its geometry on the upper half-plane $\mathbb{H}^2$. Its "lines" will turn out to be semicircles or vertical half lines that intersect orthogonally with the x-axis (Figure 4.1). Similar to the process described in Section 3.4, we can also imagine that some variation of temperature results in the geometry on $\mathbb{H}^2$. Note that the lengths of objects, including living creatures, are proportional to the absolute temperature. Then, the question would be

"What temperature variation would result in a certain geometry?"

Note that a "line" through given points p and q is also the "shortest" path from p to q, where the "distance" is measured in a way determined by its geometry. Because the "shortest" path on $\mathbb{H}^2$ is not a Euclidean line but a semicircle (Figure 4.2), the temperature is not constant. We can expect that the temperature increases as the y-coordinate increases, i.e., as the y-coordinate increases, the legs of the creatures become proportionally larger, and thus, fewer steps are required to reach point q from point p along the circle than along the Euclidean line (Figure 4.3). The simplest choice of an appropriate temperature function would be as follows:

$$T(x, y) = y.$$

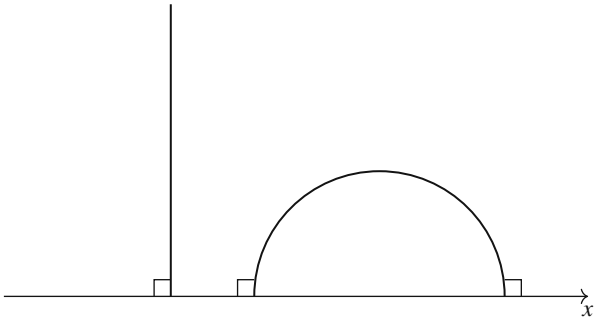


Fig. 4.1 "Lines" on the upper half-plane $\mathbb{H}^2$

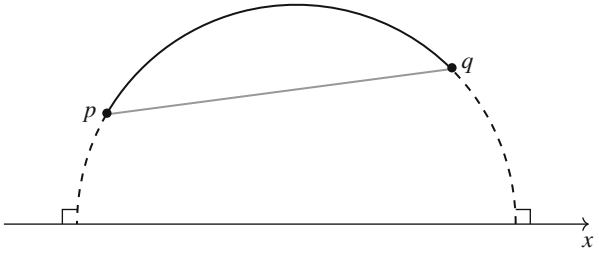


Fig. 4.2 The "shortest" path from p to q

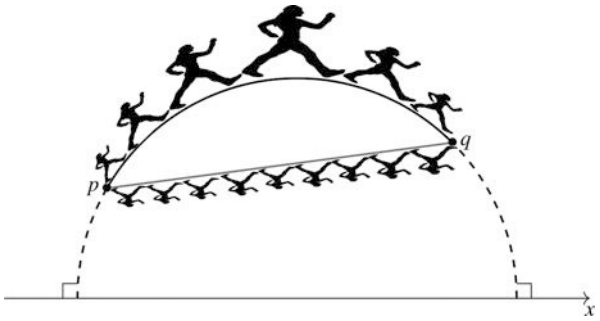


Fig. 4.3 This distance is 9 steps along the Euclidean line but only 6 steps along the semicircle!

Surprisingly, this naive choice of temperate function yields the correct geometry. Note that the temperature (also the size of the creatures) approaches zero as it approaches the x-axis.

Given this temperature function, we define the $\mathbb{H}^2$-length of a curve on $\mathbb{H}^2$ as follows:

Definition 4.1. Let $\gamma : [a, b] \to \mathbb{H}^2$ be a smooth curve on the half-plane, where $\gamma(t) = (x(t), y(t))$; then, its *hyperbolic length* is as follows:

$$l_{\mathbb{H}^2}(\gamma) = \int_a^b \frac{1}{T(x, y)} \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2}\, dt = \int_a^b \frac{1}{y} \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2}\, dt.$$

For example, consider the following:

$$\gamma_1 : [a, b] \to \mathbb{H}^2, \gamma_1(t) = (0, t), 0 < a < b.$$

Then, γ_1 is a part of a vertical line, and its hyperbolic length is

$$l_{\mathbb{H}^2}(\gamma_1) = \int_a^b \frac{\sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2}}{y}\, dt = \int_a^b \frac{1}{t}\, dt = \ln \frac{b}{a}.$$

Let

$$\gamma_2 : [a, b] \to \mathbb{H}^2, \gamma_2(t) = (t, c), 0 < a < b, 0 < c.$$

Then, it is a part of a horizontal line, and its hyperbolic length is

$$l_{\mathbb{H}^2}(\gamma_2) = \int_a^b \frac{\sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2}}{y}\, dt = \int_a^b \frac{1}{c}\, dt = \frac{b - a}{c},$$

which is inversely proportional to c.

The set $\mathbb{H}^2$ with the hyperbolic length is called the *hyperbolic plane*.

Definition 4.2. A bijective map $\phi : \mathbb{H}^2 \to \mathbb{H}^2$ is an *isometry* of the hyperbolic plane if it preserves the hyperbolic lengths of the curves, i.e.,

$$l_{\mathbb{H}^2}(\phi(\gamma)) = l_{\mathbb{H}^2}(\gamma)$$

for each smooth curve γ on $\mathbb{H}^2$.

Denote the set of all isometries of the hyperbolic plane by $\mathrm{Iso}(\mathbb{H}^2)$.

For a map

$$f : \mathbb{R}^2 \to \mathbb{R}^2,$$

if $f(\mathbb{H}^2) \subset \mathbb{H}^2$, we consider it a map from $\mathbb{H}^2$ to $\mathbb{H}^2$.

For a map $g : \mathbb{H}^2 \to \mathbb{H}^2$ and a curve $\gamma : [a, b] \to \mathbb{H}^2$ with $\gamma(t) = (x(t), y(t))$, recall that the curve $\delta := g(\gamma) : [a, b] \to \mathbb{H}^2$ is defined by

$$\delta(t) = g(\gamma(t)).$$

Example 4.1 (Translations in the x-direction). For $a \in \mathbb{R}$, $t_{(a,0)}(x, y) = (a + x, y)$.

$$\delta(t) = \left(t_{(a,0)}(\gamma)\right)(t) = (x(t) + a, y(t)).$$

$$\begin{aligned} l_{\mathbb{H}^2}\left(t_{(a,0)}(\gamma)\right) &= l_{\mathbb{H}^2}(\delta) \\ &= \int_a^b \frac{\sqrt{\left(\frac{d(x+a)}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2}}{y}\, dt \\ &= \int_a^b \frac{\sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2}}{y}\, dt \\ &= l_{\mathbb{H}^2}(\gamma). \end{aligned}$$

Hence, $t_{(a,0)}$ is an isometry of the hyperbolic plane.

Example 4.2 (Reflections in vertical lines). If L is a line defined by $x = a$, then $\bar{r}_L(x, y) = (2a - x, y)$, and

$$\delta(t) = (\bar{r}_L(\gamma))(t) = (2a - x(t), y(t)).$$

$$
\begin{aligned}
l_{\mathbb{H}^2}\left(\bar{r}_L(\gamma)\right) &= \int_a^b \frac{\sqrt{\left(\frac{d(2a-x)}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2}}{y}\,dt \\
&= \int_a^b \frac{\sqrt{\left(-\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2}}{y}\,dt \\
&= \int_a^b \frac{\sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2}}{y}\,dt \\
&= l_{\mathbb{H}^2}(\gamma).
\end{aligned}
$$

Hence, $\bar{r}_L$ is an isometry of the hyperbolic plane.

Example 4.3 (Rescaling about the origin $\mathbf{0}$*).* Let $d_r(x, y) = (rx, ry)$ for some $r > 0$.

$$
\delta(t) = \left(d_r(\gamma)\right)(t) = (rx(t), ry(t)).
$$

$$
\begin{aligned}
l_{\mathbb{H}^2}\left(d_r(\gamma)\right) &= \int_a^b \frac{\sqrt{\left(\frac{d(rx)}{dt}\right)^2 + \left(\frac{d(ry)}{dt}\right)^2}}{ry}\,dt \\
&= \int_a^b \frac{r\sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2}}{ry}\,dt \\
&= \int_a^b \frac{\sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2}}{y}\,dt \\
&= l_{\mathbb{H}^2}(\gamma).
\end{aligned}
$$

Hence, d_r is an isometry of the hyperbolic plane.

Proposition 4.3. *The inversion I is an isometry of the hyperbolic plane.*

Proof. Note that

$$I(x, y) = \frac{1}{x^2 + y^2}(x, y).$$

$$\left(\frac{d}{dt}\frac{x}{x^2+y^2}\right)^2 + \left(\frac{d}{dt}\frac{y}{x^2+y^2}\right)^2 = \frac{1}{(x^2+y^2)^2}\left(\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2\right).$$

$$\begin{aligned} l_{\mathbb{H}^2}(I(\gamma)) &= \int_a^b \frac{\sqrt{\left(\frac{d}{dt}\frac{x}{x^2+y^2}\right)^2 + \left(\frac{d}{dt}\frac{y}{x^2+y^2}\right)^2}}{\frac{y}{x^2+y^2}}\, dt \\ &= \int_a^b \frac{1}{x^2+y^2}\frac{\sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2}}{\frac{y}{x^2+y^2}}\, dt \\ &= \int_a^b \frac{\sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2}}{y}\, dt \\ &= l_{\mathbb{H}^2}(\gamma). \end{aligned}$$

Hence, I is an isometry. □

Theorem 4.4. *If C is a circle, centered at a point on the x-axis, or a vertical line, then the inversion I_C is an isometry of the hyperbolic plane.*

Proof. When C is a vertical line, this was proved in Example 4.2. If $C = C_{\mathbf{0},r}$ is a circle with radius r, centered at the origin, then

$$I_C = d_r \circ I \circ d_{1/r},$$

which is a composition of isometries. Therefore, it is an isometry. In general, if $C = C_{(a,0),r}$ is a circle of radius r, centered at $(a, 0)$, then

$$I_C = t_{(a,0)} \circ I_{C_{\mathbf{0},r}} \circ t_{(-a,0)},$$

which is again a composition of isometries. Hence, I_C is also an isometry. □

Exercises

4.1. Let L be a line segment connecting the points $(1, \sqrt{3})$ and $(0, 2)$, and let C be part of a Euclidean circle with radius 2 and center $\mathbf{0}$ connecting these two points. Show numerically that

$$l_{\mathbb{H}^2}(L) > l_{\mathbb{H}^2}(C).$$

You may use the formula

$$\int \frac{dx}{\sin x} = \ln \left| \tan \frac{x}{2} \right|$$

and a calculator.

4.2. Let C be a curve joining the origin and the point $(0, 1)$. Show that it does not have a finite hyperbolic length. What does this mean?

4.3. For each pair of points $p, q \in \mathbb{H}^2$, show that there is an isometry ϕ of the hyperbolic plane such that $\phi(p) = q$

4.2 $\mathbb{H}^2$-Shortest Paths and $\mathbb{H}^2$-Lines

Henceforth, an inversion means an inversion in a vertical line, or a circle whose center lies on the x-axis, and a translation means a translation in the x-direction. Hence, an inversion and a translation are isometries of the hyperbolic plane.

On the Euclidean plane, a line segment measures the shortest distance between two points. We develop a similar notion, corresponding to that of the line segment on the Euclidean plane. For two distinct points p_1 and p_2, a path from p_1 to p_2 is a smooth curve $\gamma : [a, b] \to \mathbb{H}^2$ such that

$$\gamma(a) = p_1, \gamma(b) = p_2.$$

Definition 4.5. A path γ from a point p_1 to another point p_2 is called an $\mathbb{H}^2$-*shortest* path from p_1 to p_2 if

$$l_{\mathbb{H}^2}(\gamma) \leq l_{\mathbb{H}^2}(\gamma')$$

for every path γ' from p_1 to p_2.

Proposition 4.6. *The curve*

$$\gamma : [a, b] \to \mathbb{H}^2, \gamma(t) = (0, t), 0 < a < b,$$

is the unique $\mathbb{H}^2$*-shortest path from* $(0, a)$ *to* $(0, b)$.

Proof. Note that $l_{\mathbb{H}^2}(\gamma) = \ln \frac{b}{a}$. Let $\delta : [c, d] \to \mathbb{H}^2$ be any path from $(0, a)$ to $(0, b)$ with $\delta(t) = (x(t), y(t))$. Then, $y(c) = a$, $y(d) = b$, and

$$l_{\mathbb{H}^2}(\delta) = \int_c^d \frac{\sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2}}{y} dt \geq \int_c^d \frac{\sqrt{\left(\frac{dy}{dt}\right)^2}}{y} dt$$

$$= \int_c^d \frac{1}{y} \left|\frac{dy}{dt}\right| dt \geq \int_{y(c)}^{y(d)} \frac{dy}{y} = \ln \frac{y(d)}{y(c)} = \ln \frac{b}{a} = l_{\mathbb{H}^2}(\gamma).$$

In summary, $l_{\mathbb{H}^2}(\delta) \geq l_{\mathbb{H}^2}(\gamma)$, and the equality holds only if

$$\frac{dx}{dt} = 0 \text{ and } \frac{dy}{dt} \geq 0,$$

which implies that δ and γ are the same curves (Figure 4.4). Thus ends the proof. □

Proposition 4.7. *Let ϕ be an isometry of the hyperbolic plane and γ be a path from p_1 to p_2. Then, γ is an $\mathbb{H}^2$-shortest path from p_1 to p_2 if and only if $\phi(\gamma)$ is an $\mathbb{H}^2$-shortest path from $\phi(p_1)$ to $\phi(p_2)$.*

Proof. First, let γ be an $\mathbb{H}^2$-shortest path from p_1 to p_2 and $\delta = \phi(\gamma)$. Let δ' be a path from $\phi(p_1)$ to $\phi(p_2)$. Then, $\gamma' := \phi^{-1}(\delta')$ is a path from p_1 to p_2. Therefore,

$$l_{\mathbb{H}^2}(\gamma) \leq l_{\mathbb{H}^2}(\gamma') = l_{\mathbb{H}^2}\left(\phi^{-1}(\delta')\right) = l_{\mathbb{H}^2}\left(\delta'\right)$$

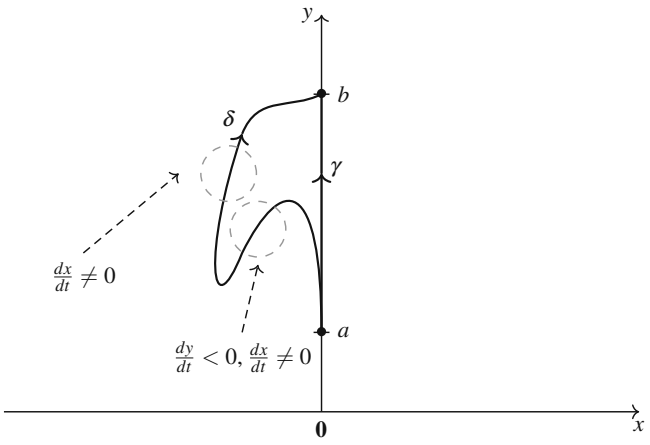


Fig. 4.4 γ is the $\mathbb{H}^2$-shortest path from $(a, 0)$ to $(b, 0)$

because ϕ^{-1} is also an isometry. Hence,

$$l_{\mathbb{H}^2}(\delta) = l_{\mathbb{H}^2}(\phi(\gamma)) = l_{\mathbb{H}^2}(\gamma) \leq l_{\mathbb{H}^2}\left(\delta'\right).$$

Therefore, $\delta = \phi(\gamma)$ is an $\mathbb{H}^2$-shortest path from $\phi(p_1)$ to $\phi(p_2)$.

Second, if we let $\psi = \phi^{-1}$, $q_1 = \phi(p_1)$, and $q_2 = \phi(p_2)$ and apply the same argument, then the converse is also proven. □

Lemma 4.8. *If Γ is either a vertical line or a circle, centered at a point on the x-axis, then there is an isometry ϕ such that $\phi(\Gamma)$ is the y-axis.*

Proof. If Γ is a vertical line, there is some isometry ϕ (for example, a translation in the x-direction) such that $\phi(\Gamma)$ is the y-axis.

Otherwise Γ would be a circle, centered at a point on the x-axis. Let q be one of the intersection points of the circle Γ and the x-axis. Let C be a circle of radius 1 with center q. Then, $I_C(\Gamma)$ is a line. Since the circle Γ intersects orthogonally with the x-axis, I_C sends the x-axis to the x-axis, and an inversion preserves angles, the line meets orthogonally with the x-axis, i.e., it is a vertical line (Figure 4.5). By applying a translation in the direction of the x-axis, one can move the line to the y-axis. The composition ϕ of the inversion and the translation maps the circle Γ to the y-axis. Note that ϕ is an isometry. □

Theorem 4.9. *For any two points $p_1, p_2 \in \mathbb{H}^2$, there exists a unique $\mathbb{H}^2$-shortest path from p_1 to p_2, and it is a part of a vertical line or a circle, centered at a point on the x-axis.*

Proof. Let γ be a path from p_1 to p_2 that is a part of a vertical line or a part of a circle, centered at a point on the x-axis.

According to Lemma 4.8, there is an isometry ϕ such that $\phi(\gamma)$ is a part of the y-axis. According to Proposition 4.6, $\phi(\gamma)$ is an $\mathbb{H}^2$-shortest path from $\phi(p_1)$ to $\phi(p_2)$, and according to Proposition 4.7,

$$\phi^{-1}(\phi(\gamma)) = \gamma$$

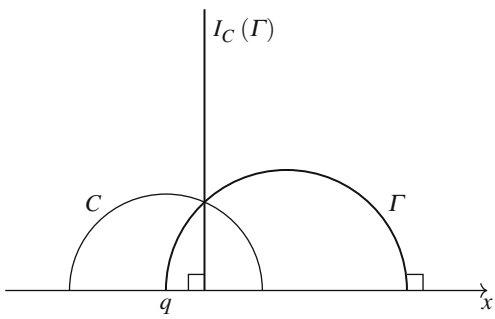


Fig. 4.5 $\phi(\Gamma)$ is a vertical line

is an $\mathbb{H}^2$-shortest path from p_1 to p_2. Suppose that there is another $\mathbb{H}^2$-shortest path δ from p_1 to p_2 that is different from γ. Then, $\phi(\delta)$ is another $\mathbb{H}^2$-shortest path from $\phi(p_1)$ to $\phi(p_2)$ that is different from $\phi(\gamma)$, which is contradictory to Proposition 4.6. Hence, γ is the unique $\mathbb{H}^2$-shortest path from p_1 to p_2. □

We can now define the hyperbolic distance between points on $\mathbb{H}^2$.

Definition 4.10. For $p_1, p_2 \in \mathbb{H}^2$, the *hyperbolic distance* between p_1 and p_2 is as follows:

$$d_{\mathbb{H}^2}(p_1, p_2) = l_{\mathbb{H}^2}(\gamma),$$

where γ is the $\mathbb{H}^2$-shortest path from p_1 to p_2.

For $p_1 = (a, b)$ and $p_2 = (a, c)$, which lie on the same vertical line,

$$d_{\mathbb{H}^2}(p_1, p_2) = \left| \ln \frac{c}{b} \right|.$$

According to Theorem 4.9, we can formulate a triangle inequality for the hyperbolic plane.

Corollary 4.11 (Triangle inequality for the hyperbolic plane). *For three points p_1, p_2, and p_3 in the hyperbolic plane,*

$$d_{\mathbb{H}^2}(p_1, p_3) \leq d_{\mathbb{H}^2}(p_1, p_2) + d_{\mathbb{H}^2}(p_2, p_3),$$

where the equality holds if and only if p_2 lies on the $\mathbb{H}^2$-shortest path from p_1 to p_3.

Proof. Let γ_{ij} be the $\mathbb{H}^2$-shortest path from p_i to p_j for $1 \leq i < j \leq 3$. Then,

$$d_{\mathbb{H}^2}(p_i, p_j) = l_{\mathbb{H}^2}(\gamma_{ij}).$$

Note that the union (denoted by δ) of γ_{12} and γ_{23} is a path from p_1 to p_3 (Figure 4.6). According to Theorem 4.9,

$$d_{\mathbb{H}^2}(p_1, p_2) + d_{\mathbb{H}^2}(p_2, p_3) = l_{\mathbb{H}^2}(\delta) \geq l_{\mathbb{H}^2}(\gamma_{13}) = d_{\mathbb{H}^2}(p_1, p_3),$$

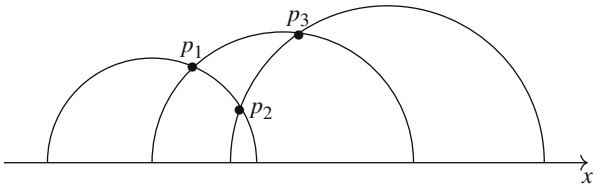


Fig. 4.6 $d_{\mathbb{H}^2}(p_1, p_3) \leq d_{\mathbb{H}^2}(p_1, p_2) + d_{\mathbb{H}^2}(p_2, p_3)$

and the equality holds if and only if δ has the same path as γ_{13}, which means that point p_2 lies on path γ_{13}. □

Corollary 4.12. *If $d_{\mathbb{H}^2}(p_1, p_2) = 0$, then $p_1 = p_2$.*

Proof. According to Lemma 4.8, there is an isometry ϕ such that $\phi(p_1)$ and $\phi(p_2)$ lie on the y-axis, i.e.,

$$\phi(p_1) = (0, a), \phi(p_2) = (0, b) \text{ for some } a, b > 0.$$

According to Proposition 4.6, the curve $\gamma : [0, 1] \to \mathbb{H}^2$ by $\gamma(t) = (0, at+b(1-t))$ is the $\mathbb{H}^2$-shortest path from $\phi(p_1)$ to $\phi(p_2)$. Hence,

$$0 = d_{\mathbb{H}^2}(p_1, p_2) = d_{\mathbb{H}^2}(\phi(p_1), \phi(p_2)) = l_{\mathbb{H}^2}(\gamma) = \left|\ln \frac{b}{a}\right|.$$

Therefore, $a = b$ and $\phi(p_1) = \phi(p_2)$. Since an isometry is injective, $p_1 = p_2$. □

Definition 4.13. For two distinct points $p_1, p_2 \in \mathbb{H}^2$, the set of points equi-$\mathbb{H}^2$-distant from p_1 and p_2 is called a *hyperbolic line* (or simply an $\mathbb{H}^2$-line), i.e.,

$$H_{p_1,p_2} = \{p \in \mathbb{H}^2 \mid d_{\mathbb{H}^2}(p, p_1) = d_{\mathbb{H}^2}(p, p_2)\}.$$

Proposition 4.14. *Let $p_1 = (-a, b)$ and $p_2 = (a, b)$ for $a, b > 0$. Then, H_{p_1,p_2} is the y-axis.*

Proof. Let L be the y-axis; then, $\bar{r}_L(p_1) = p_2$. For each point $p \in L$, $\bar{r}_L(p) = p$, and hence,

$$d_{\mathbb{H}^2}(p_1, p) = d_{\mathbb{H}^2}(\bar{r}_L(p_1), \bar{r}_L(p)) = d_{\mathbb{H}^2}(p_2, p).$$

Therefore, $p \in H_{p_1,p_2}$ and $L \subset H_{p_1,p_2}$.

Suppose that there exists some point $q \in H_{p_1,p_2}$ such that $q \notin L$. Let $q = (c, d)$; then, $c \neq 0$. Hence, we have that $c > 0$ or $c < 0$, say $c > 0$, then the $\mathbb{H}^2$-shortest path from p_1 to q intersects with the y-axis at a point p' (Figure 4.7). Let $q' = (-c, d) = \bar{r}_L(q)$. It is not difficult to see that the point p' does not lie on the $\mathbb{H}^2$-shortest path from p_1 to q'. Therefore, by Corollary 4.11,

$$d_{\mathbb{H}^2}(p_1, q') < d_{\mathbb{H}^2}(p_1, p') + d_{\mathbb{H}^2}(p', q').$$

However,

$$\begin{aligned} d_{\mathbb{H}^2}(p_1, q') &= d_{\mathbb{H}^2}\left(\bar{r}_L(p_1), \bar{r}_L(q')\right) \\ &= d_{\mathbb{H}^2}(p_2, q) \\ &= d_{\mathbb{H}^2}(p_1, q) \qquad (\because \quad q \in H_{p_1,p_2}) \end{aligned}$$

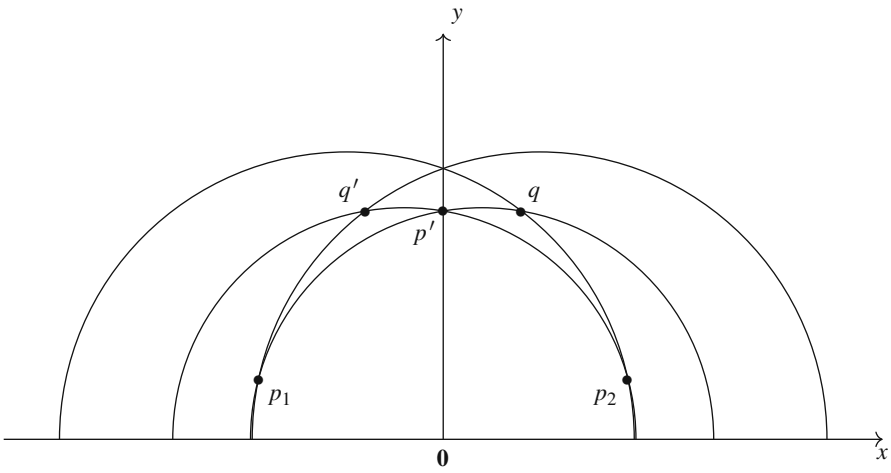


Fig. 4.7 The $\mathbb{H}^2$-shortest path from p_1 to q intersects with the y-axis at a point p'

$$
\begin{aligned}
&= d_{\mathbb{H}^2}(p_1, p') + d_{\mathbb{H}^2}(p', q) \\
&= d_{\mathbb{H}^2}(p_1, p') + d_{\mathbb{H}^2}\left(\bar{r}_L(p'), \bar{r}_L(q)\right) \\
&= d_{\mathbb{H}^2}(p_1, p') + d_{\mathbb{H}^2}(p', q'),
\end{aligned}
$$

i.e.,

$$d_{\mathbb{H}^2}(p_1, q') = d_{\mathbb{H}^2}(p_1, p') + d_{\mathbb{H}^2}(p', q'),$$

which is a contradiction. Hence, if $q \in H_{p_1,p_2}$, then $q \in L$ and $H_{p_1,p_2} \subset L$. Therefore, $H_{p_1,p_2} = L$. □

Lemma 4.15. *For two points $p_1, p_2 \in \mathbb{H}^2$, there is a composition ϕ of reflections in a circle centered at the x-axis and translations in the x-direction such that*

$$\phi(p_1) = (a, b), \quad \phi(p_2) = (-a, b)$$

for some $a, b \in \mathbb{R}$, i.e., two points $\phi(p_1)$ and $\phi(p_2)$ are symmetric with respect to the y-axis.

Proof. For a point p, let $\pi_x(p)$ and $\pi_y(p)$ be its x- and y-coordinates, respectively. If the points p_1 and p_2 lie on a circle Γ, centered at a point on the x-axis, then let q_1 and q_2 be the intersection points of Γ with the x-axis. Consider a unit circle $C_{\alpha,1}$, where the center α lies on the x-axis. The inversion $I_{C_{\alpha,1}}$ maps p_1 and p_2 to points on a vertical line when $\alpha = q_1$ or q_2, respectively. Note that the sign of

$$\pi_y\left(I_{C_{\alpha,1}}(p_1)\right) - \pi_y\left(I_{C_{\alpha,1}}(p_2)\right)$$

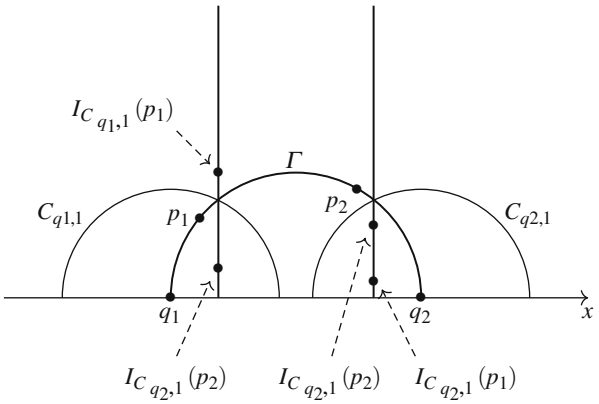


Fig. 4.8 Compare the signs of $\pi_y\left(I_{C_{\alpha,1}}(p_1)\right) - \pi_y\left(I_{C_{\alpha,1}}(p_2)\right)$ for $\alpha = q_1, q_2$

changes as α moves between q_1 and q_2 along the x-axis (Figure 4.8). Therefore, there exists some point α_0 between q_1 and q_2 such that

$$\pi_y\left(I_{C_{\alpha_0,1}}(p_1)\right) - \pi_y\left(I_{C_{\alpha_0,1}}(p_2)\right) = 0.$$

There is some translation t in the x-direction such that

$$\pi_x\left(t\left(I_{C_{\alpha_0,1}}(p_1)\right)\right) + \pi_x\left(t\left(I_{C_{\alpha_0,1}}(p_2)\right)\right) = 0.$$

We still have

$$\pi_y\left(t\left(I_{C_{\alpha_0,1}}(p_1)\right)\right) - \pi_y\left(t\left(I_{C_{\alpha_0,1}}(p_2)\right)\right) = 0.$$

Therefore, $t \circ I_{C_{\alpha_0,1}}$ is a map that we seek.

If the points p_1 and p_2 do not lie on a circle, centered at a point on the x-axis, then they lie on a vertical line. There is some inversion I_1 that maps the line to a circle. Find α_0 and a translation t for $I_1(p_1)$ and $I_1(p_2)$, as we did in the previous case. Now the map

$$t \circ I_{C_{\alpha_0,1}} \circ I_1$$

is a map that we are looking for. □

Theorem 4.16. *A hyperbolic line is either a vertical line or a circle, centered at a point on the x-axis.*

Proof. Let H_{p_1,p_2} be a hyperbolic line. According to Lemma 4.15, there is an isometry ϕ that is a composition of inversions and translations such that

$$\phi(p_1) = (a, b), \quad \phi(p_2) = (-a, b)$$

for some $a, b \in \mathbb{R}$. According to Proposition 4.14, $H_{\phi(p_1),\phi(p_2)}$ is the y-axis. Therefore,

$$H_{p_1,p_2} = \phi^{-1}(H_{\phi(p_1),\phi(p_2)})$$

is either a vertical line or a circle that intersects orthogonally with the x-axis because $H_{\phi(p_1),\phi(p_2)}$ does. Hence, the proof is complete. □

Exercises

4.4. For two points $p_1, p_2 \in \mathbb{H}^2$, show that

$$d_{\mathbb{H}^2}(p_1, p_2) = 2\tanh^{-1}\left(\frac{d(p_1, p_2)}{d(p_1, \bar{r}(p_2))}\right)$$

for each of following cases, where $d(p_1, p_2)$ is the ordinary Euclidean distance.

1. The points p_1, p_2 are on the same vertical line.
2. The points p_1, p_2 are on a Euclidean circle of Euclidean radius $\frac{1}{2}$, centered at the point $\left(\frac{1}{2}, 0\right)$.

   (*Hint.* In the polar coordinate, $r(\theta) = \cos\theta$ parameterize the circle. You may want to use the formula

   $$\int \frac{2}{\sin 2\theta} d\theta = \ln|\tan\theta|\,.$$

)
3. The points p_1, p_2 lie on the hyperbolic plane. (*Hint.* Use 1, 2 and consider suitable isometries.)

4.3 Isometries of the Hyperbolic Plane

We want to use $I_{H_{p_1,p_2}}$ as "reflections" of the hyperbolic plane. Then, the following should hold.

Lemma 4.17. *For a hyperbolic line H_{p_1,p_2},*

$$I_{H_{p_1,p_2}}(p_1) = p_2.$$

Proof. Let Γ be either a vertical line or a circle centered at the x-axis, passing through the points p_1 and p_2. According to Lemma 4.8, there is an isometry ϕ such that $\phi(\Gamma)$ is the y-axis. It is not difficult to see that

$$\phi\left(H_{p_1,p_2}\right) = H_{\phi(p_1),\phi(p_2)}.$$

By symmetry, $H_{\phi(p_1),\phi(p_2)}$ is a circle, centered at the origin $\mathbf{0}$. Then, we can choose some points q_1 and q_2 from $H_{\phi(p_1),\phi(p_2)}$ such that $q_1 = (a, b)$ and $q_2 = (-a, b)$ for some $a, b \in \mathbb{R}$. Note that

$$H_{q_1,q_2} = \phi(\Gamma)$$

and

$$\phi^{-1}(q_1), \phi^{-1}(q_2) \in H_{p_1,p_2}.$$

It is also easy to see that

$$H_{\phi^{-1}(q_1),\phi^{-1}(q_2)} = \phi^{-1}\left(\phi(\Gamma)\right) = \Gamma.$$

Let $p = I_{H_{p_1,p_2}}(p_1)$. Then $p \neq p_1$. Note that

$$\begin{aligned}
d_{\mathbb{H}^2}\left(\phi^{-1}(q_1), p\right) &= d_{\mathbb{H}^2}\left(I_{H_{p_1,p_2}}\left(\phi^{-1}(q_1)\right), I_{H_{p_1,p_2}}(p)\right) \\
&= d_{\mathbb{H}^2}\left(\phi^{-1}(q_1), p_1\right) && (\because \phi^{-1}(q_1) \in H_{p_1,p_2}) \\
&= d_{\mathbb{H}^2}\left(q_1, \phi(p_1)\right) \\
&= d_{\mathbb{H}^2}\left(q_2, \phi(p_1)\right) \\
&= d_{\mathbb{H}^2}\left(\phi^{-1}(q_2), p_1\right) \\
&= d_{\mathbb{H}^2}\left(I_{H_{p_1,p_2}}\left(\phi^{-1}(q_2)\right), I_{H_{p_1,p_2}}(p_1)\right) \\
&= d_{\mathbb{H}^2}\left(\phi^{-1}(q_2), p\right).
\end{aligned}$$

Hence, $p \in H_{\phi^{-1}(q_1),\phi^{-1}(q_2)} = \Gamma$, and thus, $\phi(p)$ lies on the y-axis. Let q be the intersection point of the y-axis and $H_{\phi(p_1),\phi(p_2)}$. Then, $\phi^{-1}(q)$ is the intersection point of Γ and H_{p_1,p_2}, and

$$d_{\mathbb{H}^2}\left(q, \phi(p)\right) = d_{\mathbb{H}^2}\left(q, \phi(p_1)\right) = d_{\mathbb{H}^2}\left(q, \phi(p_2)\right)$$

because

$$\begin{aligned}
d_{\mathbb{H}^2}\left(q, \phi(p)\right) &= d_{\mathbb{H}^2}\left(\phi^{-1}(q), p\right) \\
&= d_{\mathbb{H}^2}\left(I_{H_{p_1,p_2}}\left(\phi^{-1}(q)\right), I_{H_{p_1,p_2}}(p)\right)
\end{aligned}$$

$$= d_{\mathbb{H}^2}\left(\phi^{-1}(q),\, p_1\right)$$
$$= d_{\mathbb{H}^2}\left(q, \phi(p_1)\right).$$

Since the points $\phi(p)$, $\phi(p_1)$, and $\phi(p_2)$ lie on the y-axis, we conclude that $\phi(p) = \phi(p_1)$ or $\phi(p_2)$. Note that $p \neq p_1$. Hence, we have $\phi(p) = \phi(p_2)$. Finally,

$$I_{H_{p_1,p_2}}(p_1) = p = p_2.$$

□

Points on the hyperbolic plane are said to be *collinear* if they lie on the same hyperbolic line.

Lemma 4.18. *An isometry of the hyperbolic plane maps non-collinear points p_1, p_2, and p_3 to non-collinear points.*

Proof. Let $\phi : \mathbb{H}^2 \to \mathbb{H}^2$ be an isometry. Suppose that the points $\phi(p_1)$, $\phi(p_2)$ and $\phi(p_3)$ are collinear. Then, they lie on an $\mathbb{H}^2$-line $H_{p,q}$. Therefore,

$$d_{\mathbb{H}^2}(p, \phi(p_1)) = d_{\mathbb{H}^2}(q, \phi(p_1)),$$

$$d_{\mathbb{H}^2}(p, \phi(p_2)) = d_{\mathbb{H}^2}(q, \phi(p_2)),$$

and

$$d_{\mathbb{H}^2}(p, \phi(p_3)) = d_{\mathbb{H}^2}(q, \phi(p_3)).$$

However, then,

$$d_{\mathbb{H}^2}(\phi^{-1}(p),\, p_1) = d_{\mathbb{H}^2}(\phi^{-1}(q),\, p_1),$$

$$d_{\mathbb{H}^2}(\phi^{-1}(p),\, p_2) = d_{\mathbb{H}^2}(\phi^{-1}(q),\, p_2),$$

and

$$d_{\mathbb{H}^2}(\phi^{-1}(p),\, p_3) = d_{\mathbb{H}^2}(\phi^{-1}(q),\, p_3),$$

which implies that the points p_1, p_2, and p_3 lie on an $\mathbb{H}^2$-line $H_{\phi^{-1}(p),\phi^{-1}(q)}$. This is a contradiction, and the points $\phi(p_1)$, $\phi(p_2)$, and $\phi(p_3)$ are non-collinear. □

Theorem 4.19 (Three points theorem for the hyperbolic plane). *If two isometries ϕ and ψ coincide at three non-collinear points p_1, p_2, and p_3, then $\phi = \psi$.*

Proof. According to Lemma 4.18, the points $\phi(p_1)$, $\phi(p_2)$, and $\phi(p_3)$ are also non-collinear.

Suppose that $\phi \neq \psi$. Then, there is a point p such that

$$\phi(p) \neq \psi(p),$$

and we can define an $\mathbb{H}^2$-line $H = H_{\phi(p),\psi(p)}$. Note that

$$\begin{aligned} d_{\mathbb{H}^2}(\phi(p), \phi(p_1)) &= d_{\mathbb{H}^2}(p, p_1) && (\because \ \phi \text{ is an isometry}) \\ &= d_{\mathbb{H}^2}(\psi(p), \psi(p_1)) && (\because \ \psi \text{ is an isometry}) \\ &= d_{\mathbb{H}^2}(\psi(p), \phi(p_1)), \end{aligned}$$

i.e., $d_{\mathbb{H}^2}(\phi(p), \phi(p_1)) = d_{\mathbb{H}^2}(\psi(p), \phi(p_1))$. Therefore, $\phi(p_1) \in H$. Similarly, we have

$$\phi(p_2) \in H \text{ and } \phi(p_3) \in H.$$

But then, the points $\phi(p_1)$, $\phi(p_2)$, and (p_3) are collinear, which is a contradiction. □

Theorem 4.20 (Three inversions theorem for the hyperbolic plane). *An isometry of the hyperbolic plane is a composition of at most three inversions.*

Proof. Let $\phi : \mathbb{H}^2 \to \mathbb{H}^2$ be an isometry and p_1, p_2, and p_3 be non-collinear points. We divide the situation into four cases.

Case 1. First, assume that

$$\phi(p_1) = p_1, \phi(p_2) = p_2 \text{ and } \phi(p_3) = p_3;$$

then, $\phi = \mathrm{id}_{\mathbb{H}^2}$, letting $\psi = \mathrm{id}_{\mathbb{H}^2}$ in Theorem 4.19.

Case 2. If only two of p_1, p_2, and p_3 coincide with their images under ϕ, say

$$\phi(p_1) = p_1 \text{ and } \phi(p_2) = p_2 \text{ but } \phi(p_3) \neq p_3,$$

then

$$d_{\mathbb{H}^2}(p_3, p_1) = d_{\mathbb{H}^2}(\phi(p_3), \phi(p_1)) = d_{\mathbb{H}^2}(\phi(p_3), p_1).$$

Therefore, letting $H = H_{p_3,\phi(p_3)}$, we have $p_1 \in H$. Similarly, we have $p_2 \in H$. Let $\psi = I_H \circ \phi$; then,

$$\begin{aligned} \psi(p_1) &= I_H(\phi(p_1)) \\ &= I_H(p_1) \\ &= p_1. && (\because \ p_1 \in H) \end{aligned}$$

Similarly, $\psi(p_2) = p_2$. Note also that

$$\psi(p_3) = I_H(\phi(p_3)) = p_3. \qquad (\because \ H = H_{p_3,\phi(p_3)})$$

Again, $\psi = \mathrm{id}_{\mathbb{H}^2}$ by Theorem 4.19. Therefore, $I_H \circ \phi = \mathrm{id}_{\mathbb{H}^2}$. Hence,

$$\phi = I_H^{-1} = I_H,$$

and ϕ is an inversion.

Case 3. If only one of p_1, p_2, and p_3 coincides with its image under ϕ, say

$$\phi(p_1) = p_1 \text{ but } \phi(p_2) \neq p_2 \text{ and } \phi(p_3) \neq p_3,$$

then

$$d_{\mathbb{H}^2}(p_2, p_1) = d_{\mathbb{H}^2}(\phi(p_2), \phi(p_1)) = d_{\mathbb{H}^2}(\phi(p_2), p_1).$$

Therefore, letting $M = H_{p_2,\phi(p_2)}$, we have $p_1 \in M$. Let $\phi' = I_M \circ \phi$; then,

$$\phi'(p_1) = I_M(\phi(p_1)) = I_M(p_1) = p_1,$$

and

$$\phi'(p_2) = I_M(\phi(p_2)) = p_2,$$

converting to Case 1 or Case 2. Therefore, $\phi' = \mathrm{id}_{\mathbb{H}^2}$ or $\phi' = I_H$ for some circle H, i.e., $\phi = I_M$ or $\phi = I_M \circ I_H$.

Case 4. Finally, assume that

$$\phi(p_1) \neq p_1, \phi(p_2) \neq p_2 \text{ and } \phi(p_3) \neq p_3.$$

Let $N = H_{p_1,\phi(p_1)}$ and $\phi'' = I_N \circ \phi$. Then,

$$\phi''(p_1) = I_N(\phi(p_1)) = p_1,$$

again converting to Case 1, Case 2, or Case 3. Therefore,

$$\phi'' = \mathrm{id}_{\mathbb{H}^2}, \phi'' = I_M \text{ or } \phi'' = I_M \circ I_H$$

for some circles H and M, i.e.,

$$\phi = I_N, \phi = I_N \circ I_M \text{ or } \phi = I_N \circ I_M \circ I_H.$$

□

Since an isometry is a composition of inversions and an inversion preserves angles, we pose the following corollary.

Corollary 4.21. *An isometry of the hyperbolic plane preserves angles.*

Let $\mathrm{Iso}^+(\mathbb{H}^2)$ be the set of all isometries that are compositions of even numbers of inversions inversions and $\mathrm{Iso}^-(\mathbb{H}^2)$ be the set of isometries that are compositions of odd numbers of isometry in $\mathrm{Iso}^+(\mathbb{H}^2)$ is said to be orientation-preserving and an isometry in $\mathrm{Iso}^-(\mathbb{H}^2)$ is said to be orientation-reversing.

Exercises

4.5. Suppose that two distinct points p and q are fixed points of an isometry

$$\phi : \mathbb{H}^2 \to \mathbb{H}^2.$$

Show that every point on the hyperbolic line through p and q is also a fixed point of ϕ.

4.6. Recall that given the isometries ϕ and ψ of $\mathbb{H}^2$, the conjugation of ψ by ϕ is the isometry

$${}^{\phi}\psi = \phi \circ \psi \circ \phi^{-1}.$$

(a) For inversions I_H and I_M in hyperbolic lines H and M, show that

$${}^{I_M}I_H = I_{H'},$$

where $H' = I_M(H)$.

(b) For an isometry ϕ and a hyperbolic line H, show that

$${}^{\phi}I_H = I_{H'},$$

where $H' = \phi(H)$.

4.7. Show that a translation in the x-direction and rescaling about the origin are compositions of two inversions.

4.8. Two $\mathbb{H}^2$-lines L and M are said to be *ultraparallel* if there is another $\mathbb{H}^2$-line that is orthogonal to both the $\mathbb{H}^2$-lines L, M. Consider an isometry $\phi = I_M \circ I_L$ for ultraparallel $\mathbb{H}^2$-lines L, M. Show that there is an isometry ψ such that

$${}^{\psi}\phi = d_r$$

for some $r > 0$.

(*Hint.* Consider an isometry that sends the common perpendicular $\mathbb{H}^2$-line to the y-axis.)

4.9. Note that an isometry ϕ of $\mathbb{H}^2$ can be defined for points on the x-axis, possibly except for one point, if one sets the image of that point under the isometry as ∞.

1. Assume that $\phi(1,0) = (1,0), \phi(0,0) = (0,0)$ and $\phi(\infty) = \infty$. Show that $\phi = \mathrm{id}_{\mathbb{H}^2}$.
2. For each linear fractional transformation $f : \mathbb{R}_\infty \to \mathbb{R}_\infty$ (see Exercise 3.13), show that there exists a unique isometry ϕ of $\mathbb{H}^2$ such that

$$\phi(x,0) = (f(x),0)$$

for any $x \in \mathbb{R}_\infty$, where we regard $(\infty,0)$ (for $\infty \in \mathbb{R}_\infty$) as ∞ (for $\infty \in \mathbb{R}^2_\infty$).

4.4 Hyperbolic Triangle and Hyperbolic Area

An $\mathbb{H}^2$-triangle consists of three edges of $\mathbb{H}^2$-lines. An $\mathbb{H}^2$-polygon is a figure on $\mathbb{H}^2$ that can be expressed as the union of a finite number of $\mathbb{H}^2$-triangles, overlapping only in edges and vertices (Figure 4.9).

Two regions on $\mathbb{H}^2$ are said to be *congruent* if they are images of each other by some isometry of the hyperbolic plane.

Different from triangles on the Euclidean plane and sphere, we do not have an intuitive notion for an area for a region on $\mathbb{H}^2$. However, we will show that we can define a *hyperbolic area* satisfying the following conditions.

A1. For each $\mathbb{H}^2$-polygon R, the $\mathbb{H}^2$-area of R (denote $\mathrm{Area}_{\mathbb{H}^2}(R)$) is a positive number.
A2. (Congruence) If two $\mathbb{H}^2$-triangles are congruent, then they have equal $\mathbb{H}^2$-areas.
A3. (Additivity) If $R = R_1 \cup R_2$ is the union of two $\mathbb{H}^2$-polygons, overlapping only in edges and vertices, then

$$\mathrm{Area}_{\mathbb{H}^2}(R) = \mathrm{Area}_{\mathbb{H}^2}(R_1) + \mathrm{Area}_{\mathbb{H}^2}(R_2).$$

(continued)

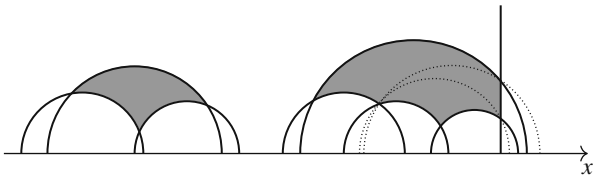


Fig. 4.9 Hyperbolic triangle and hyperbolic polygon

A4. Let T be a right-angled (i.e., one angle is $\frac{\pi}{2}$) $\mathbb{H}^2$-triangle with like sides of $\mathbb{H}^2$-length a as in Figure 4.10. Then,

$$\frac{\text{Area}_{\mathbb{H}^2}(T)}{\frac{1}{2}a^2}$$

converges to one as a goes to zero.

When a right-angled triangle is very tiny, then it is very similar to one in the Euclidean plane (Figure 4.10). This is why we require condition A4. Note that area on the sphere also satisfies condition A4 (Exercise 2.14).

For a region, which can be approximated by an $\mathbb{H}^2$-polygon, its area can be defined as the limit of $\mathbb{H}^2$-areas of $\mathbb{H}^2$-polygons. Note that the ordinary Euclidean area satisfies A2. There is an elegant way to deduce a formula for $\mathbb{H}^2$-area only from the conditions A1 $\sim$ A4. Here, we give a somewhat crude but simpler definition by double integration.

Definition 4.22. Let R be an (integrable) region on $\mathbb{H}^2$; then, the $\mathbb{H}^2$-*area* of R is

$$\text{Area}_{\mathbb{H}^2}(R) = \int\int_R \frac{dxdy}{y^2}.$$

Then, conditions A1 and A3 are immediately satisfied.

Lemma 4.23. *The unbounded region Ω with interior angles α and β as in Figure 4.11 has a finite $\mathbb{H}^2$-area,*

$$\text{Area}_{\mathbb{H}^2}(\Omega) = \pi - (\alpha + \beta).$$

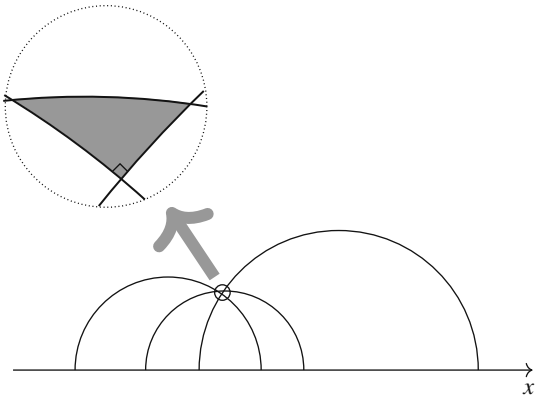


Fig. 4.10 Very tiny right-angled $\mathbb{H}^2$-triangle and its enlargement

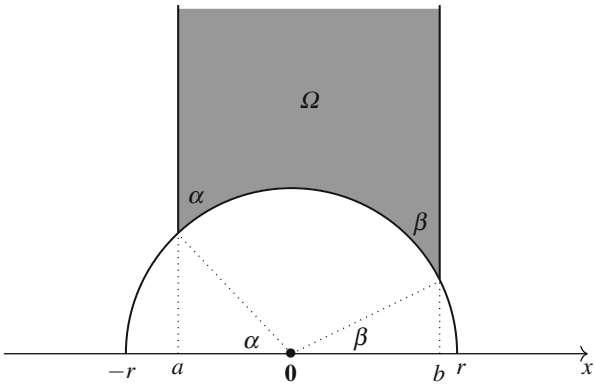


Fig. 4.11 Unbounded region Ω with interior angles α and β

Proof. Note that

$$a = r\cos(\pi - \alpha) \text{ and } b = r\cos\beta.$$

By Definition 4.22,

$$\text{Area}_{\mathbb{H}^2}(\Omega) = \int\int_{\Omega} \frac{dxdy}{y^2} = \int_a^b \int_{\sqrt{r^2-x^2}}^{\infty} \frac{1}{y^2} dydx = \int_a^b \frac{dx}{\sqrt{r^2-x^2}},$$

and making the change of variables $x = r\cos\theta$, we have

$$\int_a^b \frac{dx}{\sqrt{r^2-x^2}} = \int_{\pi-\alpha}^{\beta} \frac{-r\sin\theta}{r\sin\theta} d\theta = \pi - (\alpha + \beta).$$

□

Corollary 4.24. *The $\mathbb{H}^2$-area of an $\mathbb{H}^2$-triangle with interior angles α, β, and γ is*

$$\pi - (\alpha + \beta + \gamma).$$

Proof. Let T be an $\mathbb{H}^2$-triangle with interior angles α, β, and γ. If one of the edges is a part of a vertical line as in Figure 4.12, we obtain

$$\begin{aligned}\text{Area}_{\mathbb{H}^2}(T) &= \text{Area}_{\mathbb{H}^2}(T \cup \Omega) - \text{Area}_{\mathbb{H}^2}(\Omega) \\ &= (\pi - (\gamma + (\beta + \delta))) - (\pi - ((\pi - \alpha) + \delta)) \\ &= \pi - (\alpha + \beta + \gamma),\end{aligned}$$

where we used Lemma 4.23. If none of the edges are parts of vertical lines, we can divide the triangle as in Figure 4.13, where $\beta = \beta_1 + \beta_2$.

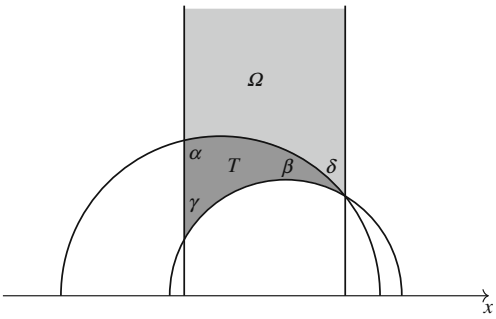


Fig. 4.12 Hyperbolic area of a hyperbolic triangle

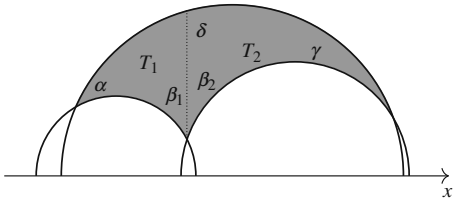


Fig. 4.13 Division of a hyperbolic square into two hyperbolic triangles T_1 and T_2

Note that $T = T_1 \cup T_2$ with one edge in common. Thus,

$$\begin{aligned}\text{Area}_{\mathbb{H}^2}(T) &= \text{Area}_{\mathbb{H}^2}(T_1) + \text{Area}_{\mathbb{H}^2}(T_2)\\ &= (\pi - (\alpha + \beta_1 + (\pi - \delta))) + (\pi - (\delta + \beta_2 + \gamma))\\ &= \pi - (\alpha + \beta_1 + \beta_2 + \gamma)\\ &= \pi - (\alpha + \beta + \gamma).\end{aligned}$$

□

Theorem 4.25. *The $\mathbb{H}^2$-area, defined in Definition 4.22, satisfies all conditions A1, A2, A3, and A4.*

Proof. As previously discussed, conditions A1 and A3 come immediately from Definition 4.22. Two congruent $\mathbb{H}^2$-triangles have the same interior angles (note that an isometry of $\mathbb{H}^2$ preserves angles); therefore, they have the same $\mathbb{H}^2$-area according to Corollary 4.24, which is condition A2.

To show A4, let T be an isosceles right-angled $\mathbb{H}^2$-triangle with like sides of $\mathbb{H}^2$-length a. Suppose that T shrinks to a point $p = (x, y)$. From the mean value theorem for the integration, there is a point $(u, v) \in T$ such that

$$\text{Area}_{\mathbb{H}^2}(T) = \text{Area}(T)\frac{1}{v^2},$$

where $\text{Area}(T)$ is the ordinary Euclidean area of T. Let Γ_1 and Γ_2 be two sides of $\mathbb{H}^2$-length a. Then, there are points $(u_1, v_1) \in \Gamma_1$ and $(u_2, v_2) \in \Gamma_2$ such that

$$a = \frac{l(\Gamma_1)}{v_1} = \frac{l(\Gamma_2)}{v_2},$$

where $l(\Gamma_i)$ is the Euclidean length of Γ_i. As T shrinks to a point (i.e., a approaches zero),

$$\text{Area}_{\mathbb{H}^2}(T) = \text{Area}(T)\frac{1}{v^2}$$

is approximated by

$$\frac{1}{2}l(\Gamma_1)l(\Gamma_2)\frac{1}{v^2} = \frac{1}{2}v_1 v_2 a^2 \frac{1}{v^2}.$$

Note that v_1, v_2, and v approach y as T shrinks to the point $p = (x, y)$. Therefore, as T shrinks to a point, the ratio

$$\frac{\text{Area}_{\mathbb{H}^2}(T)}{\frac{1}{2}a^2}$$

approaches

$$\frac{\frac{1}{2}y \cdot y \cdot a^2 \frac{1}{y^2}}{\frac{1}{2}a^2} = 1,$$

meeting condition A4. □

Corollary 4.26. *The sum of the interior angles of a hyperbolic triangle is less than* π.

Proof. Let T be an $\mathbb{H}^2$-triangle with interior angles α, β, and γ. According to condition A1, its $\mathbb{H}^2$-area is positive:

$$\text{Area}_{\mathbb{H}^2}(T) = \pi - (\alpha + \beta + \gamma) > 0,$$

concluding the proof. □

Exercises

4.10. Find a formula for the $\mathbb{H}^2$-area of an $\mathbb{H}^2$-polygon with interior angles $\theta_1, \theta_2, \cdots, \theta_n$.

4.11. Show that the $\mathbb{H}^2$-areas of $\mathbb{H}^2$-triangles are bounded, and find the least upper bound.

4.12. We call

$$\partial\mathbb{H}^2 := \{(x, y) \in \mathbb{R}^2 \mid y = 0\} \cup \{\infty\}$$

the *boundary* of $\mathbb{H}^2$. Note that the boundary points do not belong to $\mathbb{H}^2$. One can regard each $\mathbb{H}^2$-line as having two endpoints at the boundary $\partial\mathbb{H}^2$ (Figure 4.14).

An *ideal* $\mathbb{H}^2$-triangle is a triangle with three vertices at the boundary, a 2/3-ideal $\mathbb{H}^2$-triangle has two points at the boundary, and a 1/3-ideal $\mathbb{H}^2$-triangle has one point at the boundary (Figure 4.15).

Note that an ideal, a 2/3-ideal or a 1/3-ideal $\mathbb{H}^2$-triangle's vertices do not lie on $\mathbb{H}^2$ (Figures 4.16 and 4.17). Therefore, they are not real $\mathbb{H}^2$-triangles; they are not even bounded. However, we can show that they have finite $\mathbb{H}^2$-areas.

1. Show that all ideal $\mathbb{H}^2$-triangles are congruent.
2. Find the $\mathbb{H}^2$-area of an ideal $\mathbb{H}^2$-triangle.

4.13. Two $\mathbb{H}^2$-lines L and M are said to be *asymptotically parallel* if they intersect on the boundary of $\mathbb{H}^2$. Consider an isometry $\phi = I_M \circ I_L$ for asymptotically parallel $\mathbb{H}^2$-lines L, M. Show that there is an isometry ψ such that

$$^{\psi}\phi = t_{(1,0)}.$$

(*Hint.* Consider an isometry that sends the intersection point on $\partial\mathbb{H}^2$ to ∞.)

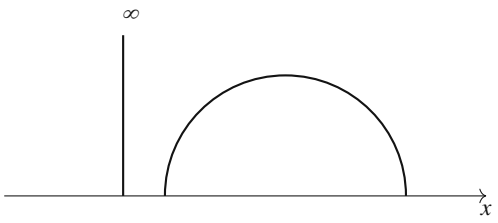


Fig. 4.14 Each $\mathbb{H}^2$-line has two endpoints at the boundary $\partial\mathbb{H}^2$

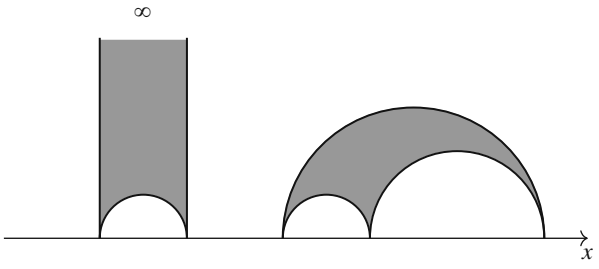


Fig. 4.15 Ideal $\mathbb{H}^2$-triangles

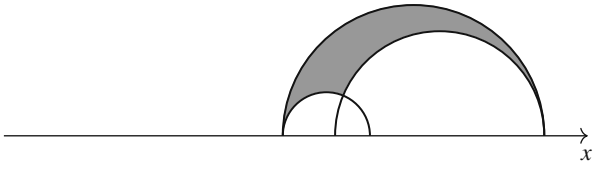


Fig. 4.16 2/3-ideal $\mathbb{H}^2$-triangles

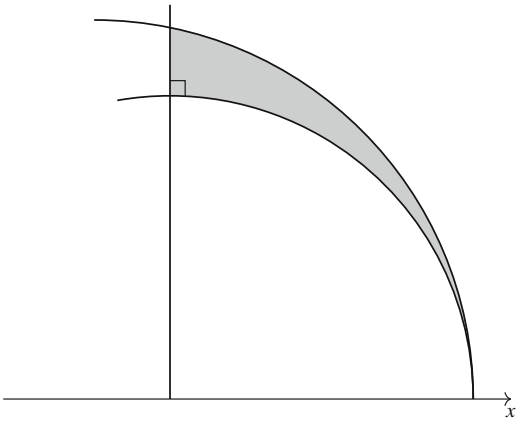


Fig. 4.17 1/3-ideal $\mathbb{H}^2$-triangles with a right angle

4.14. Two $\mathbb{H}^2$-lines are said to be *parallel* if they are disjoint. Show that two parallel $\mathbb{H}^2$-lines are ultraparallel or asymptotically parallel.

4.5 Poincaré Disk

Let C be a Euclidean circle of radius $\sqrt{2}$, centered at $(0, -1)$. Let

$$J = \bar{r} \circ I_C,$$

the composition of the inversion in C and reflection in the x-axis. Then, concretely, J is as follows:

$$J(x, y) = \left(\frac{2x}{x^2 + (1+y)^2}, \frac{x^2 + y^2 - 1}{x^2 + (1+y)^2} \right).$$

It is easy to verify that J maps $\mathbb{H}^2$ onto a unit disk

$$\mathbb{B}^2 := \{(x, y) \mid x^2 + y^2 < 1\}.$$

Let $\partial\mathbb{B}^2 := \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 = 1\}$, and call it the boundary of $\mathbb{B}^2$. As in Figure 4.18, $\partial\mathbb{H}^2$ is mapped to $\partial\mathbb{B}^2$ via J.

Let $\gamma : [a, b] \to \mathbb{B}^2$ be a smooth curve on $\mathbb{B}^2$. Then, a curve $J^{-1}(\gamma) : [a, b] \to \mathbb{H}^2$ on $\mathbb{H}^2$ is defined as follows:

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

We define the $\mathbb{D}^2$-length of γ as follows:

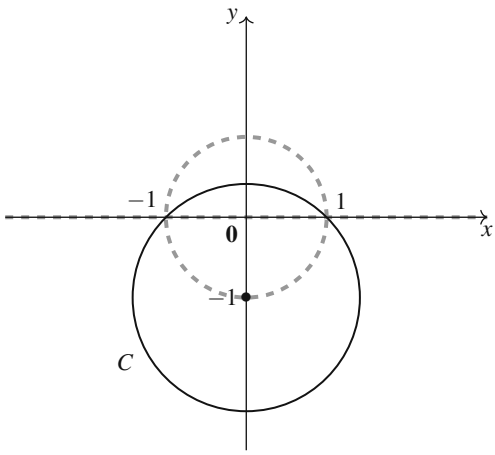


Fig. 4.18 J maps $\mathbb{H}^2$ onto a unit disk

$$l_{\mathbb{D}^2}(\gamma) = l_{\mathbb{H}^2}(J^{-1}(\gamma)).$$

The set $\mathbb{B}^2$ with this $\mathbb{D}^2$-length is denoted by $\mathbb{D}^2$ and is known as the *Poincaré disk*.

Let $\gamma : [a, b] \to \mathbb{D}^2$ be a smooth curve with $\gamma(t) = (u(t), v(t))$. Let $(u, v) = J(x, y)$. Then, it is not hard to verify that

$$x = \frac{2u}{u^2 + (v-1)^2},$$

$$\frac{dx}{dt} = \frac{2(-u^2 + (v-1)^2)}{(u^2 + (v-1)^2)^2}\frac{du}{dt} - \frac{4u(v-1)}{(u^2 + (v-1)^2)^2}\frac{dv}{dt},$$

and

$$y = \frac{-u^2 - v^2 + 1}{u^2 + (v-1)^2},$$

$$\frac{dy}{dt} = \frac{4u\,(v-1)}{\left(1 + u^2 - 2v + v^2\right)^2}\frac{du}{dt} + \frac{2\left(-u^2 + (v-1)^2\right)}{\left(1 + u^2 - 2v + v^2\right)^2}\frac{dv}{dt}.$$

A lengthy but direct calculation yields the following:

$$\frac{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2}{y^2} = 4\frac{\left(\frac{du}{dt}\right)^2 + \left(\frac{dv}{dt}\right)^2}{\left(1 - (u^2 + v^2)\right)^2}.$$

Therefore,

$$l_{\mathbb{D}^2}(\gamma) = l_{\mathbb{H}^2}(J^{-1}(\gamma)) = \int_a^b \frac{\sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2}}{y}\, dt = \int_a^b \frac{2\sqrt{\left(\frac{du}{dt}\right)^2 + \left(\frac{dv}{dt}\right)^2}}{1-(u^2+v^2)}\, dt.$$

Hence, we have

$$l_{\mathbb{D}^2}(\gamma) = \int_a^b \frac{2}{1-(u^2+v^2)}\sqrt{\left(\frac{du}{dt}\right)^2 + \left(\frac{dv}{dt}\right)^2}\, dt. \tag{4.1}$$

For two points p and q on $\mathbb{D}^2$, the $\mathbb{D}^2$-distance $d_{\mathbb{D}^2}(p, q)$ is defined as follows:

$$d_{\mathbb{D}^2}(p, q) := d_{\mathbb{H}^2}(J^{-1}(p), J^{-1}(q)).$$

An isometry of $\mathbb{D}^2$ is a bijective map from $\mathbb{D}^2$ to itself that preserves the $\mathbb{D}^2$-distance. Denote the set of all isometries of $\mathbb{D}^2$ by $\mathrm{Iso}(\mathbb{D}^2)$.

Proposition 4.27. *A bijection $\phi : \mathbb{D}^2 \to \mathbb{D}^2$ is an isometry of $\mathbb{D}^2$ if and only if*

$$J^{-1} \circ \phi \circ J : \mathbb{H}^2 \to \mathbb{H}^2$$

is an isometry of $\mathbb{H}^2$.

Proof. Let ϕ be an isometry of $\mathbb{D}^2$ and $\psi = J^{-1} \circ \phi \circ J$. For any two points $p, q \in \mathbb{H}^2$,

$$\begin{aligned} d_{\mathbb{H}^2}(\psi(p), \psi(q)) &= d_{\mathbb{H}^2}\left(J^{-1}(\phi(J(p))), J^{-1}(\phi(J(q)))\right) \\ &= d_{\mathbb{D}^2}\left(\phi(J(p)), \phi(J(q))\right) \\ &= d_{\mathbb{D}^2}\left(J(p), J(q)\right) \qquad (\because \phi \in \mathrm{Iso}(\mathbb{D}^2)) \\ &= d_{\mathbb{H}^2}\left(J^{-1}(J(p)), J^{-1}(J(q))\right) \\ &= d_{\mathbb{H}^2}(p, q). \end{aligned}$$

Therefore, ψ is an isometry of $\mathbb{H}^2$. The converse can be similarly shown. □

One can also define a $\mathbb{D}^2$-line.

Definition 4.28. For two distinct points p_1 and p_2 of $\mathbb{D}^2$, the set of points equi-$\mathbb{D}^2$-distant from p_1 and p_2 is called a $\mathbb{D}^2$-*line*:

$$H'_{p_1,p_2} = \{p \in \mathbb{D}^2 \mid d_{\mathbb{D}^2}(p, p_1) = d_{\mathbb{D}^2}(p, p_2)\}.$$

Proposition 4.29. *A subset H' of $\mathbb{D}^2$ is a $\mathbb{D}^2$-line if and only if $J^{-1}(H')$ is an $\mathbb{H}^2$-line.*

Proof. Note, for $p_1, p_2 \in \mathbb{D}^2$,

$$\begin{aligned} p \in H'_{p_1,p_2} &\Leftrightarrow d_{\mathbb{D}^2}(p, p_1) = d_{\mathbb{D}^2}(p, p_2) \\ &\Leftrightarrow d_{\mathbb{H}^2}\left(J^{-1}(p), J^{-1}(p_1)\right) = d_{\mathbb{H}^2}\left(J^{-1}(p), J^{-1}(p_2)\right) \\ &\Leftrightarrow J^{-1}(p) \in H_{J^{-1}(p_1),J^{-1}(p_2)}. \end{aligned}$$

Therefore, if H' is a $\mathbb{D}^2$-line, then $J^{-1}(H') = H_{J^{-1}(p_1),J^{-1}(p_2)}$, which is an $\mathbb{H}^2$-line. Conversely, let $J^{-1}(H')$ be an $\mathbb{H}^2$-line; then, there are points $q_1, q_2 \in \mathbb{H}^2$ such that

$$J^{-1}(H') = H_{q_1,q_2} = H_{J^{-1}(p_1),J^{-1}(p_2)},$$

where $p_1 = J(q_1)$ and $p_2 = J(q_2)$. Therefore,

$$H' = J\left(J^{-1}(H')\right) = J\left(H_{J^{-1}(p_1),J^{-1}(p_2)}\right) = H'_{p_1,p_2},$$

which is a $\mathbb{D}^2$-line. □

As sets, the boundary of $\mathbb{D}^2$ is the same as that of $\mathbb{B}^2$, i.e., $\partial\mathbb{D}^2 = \partial\mathbb{B}^2$.

Theorem 4.30. *A $\mathbb{D}^2$-line is a circline, intersecting the boundary $\partial\mathbb{D}^2$ of $\mathbb{D}^2$ orthogonally (Figure 4.19).*

Proof. Note that J is a composition of a reflection in a line and an inversion, which maps a circline to a circline and preserves angles. Hence, J also maps a circline to a circline and preserves angles. Since every $\mathbb{H}^2$-line is a circline that orthogonally intersects $\partial\mathbb{H}^2$ and J maps $\mathbb{H}^2$ onto $\mathbb{D}^2$, the proof is complete by Proposition 4.29. □

The following theorem can be showed easily.

Theorem 4.31. *An isometry of $\mathbb{D}^2$ is a composition of at most three inversions in $\mathbb{D}^2$-lines.*

Proof. Let ϕ be an isometry of $\mathbb{D}^2$, then, by Proposition 4.27,

$$J^{-1} \circ \phi \circ J : \mathbb{H}^2 \to \mathbb{H}^2$$

is an isometry of $\mathbb{S}^2$. By Theorem 4.20, $J^{-1} \circ \phi \circ J$ is a composition of at most three inversions in $\mathbb{H}^2$-lines:

$$J^{-1} \circ \phi \circ J = I_1 \circ I_2 \circ \cdots \circ I_n,$$

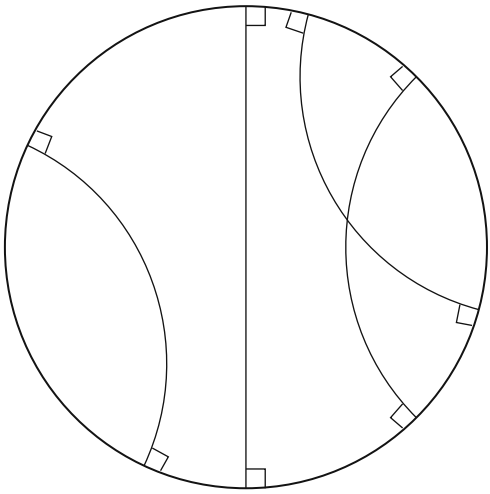


Fig. 4.19 $\mathbb{D}^2$-lines on $\mathbb{D}^2$

where $I_1, I_2, \cdots, I_n$ are some inversions in $\mathbb{H}^2$-lines with $n \leq 3$. Then

$$\begin{aligned}\phi &= J \circ I_1 \circ I_2 \circ \cdots \circ I_n \circ J^{-1} \\ &= J \circ\ I_1 \circ J^{-1} \circ J \circ I_2 \circ J^{-1} \circ J \circ \cdots \circ J^{-1} \circ J \circ I_n \circ J^{-1} \\ &= f_1 \circ f_2 \circ \cdots \circ f_n,\end{aligned}$$

where $f_i := J \circ I_i \circ J^{-1}$. Note that $I_i = I_{\Gamma_i}$ for some $\mathbb{H}^2$-line Γ_i.

$$\begin{aligned}f_i &= J \circ I_i \circ J^{-1} \\ &= \bar{r} \circ I_C \circ I_{\Gamma_i} \circ I_C \circ \bar{r} \\ &= \bar{r} \circ I_{I_C(\Gamma_i)} \circ \bar{r} \\ &= I_{\bar{r}(I_C(\Gamma_i))} \\ &= I_{J(\Gamma_i)},\end{aligned}$$

where we used Lemma 3.15 twice. By Proposition 4.29, $J(\Gamma_i)$ is a $\mathbb{D}^2$-line and f_i is an inversion in a $\mathbb{D}^2$-line. □

As usual, we define $\mathbb{D}^2$-circles (or $\mathbb{H}^2$ circles) as sets of points equi-$\mathbb{D}^2$-distant (or equi-$\mathbb{H}^2$-distant) from a certain point.

Lemma 4.32. *The $\mathbb{D}^2$-circle of $\mathbb{D}^2$-radius ρ, centered at the origin, is a Euclidean circle, centered at the origin. Its $\mathbb{D}^2$-circumference is $2\pi \sinh \rho$.*

Proof. It is not difficult to see that the $\mathbb{D}^2$-shortest path from a point in $\mathbb{D}^2$ to the origin is a part of a $\mathbb{D}^2$-line. Note also that every $\mathbb{D}^2$-line through the origin is a line. Let Γ be a $\mathbb{D}^2$-circle. For each point $p \in \Gamma$, let

$$p = (r\cos\theta, r\sin\theta)$$

for some θ, where r is the Euclidean radius of Γ. Then,

$$\gamma : [0, r] \to \mathbb{D}^2 \text{ by } \gamma(t) = (t\cos\theta, t\sin\theta) \tag{4.2}$$

is the $\mathbb{D}^2$-shortest path from the origin to point p. By the formula in (4.1), the $\mathbb{D}^2$-radius ρ of Γ is as follows:

$$\begin{aligned}
\rho &= d_{\mathbb{D}^2}(\mathbf{0}, p) \\
&= \int_0^r \frac{2\sqrt{\left(\frac{du}{dt}\right)^2 + \left(\frac{dv}{dt}\right)^2}}{1 - (u^2 + v^2)}\, dt \\
&= \int_0^r \frac{2\sqrt{\cos^2\theta + \sin^2\theta}}{1 - (t^2\cos^2\theta + t^2\sin^2\theta)}\, dt \\
&= \int_0^r \frac{2}{1 - t^2} dt = \ln(1 + r) - \ln(1 - r) \\
&= \ln\left(\frac{1 + r}{1 - r}\right) \\
&= \ln\left(\frac{2}{1 - r} - 1\right).
\end{aligned}$$

Hence,

$$e^\rho = \frac{2}{1 - r} - 1,$$

and

$$r = \frac{e^\rho - 1}{e^\rho + 1}.$$

Let us now determine the $\mathbb{D}^2$-circumference of Γ. A curve $\delta : [0, 2\pi] \to \mathbb{D}^2$, defined as

$$\delta(t) = (r\cos t, r\sin t),$$

parameterizes the circumference. Therefore, by the formula in (4.1),the $\mathbb{D}^2$ circumference is as follows:

$$\begin{aligned}\int_0^{2\pi} \frac{2\sqrt{\left(\frac{du}{dt}\right)^2+\left(\frac{dv}{dt}\right)^2}}{1-(u^2+v^2)}\,dt &= \int_0^{r} \frac{2\sqrt{r^2\sin^2 t + r^2\cos^2 t}}{1-(r^2\cos^2 t + r^2\sin^2 t)}\,dt \\ &= \int_0^{2\pi} \frac{2r}{1-r^2}dt \\ &= \pi\left(e^{\rho}-e^{-\rho}\right) \\ &= 2\pi\sinh\rho.\end{aligned}$$

□

Theorem 4.33.

1. *A $\mathbb{D}^2$-circle is a Euclidean circle, and its $\mathbb{D}^2$-circumference is $2\pi\sinh\rho$, where radius ρ is its $\mathbb{D}^2$-radius.*
2. *An $\mathbb{H}^2$ circle is a Euclidean circle, and its $\mathbb{H}^2$-circumference is $2\pi\sinh\rho$, where radius ρ is its $\mathbb{H}^2$-radius.*

Proof. Let C be a $\mathbb{D}^2$-circle, centered at a point $p \in \mathbb{D}^2$ with $\mathbb{D}^2$-radius ρ. We can choose an isometry ϕ of $\mathbb{D}^2$ such that $\phi(p) = \mathbf{0}$. Now $\phi(C)$ is centered at the origin, and its $\mathbb{D}^2$-circumference is $2\pi\sinh\rho$ according to Lemma 4.32. Hence, the $\mathbb{D}^2$-circumference of C is $2\pi\sinh\rho$.

Let Γ be an $\mathbb{H}^2$ circle with $\mathbb{H}^2$-radius ρ. Then, $J(\Gamma)$ is a $\mathbb{D}^2$-circle with $\mathbb{D}^2$-radius ρ, and its $\mathbb{D}^2$-circumference is $2\pi\sinh\rho$. Therefore, the $\mathbb{H}^2$-circumference Γ is $2\pi\sinh\rho$. □

Definition 4.34. The *Gaussian curvature* of a surface S at $p \in S$ is as follows:

$$K = \lim_{\rho\to 0^+} 3\cdot\frac{2\pi\rho - \Gamma(\rho)}{\pi\rho^3},$$

where $\Gamma(\rho)$ is the circumference (measured in S) of the circle of radius ρ centered at p.

The Gaussian curvature is a fundamental geometric invariant. It is not difficult to show that two isometric surfaces have the same Gaussian curvature at their corresponding points (Exercise 4.18). Accordingly, the following theorem implies that the Euclidean plane, the sphere, and the hyperbolic plane are not isometric, i.e., they are geometrically different.

Theorem 4.35.

1. *The Gaussian curvature of the Euclidean plane is* 0 *at each of its points.*
2. *The Gaussian curvature of the sphere is* $+1$ *at each of its points.*
3. *The Gaussian curvature of the hyperbolic plane is* -1 *at each of its points.*

Proof. The first statement is obvious. The circumference of a circle on $\mathbb{S}^2$ with spherical radius ρ is $2\pi \sin\rho$. Therefore, the Gaussian curvature of the sphere is $+1$.

Finally, according to point 2 in Theorem 4.33, the Gaussian curvature of the hyperbolic plane is -1 at each of its points. □

Exercises

4.15. Note that a Euclidean rotation r_θ is also an isometry of $\mathbb{D}^2$. Consider an isometry $\phi = I_M \circ I_H$ of $\mathbb{D}^2$ for $\mathbb{D}^2$-lines H and M that intersect each other. Show that there is an isometry ψ of $\mathbb{D}^2$ such that

$$^{\psi}\phi = r_\theta$$

for some angle θ.

Show also that θ is twice the angle between H and M at the intersection point.

4.16. Show that the $\mathbb{H}^2$-radius of the inscribed circle of an $\mathbb{H}^2$-triangle is bounded. Find its least upper bound.

(*Hint.* Try to solve the problem on $\mathbb{D}^2$.)

4.17. Show that the length of the altitude of any isosceles right-angled $\mathbb{H}^2$-triangle is bounded. Find its least upper bound, which is the *Schweikart's constant*.

(*Hint.* Try to solve the problem on $\mathbb{D}^2$ and consider a $\mathbb{D}^2$-triangle whose right-angled vertex is the origin.)

4.6 Klein Disk

We introduce another model of a hyperbolic surface. Consider a hemisphere $\mathbb{J}^2 = \{(x, y, z) \in \mathbb{S}^2 \mid z > 0\}$, with an equator G of $\mathbb{S}^2$ and a reflection $\bar{r}_G$ of $\mathbb{S}^2$ in G. Then, we have a bijective map

$$f = \bar{r}_G \circ (\Phi^{-1}|_{\mathbb{B}^2}) : \mathbb{B}^2 \to \mathbb{J}^2,$$

$$g = \pi_z|_{\mathbb{J}^2} : \mathbb{J}^2 \to \mathbb{B}^2,$$

and

$$K = g \circ f : \mathbb{B}^2 \to \mathbb{B}^2,$$

where π_z is the projection of $\mathbb{R}^3$ to the xy-plane along the z-axis. More concretely,

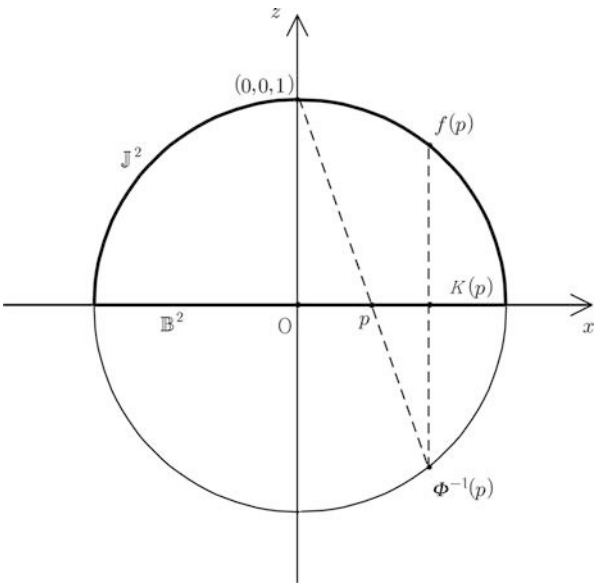


Fig. 4.20 The maps f, g, and K in the zx-plane, cross-section view

$$K(x, y) = \left(\frac{2x}{1 + x^2 + y^2}, \frac{2y}{1 + x^2 + y^2} \right).$$

In Figure 4.20, these maps are illustrated (zx-plane cross-section view).

Let Γ be a curve on $\mathbb{B}^2$. Then, we have another curve $K^{-1}(\Gamma)$ on $\mathbb{B}^2$. We define the $\mathbb{K}^2$-length of Γ as follows:

$$l_{\mathbb{K}^2}(\Gamma) := l_{\mathbb{D}^2}(K^{-1}(\Gamma)).$$

The set $\mathbb{B}^2$ with this $\mathbb{K}^2$-length, denoted $\mathbb{K}^2$, is called the *Klein disk*. The notions of the $\mathbb{K}^2$-distance $d_{\mathbb{K}^2}$, $\mathbb{K}^2$-line, and $\mathbb{K}^2$-circle are similarly defined.

Recall that a set M with a distance d_M is called a (metric) space. Recall also that if there is a bijective map ϕ from a space M_1 to another space M_2 that preserves the distances d_{M_1} and d_{M_2}, i.e.,

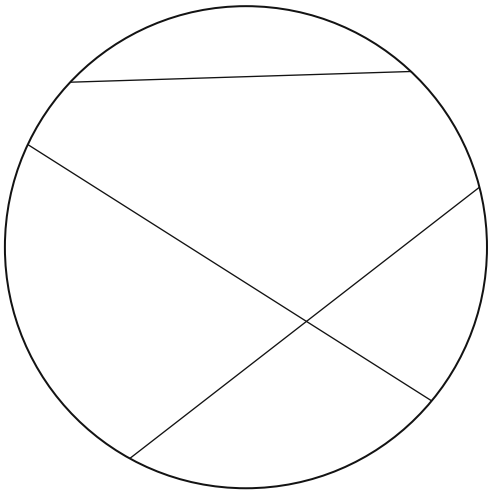
$$d_{M_2}(\phi(p), \phi(q)) = d_{M_1}(p, q)$$

for any $p, q \in M_1$, then the spaces M_1 and M_2 are said to be isometric.

Hence $\mathbb{H}^2$, $\mathbb{D}^2$ and $\mathbb{K}^2$ are isometric spaces that represent the same geometry:

$$\mathbb{H}^2 \xrightarrow{J} \mathbb{D}^2 \xrightarrow{K} \mathbb{K}^2.$$

Fig. 4.21 $\mathbb{K}^2$-lines on $\mathbb{K}^2$



The $\mathbb{K}^2$ length of a curve $\gamma : [a, b] \to \mathbb{K}^2$, $\gamma(t) = (x(t), y(t))$, can be shown to have the following form:

$$l_{\mathbb{K}^2}(\gamma) = \int_a^b \sqrt{\frac{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2}{1 - x^2 - y^2} + \frac{\left(x\frac{dx}{dt} + y\frac{dy}{dt}\right)^2}{(1 - x^2 - y^2)^2}}\, dt, \tag{4.3}$$

which is quite complicated. However, the Klein disk model has an advantage—a $\mathbb{K}^2$-line is a Euclidean line.

Theorem 4.36. *A $\mathbb{K}^2$-line is a Euclidean line. Conversely, a Euclidean line on $\mathbb{K}^2$ is a $\mathbb{K}^2$-line (Figure 4.21).*

Proof. If a curve Γ on $\mathbb{K}^2$ is a $\mathbb{K}^2$-line, then the curve

$$\Gamma' = K^{-1}(\Gamma)$$

is a $\mathbb{D}^2$-line. Hence, Γ' is a circline that intersects orthogonally with $\partial\mathbb{B}^2$. Note that the map f is angle-preserving and sends one circline to another one. Hence, $f(\Gamma')$ is a circline on $\mathbb{S}^2$ that intersects orthogonally with the equator G of $\mathbb{S}^2$. The plane that cuts $\mathbb{S}^2$ along $f(\Gamma')$ meets orthogonally with the xy-plane. Now it is easy to see that the curve

$$\Gamma = K(\Gamma') = g(f(\Gamma'))$$

is the intersection of this plane with the xy-plane, which is a Euclidean line. □

Note that the projection g is not angle-preserving and does not send a circle to a circle. Hence, a $\mathbb{K}^2$ circle does not need to be a Euclidean circle (Exercise 4.20).

Exercises

4.18. Prove that the Gaussian curvature is a geometric invariant. In concrete words, for an isometry $\phi : S \to S'$ from a surface S with distance d to another surface S' with distance d' and a given point $p \in S$, show that the Gaussian curvature of S at point p is equal to that of S' at points $\phi(p)$.

4.19. Show that the map K is not angle-preserving.

4.20. Show that a $\mathbb{K}^2$ circle is a Euclidean ellipse. When is it a Euclidean circle?

4.7 Euclid's Fifth Postulate: The Parallel Postulate

At approximately 300 BC, Euclid wrote a book called "Elements" that encompassed all the mathematics known to the Greeks up to that time. Numerous theorems about geometry are derived from the following five postulates—postulates are so self-evident that they must be accepted without proof.

1. Two distinct points on the plane determine a line segment.
2. Line segments can be extended indefinitely in a straight line.
3. For any straight line segment, a circle can be drawn having the segment as the radius and one endpoint as the center.
4. All right angles are congruent.
5. If a line segment intersects two straight lines forming two interior angles on the same side that sum to less than π, then the two lines, if extended indefinitely, meet on the side on which the angles sum to less than π.

The fifth postulate (commonly called the *parallel postulate*) is more complicated than the preceding four postulates. Euclid used the previous four postulates to obtain 28 theorems. The geometry based on the axiom system without the fifth postulate is called the *absolute geometry*.

Until the 1800s, people believed that this absolute geometry was actually the Euclidean geometry and thought that the fifth postulate would be a mere theorem of the absolute geometry. They attempted to "prove" the fifth postulate, using only the other four postulates, for 2,000 years. Many people claimed that they actually did so, but they were, in fact, assuming some hypotheses that were equivalent to the fifth postulate. Some of those hypotheses are the following:

- There is at most one line that can be drawn parallel to another given line through an external point.

- The sum of the interior angles in every triangle is exactly π.
- There exists a pair of similar, but not congruent, triangles.
- The Pythagorean Theorem.
- There is no upper limit to the area of a triangle.
- The circumference of any circle of radius r is $2\pi r$.
- There is a quadrilateral whose interior angles are all $\frac{\pi}{2}$.

Now, it is easy to see that all the postulates except the fifth one hold for the hyperbolic plane. Hence, the existence of hyperbolic geometry can be regarded as the evidence that Euclid's fifth postulate is independent of the other four postulates. For a more detailed history of the development of hyperbolic geometry and non-Euclidean geometry, we refer to [15].

There are many differences between the geometries of the Euclidean plane, the sphere, and the hyperbolic plane. The most intrinsic one is the difference between their Gaussian curvatures, as shown in Theorem 4.35 (see also Exercise 4.18). One of the other differences is the tessellation on them. A *tessellation* is an arrangement of flat shapes, called *tiles*, without overlaps or gaps. Some tessellations involve many types of tiles; however, the most interesting tessellations use only one or a few different types of tiles. A *regular tessellation* is a pattern made by repeating a regular polygon. A regular tessellation is described by two positive integers, $[s, t]$, where s is the number of sides on the regular tiling polygon and t is the number of these polygons that meet at a vertex.

In the Euclidean plane, the vertex angle of a regular s-sided polygon is equal to

$$\pi - \frac{2\pi}{s}.$$

For a regular tessellation on the Euclidean plane where t polygons meet at a vertex,

$$t\left(\pi - \frac{2\pi}{s}\right) = 2\pi$$

or, equivalently,

$$(s-2)(t-2) = 4.$$

Hence, there are three possibilities:

$$[s, t] = [3, 6], [4, 4], [6, 3].$$

This is why there are only three types of regular tessellations, whose tiles are triangles, squares, and hexagons (Figure 4.22).

For regular tessellations on the sphere, the vertex angle of a regular s-sided polygon is greater than $\pi - \frac{2\pi}{s}$. Hence, we have

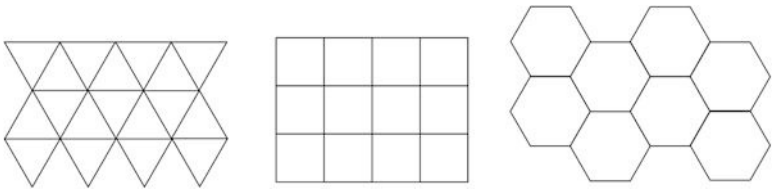


Fig. 4.22 Regular tessellations of types [3, 6], [4, 4], and [6, 3] on $\mathbb{R}^2$

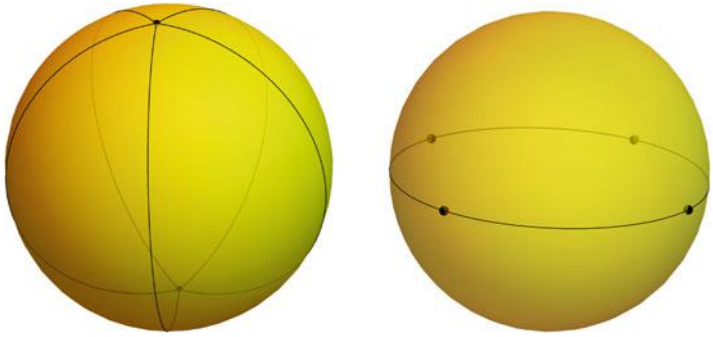


Fig. 4.23 Regular spherical tessellations of types [2, 5] and [4, 2]

$$t\left(\pi - \frac{2\pi}{s}\right) < 2\pi$$

or, equivalently,

$$(s-2)(t-2) < 4.$$

Hence, in contrast with the Euclidean plane, $s = 2$ or $t = 2$ can be a solution for the equation. For the sphere, a spherical lune can be regarded as a regular 2-gon. The combination of t spherical lunes with interior angles $\frac{2\pi}{t}$ forms a regular tessellation of the sphere with t tiles. This is a regular spherical tessellation of type $[2, t]$ (Figure 4.23). However, a hemisphere can be regarded as an s-gon by placing s vertices on its boundary. The use of two hemispheres forms a regular spherical tessellation of the sphere with two tiles of regular s-gons of types $[s, 2]$ (Figure 4.23). Other than these two classes of regular spherical tessellations, a regular spherical tessellation of the sphere has a one-to-one correspondence to a Platonic solid, which is a projection of the boundary of the solid from the origin to the sphere. We have seen that there are exactly five Platonic solids (Theorem 2.14). Hence, there are five regular spherical tessellations of spheres of this type.

In the case of the hyperbolic plane, the vertex angle of a regular s-sided polygon is less than $\pi - \frac{2\pi}{s}$. Hence,

$$t\left(\pi - \frac{2\pi}{s}\right) > 2\pi$$

or, equivalently,

$$(s-2)(t-2) > 4.$$

Note that there are infinitely many possibilities. The following proposition guarantees that all these possibilities are realizable (Exercise 4.21).

Proposition 4.37. *For a given integer n with $n \geq 3$ and a given angle θ with*

$$0 < \theta < \frac{(n-2)\pi}{n},$$

there is a regular n-gon on the hyperbolic plane whose interior angle is θ.

Proof. We will show it on the Poincaré disk $\mathbb{D}^2$. Draw a Euclidean circle Γ that intersects orthogonally with $\partial\mathbb{D}^2$ such that the Euclidean center of the circle lies on the x-axis. Note that Γ is a $\mathbb{D}^2$-line. We can also require that the circle intersects with the line segment $\overline{\mathbf{0}p}$ with the angle $\frac{\theta}{2}$, where $p = (\cos\varphi, \sin\varphi)$ with $\varphi = \frac{\pi}{n}$ (Figure 4.24). Let

$$\Gamma_k = r_{\frac{2k\pi}{n}}(\Gamma).$$

Then, $\Gamma_0, \Gamma_1, \cdots, \Gamma_{n-1}$ together form a regular n-gon whose interior angle is θ. □

The regular hyperbolic tessellations shown in Figure 4.25 are of type [4, 5] which are composed of $\mathbb{D}^2$-squares.

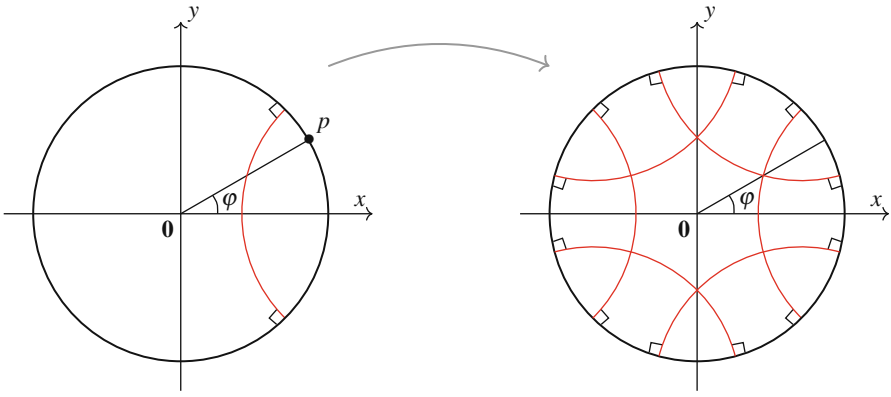


Fig. 4.24 A construction of a regular $\mathbb{D}^2$ hexagon with interior angle θ

The regular hyperbolic tessellations shown in Figure 4.26 are of types [5, 4] and [7, 4], respectively. Note that all the interior angles of the tiles are $\frac{\pi}{2}$.

The regular hyperbolic tessellations in Figure 4.27 are of types [7, 3] and [7, 7], respectively.

Circle Limit IV (Heaven and Hell) by M.C. Escher(1960) is related with a regular hyperbolic tessellation of type [6, 4][1].

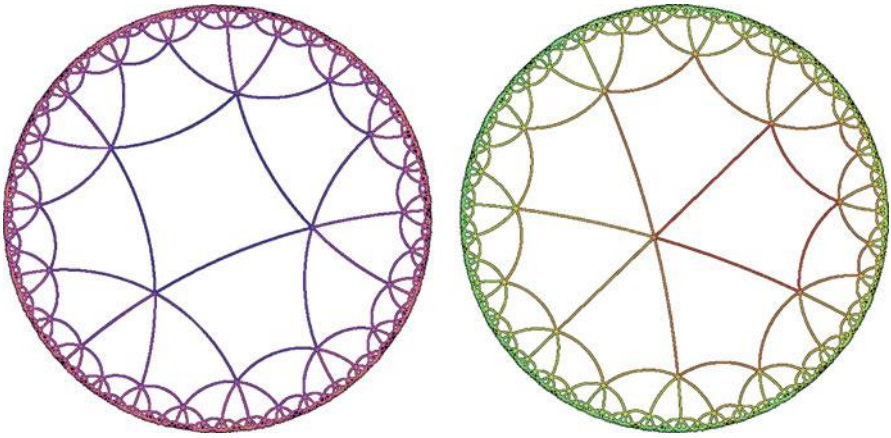


Fig. 4.25 Two regular hyperbolic tessellations of type [4, 6]

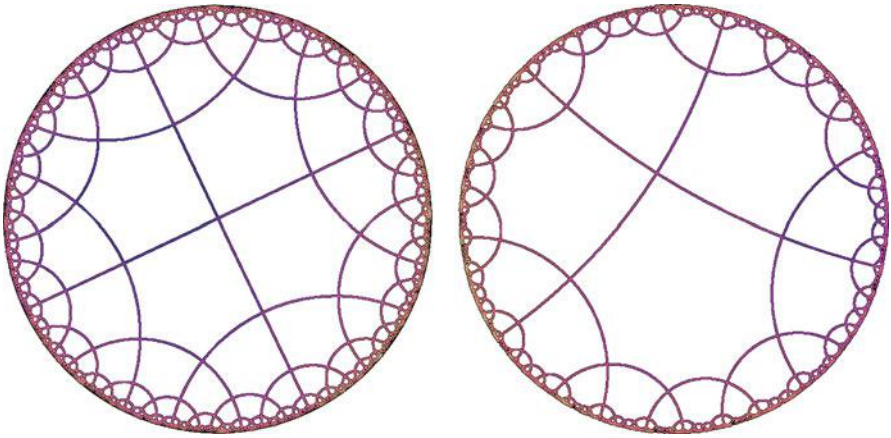


Fig. 4.26 Regular hyperbolic tessellations of types [5, 4] and [7, 4]

[1] https://www.escherinhetpaleis.nl/escher-today/circle-limit-iv-heaven-and-hell/?lang=en

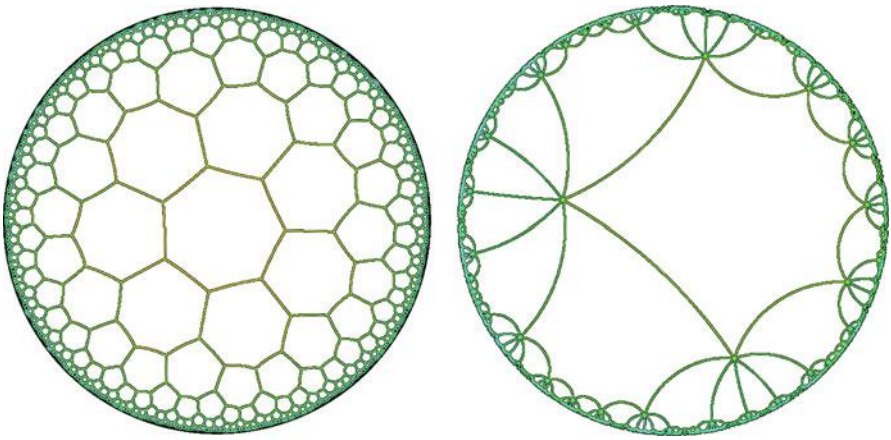


Fig. 4.27 Regular hyperbolic tessellations of types [7, 3] and [7, 7]

Exercises

4.21. For each pair of non-negative integers s and t with

$$(s-2)(t-2) > 4,$$

show that there is a regular hyperbolic tessellation of type $[s, t]$.

Chapter 5
Lorentz–Minkowski Plane

"The essence of mathematics is its freedom."

Georg Cantor (1845–1918)

"Do not fear to be eccentric in opinion, for every opinion now accepted was once eccentric."

Bertrand Russell (1872–1970)

Special relativity (also known as the special theory of relativity) is an experimentally well-confirmed and universally accepted physical theory that explains how space and time are linked. It was originally proposed by Albert Einstein. Today, special relativity is accepted as the most accurate theory of motion at any speed when gravitational forces are negligible. Special relativity leads to a wide range of consequences, which have been experimentally confirmed, including length contraction, time dilation, relativity of simultaneity, a universal speed limit and the mass-energy equivalence. It has replaced the long-standing notion of an absolute universal time by the notion of a relative time that is not independent of a reference frame and spatial position. Rather than treating the invariant time and the invariant spatial intervals between two events separately, we must consider an invariant spacetime interval, which enables us to understand spacetime from a geometric view of distances and isometries. In this chapter, we study the simple two-dimensional case. The goal is to formulate the geometry of special relativity in a truly equal setting along with other classical geometries. We will also see that hyperbolic geometry and relativistic geometry are intrinsically related.

N.-H. Lee, *Geometry: from Isometries to Special Relativity*, Undergraduate Texts in Mathematics, https://doi.org/10.1007/978-3-030-42101-4_5

5.1 Lorentz–Minkowski Distance

For $e_1 = (x_1, \tau_1)$, $e_2 = (x_2, \tau_2)$ in $\mathbb{R}^2$, the *Lorentz–Minkowski distance* between e_1 and e_2 is as follows[1]:

$$d^{\mathrm{II}}(e_1, e_2) = (x_1 - x_2)^2 - (\tau_1 - \tau_2)^2.$$

The set $\mathbb{R}^2$, together with the *Lorentz–Minkowski distance*, is called the Lorentz–Minkowski plane. To distinguish the Lorentz–Minkowski plane from the Euclidean plane, we denote the set by $\mathbb{R}^{1,1}$ (Figure 5.1) and call its elements *events*. A bijective map $\phi : \mathbb{R}^{1,1} \to \mathbb{R}^{1,1}$ is called an *isometry* of the Lorentz–Minkowski plane if it preserves the Lorentz–Minkowski distance, i.e.,

$$d^{\mathrm{II}}(\phi(e_1), \phi(e_2)) = d^{\mathrm{II}}(e_1, e_2)$$

for any two events $e_1, e_2 \in \mathbb{R}^{1,1}$. We denote by $\mathrm{Iso}(\mathbb{R}^{1,1})$ the set of all isometries of the Lorentz–Minkowski plane. A line on the Lorentz–Minkowski plane can be regarded as the set of events equidistant from two distinct events e_1 and e_2:

$$L_{e_1,e_2} := \{e \in \mathbb{R}^{1,1} \mid d^{\mathrm{II}}(e_1, e) = d^{\mathrm{II}}(e_2, e)\}.$$

Let α be an event and $t_\alpha(e) = \alpha + e$; then, t_α is an isometry that is called a *translation*. The following properties can be readily proven:

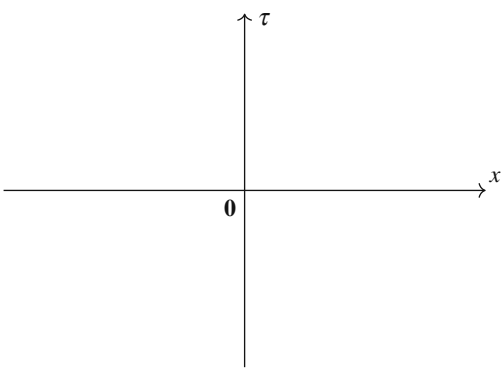


Fig. 5.1 Lorentz–Minkowski plane $\mathbb{R}^{1,1}$

[1] In a general system of units,

$$d^{\mathrm{II}}(e_1, e_2) = (x_1 - x_2)^2 - c^2(t_1 - t_2)^2$$

for $e_1 = (x_1, t_1)$, $e_2 = (x_2, t_2)$, where c is the speed of light. In this book, we are using a timescale so that $c = 1$.

$$d^{\mathrm{II}}(e_1, e_2) = d^{\mathrm{II}}(e_2, e_1),$$

$$d^{\mathrm{II}}(e_1, e_2) = d^{\mathrm{II}}(e_2 - e_1, \mathbf{0}),$$

and

$$d^{\mathrm{II}}(ae_1, ae_2) = a^2 d^{\mathrm{II}}(e_1, e_2)$$

for $a \in \mathbb{R}$.

Lemma 5.1. *An isometry of the Lorentz–Minkowski plane maps a line to a line.*

Proof. Let L be a line and ϕ be an isometry; then, there are two distinct events e_1 and e_2 such that

$$L = \{p \in \mathbb{R}^{1,1} \mid d^{\mathrm{II}}(e_1, p) = d^{\mathrm{II}}(e_2, p)\}.$$

Since ϕ is injective, the events $\phi(e_1)$ and $\phi(e_2)$ are distinct. Let

$$L' = \{q \in \mathbb{R}^{1,1} \mid d^{\mathrm{II}}(\phi(e_1), q) = d^{\mathrm{II}}(\phi(e_2), q)\}$$

such that L' is also a line. Then, it is adequate to show that

$$L' = \phi(L).$$

Let $p \in L$ be an event such that

$$d^{\mathrm{II}}(e_1, p) = d^{\mathrm{II}}(e_2, p).$$

Since ϕ preserves the Lorentz–Minkowski distance,

$$d^{\mathrm{II}}(\phi(e_1), \phi(p)) = d^{\mathrm{II}}(e_1, p) = d^{\mathrm{II}}(e_2, p) = d^{\mathrm{II}}(\phi(e_2), \phi(p)),$$

i.e.,

$$d^{\mathrm{II}}(\phi(e_1), \phi(p)) = d^{\mathrm{II}}(\phi(e_2), \phi(p));$$

therefore, $\phi(p) \in L'$. Conversely, let $q \in L'$; then,

$$d^{\mathrm{II}}(\phi(e_1), q) = d^{\mathrm{II}}(\phi(e_2), q).$$

Since ϕ is surjective, there is an event p such that $q = \phi(p)$. Accordingly,

$$d^{\mathrm{II}}(\phi(e_1), \phi(p)) = d^{\mathrm{II}}(\phi(e_2), \phi(p)).$$

Again,

$$d^{\mathrm{II}}(e_1, p) = d^{\mathrm{II}}(\phi(e_1), \phi(p)) = d^{\mathrm{II}}(\phi(e_2), \phi(p)) = d^{\mathrm{II}}(e_2, p),$$

i.e., $d^{\mathrm{II}}(e_1, p) = d^{\mathrm{II}}(e_2, p)$. Thus, $p \in L$, and then,

$$q = \phi(p) \in \phi(L).$$

□

Events in $\mathbb{R}^{1,1}$ are said to be collinear if they lie on the same line. The following lemma is the Lorentz–Minkowski plane version of the "three points theorem."

Lemma 5.2 (Three events theorem). *Let ϕ and ψ be isometries of the Lorentz–Minkowski plane. If*

$$\phi(e_1) = \psi(e_1), \phi(e_2) = \psi(e_2) \text{ and } \phi(e_3) = \psi(e_3)$$

for some set of non-collinear events e_1, e_2, and e_3, then $\phi = \psi$.

Proof. It is not difficult to show that the events $\phi(e_1)$, $\phi(e_2)$, and $\phi(e_3)$ are also non-collinear.

Suppose that $\phi \neq \psi$. Then, there is an event e such that $\phi(e) \neq \psi(e)$, and we can define a line $L = L_{\phi(e),\psi(e)}$. Note that

$$d^{\mathrm{II}}(\phi(e), \phi(e_1)) = d^{\mathrm{II}}(e, e_1) = d^{\mathrm{II}}(\psi(e), \psi(e_1)) = d^{\mathrm{II}}(\psi(e), \phi(e_1)),$$

i.e., $d^{\mathrm{II}}(\phi(e), \phi(e_1)) = d^{\mathrm{II}}(\psi(e), \phi(e_1))$. Therefore, $\phi(e_1) \in L$. Similarly, $\phi(e_2) \in L$ and $\phi(e_3) \in L$. Then the events $\phi(e_1)$, $\phi(e_2)$, and $\phi(e_3)$ are collinear, which is a contradiction. □

For $e_1 = (x_1, \tau_1), e_2 = (x_2, \tau_2) \in \mathbb{R}^{1,1}$, the Minkowski inner product is defined by

$$e_1 \cdot e_2 = x_1 x_2 - \tau_1 \tau_2.$$

The Minkowski inner product possesses similar properties as the ordinary inner product:

$$e_1 \cdot e_2 = e_2 \cdot e_1,$$

$$e_1 \cdot (e_2 + e_3) = e_1 \cdot e_2 + e_1 \cdot e_3.$$

Define the norm-square $\| \ \|^2$ on $\mathbb{R}^{1,1}$ by $\|e\|^2 = e \cdot e$. We can see that

$$d^{\mathrm{II}}(e_1, e_2) = \|(e_1 - e_2)\|^2.$$

If an event e satisfies $\|e\|^2 > 0$, $\|e\|^2 = 0$ or $\|e\|^2 < 0$, it is called *spacelike*, *lightlike*, or *timelike*, respectively. If two events e_1 and e_2 satisfy $e_1 \cdot e_2 = 0$, then they are said to be *orthogonal* to one another.

Theorem 5.3. *Let $\phi : \mathbb{R}^{1,1} \to \mathbb{R}^{1,1}$ be a surjective map with $\phi(\mathbf{0}) = \mathbf{0}$. The map ϕ is an isometry if and only if it preserves the Minkowski inner product, i.e.,*

$$\phi(e_1) \cdot \phi(e_2) = e_1 \cdot e_2$$

for any $e_1, e_2 \in \mathbb{R}^{1,1}$.

Proof. Assume first that ϕ is an isometry. Note that

$$\|\phi(e)\|^2 = d^{\mathrm{II}}(\phi(e), \mathbf{0}) = d^{\mathrm{II}}(\phi(e), \phi(\mathbf{0})) = d^{\mathrm{II}}(e, \mathbf{0}) = \|e\|^2$$

for every event e. From

$$\|e_1 - e_2\|^2 = (e_1 - e_2) \cdot (e_1 - e_2) = \|e_1\|^2 - 2e_1 \cdot e_2 + \|e_2\|^2,$$

$e_1 \cdot e_2 = \frac{1}{2}\left(\|e_1\|^2 + \|e_2\|^2 - \|e_1 - e_2\|^2\right)$. Hence,

$$\begin{aligned}
\phi(e_1) \cdot \phi(e_2) &= \frac{1}{2}\left(\|\phi(e_1)\|^2 + \|\phi(e_2)\|^2 - \|\phi(e_1) - \phi(e_2)\|^2\right) \\
&= \frac{1}{2}\left(\|e_1\|^2 + \|e_2\|^2 - d^{\mathrm{II}}(\phi(e_1), \phi(e_2))\right) \\
&= \frac{1}{2}\left(\|e_1\|^2 + \|e_2\|^2 - d^{\mathrm{II}}(e_1, e_2)\right) \\
&= \frac{1}{2}\left(\|e_1\|^2 + \|e_2\|^2 - \|e_1 - e_2\|^2\right) \\
&= e_1 \cdot e_2,
\end{aligned}$$

and therefore, ϕ preserves the Minkowski inner product.

Conversely, assume that ϕ preserves the Minkowski inner product. Then,

$$\begin{aligned}
d^{\mathrm{II}}(\phi(e_1), \phi(e_2)) &= \|\phi(e_1) - \phi(e_2)\|^2 \\
&= \phi(e_1) \cdot \phi(e_1) - 2\phi(e_1) \cdot \phi(e_2) + \phi(e_2) \cdot \phi(e_2) \\
&= e_1 \cdot e_1 - 2e_1 \cdot e_2 + e_2 \cdot e_2 \\
&= \|e_1 - e_2\|^2 = d^{\mathrm{II}}(e_1, e_2).
\end{aligned}$$

Therefore, ϕ is an isometry. □

Consider a map $b_\lambda : \mathbb{R}^{1,1} \to \mathbb{R}^{1,1}$, defined as follows:

$$b_\lambda(x, \tau) = (x \cosh\lambda + \tau \sinh\lambda, x \sinh\lambda + \tau \cosh\lambda),$$

where $\lambda \in \mathbb{R}$ and $(x, \tau) \in \mathbb{R}^{1,1}$. Let us show that this is an isometry. Since $\phi(\mathbf{0}) = \mathbf{0}$, it is sufficient to show that

$$b(e_1) \cdot b(e_2) = e_1 \cdot e_2$$

for any $e_1 = (x_1, \tau_1), e_2 = (x_2, \tau_2) \in \mathbb{R}^{1,1}$. Noting that $\cosh^2 \lambda - \sinh^2 \lambda = 1$,

$$\begin{aligned} b_\lambda(e_1) \cdot b_\lambda(e_2) &= (x_1 \cosh\lambda + \tau_1 \sinh\lambda)(x_2 \cosh\lambda + \tau_2 \sinh\lambda) \\ &\quad - (\tau_1 \cosh\lambda + x_1 \sinh\lambda)(\tau_2 \cosh\lambda + x_2 \sinh\lambda) \\ &= x_1 x_2 \cosh^2\lambda - x_1 x_2 \sinh^2\lambda - \tau_1\tau_2 \cosh^2\lambda + \tau_1\tau_2 \sinh^2\lambda \\ &= (x_1 x_2 - \tau_1\tau_2)(\cosh^2\lambda - \sinh^2\lambda) \\ &= x_1 x_2 - \tau_1\tau_2 = e_1 \cdot e_2. \end{aligned}$$

Hence, b_λ is an isometry, referred to as a *Lorentz boost*.[2] It is not difficult to show that (Exercise 5.5)

$$b_{\lambda_1+\lambda_2} = b_{\lambda_1} \circ b_{\lambda_2}.$$

Exercises

5.1. For some $a, b, c \in \mathbb{R}$ with $(a, b) \neq (0, 0)$, let

$$L = \{(x, \tau) \in \mathbb{R}^{1,1} \mid ax + b\tau = c\}.$$

Find some events $e_1, e_2 \in \mathbb{R}^{1,1}$ such that

$$L = L_{e_1,e_2}.$$

5.2. For two distinct events e_1 and e_2, show that

1. $$L_{e_1,e_2} = \{e \in \mathbb{R}^{1,1} \mid (e - u) \cdot v = 0\},$$

   where $u = \frac{1}{2}(e_1 + e_2)$ and $v = \frac{1}{2}(e_1 - e_2)$.
2. for any $u' \in L_{e_1,e_2}$ and $a \in \mathbb{R}$ with $a \neq 0$,

   $$L_{e_1,e_2} = \{e \in \mathbb{R}^{1,1} \mid (e - u') \cdot v' = 0\},$$

   where $v' = av$.

[2] A Lorentz boost is also called a *hyperbolic rotation*.

5.3. Show that a Euclidean reflection

$$\phi(x, \tau) = (\tau, x)$$

is not an isometry of $\mathbb{R}^{1,1}$.

5.4. Show that a Euclidean rotation

$$\phi(x, \tau) = (x \cos\theta - \tau \sin\theta, x \sin\theta + \tau \cos\theta)$$

is not an isometry of $\mathbb{R}^{1,1}$ unless $\theta = n\pi$ for some $n \in \mathbb{Z}$.

5.5. Show that a composition of two Lorentz boosts is a Lorentz boost. More precisely, show that

$$b_{\lambda_2} \circ b_{\lambda_1} = b_{\lambda_1+\lambda_2}.$$

5.6. Suppose that a map $\phi : \mathbb{R}^{1,1} \to \mathbb{R}^{1,1}$ preserves the Lorentz–Minkowski distance. Show that it is injective.

Note that the solution for a similar-looking Exercise 1.11 does not work here.

5.2 Relativistic Reflections

Since reflections play a key role in studying isometries of the Euclidean plane, it is natural to also consider reflections in the Lorentz–Minkowski plane. For a line L_{e_1,e_2}, let us denote by $\bar{r}_{L_{e_1,e_2}}$ (if it exists) the *relativistic reflection* of the Lorentz–Minkowski plane in the line L_{e_1,e_2}. We seek to define $\bar{r}_{L_{e_1,e_2}}$ such that it satisfies all the properties in Remark 5.4, similar to a reflection of the Euclidean plane.

Remark 5.4.

(a) $\bar{r}_{L_{e_1,e_2}}(e_1) = e_2$, and $\bar{r}_{L_{e_1,e_2}}(e_2) = e_1$.
(b) $\bar{r}_{L_{e_1,e_2}}(e) = e$ for each event $e \in L_{e_1,e_2}$.
(c) $\bar{r}^2_{L_{e_1,e_2}} = \mathrm{id}_{\mathbb{R}^{1,1}}$.

Let $e_1 = (0, 0)$ and $e_2 = (1, 1)$. Then, the events e_1 and e_2 belong to the line L_{e_1,e_2}. By property (a), $\bar{r}_{L_{e_1,e_2}}(e_1) = e_2$, and by property (b), $\bar{r}_{L_{e_1,e_2}}(e_1) = e_1$. Therefore, $e_1 = e_2$, but clearly $e_1 \neq e_2$. Therefore, a relativistic reflection in such a line cannot exist. We call the line described above *lightlike*, i.e., a line L_{e_1,e_2} such that $e_1, e_2 \in L_{e_1,e_2}$. It is not difficult to see that L_{e_1,e_2} is lightlike if and only if $e_1 - e_2$ is lightlike (Lemma 5.6). For a non-lightlike line $L = L_{e_1,e_2}$, define the relativistic reflection as follows:

$$\bar{r}_L(e) = e - \frac{2(e-u)\cdot v}{\|v\|^2} v,$$

where $u = \frac{1}{2}(e_1 + e_2)$ and $v = \frac{1}{2}(e_1 - e_2)$ (see Exercise 1.14). We call v a *normal event* of the line L.

Theorem 5.5. *The relativistic reflection $\bar{r}_{L_{e_1,e_2}}$ is an isometry of the Lorentz–Minkowski plane, and it satisfies the properties in Remark 5.4.*

Proof. First, for two events p_1 and p_2,

$$\begin{aligned} d^{\mathrm{II}}(\bar{r}_{L_{e_1,e_2}}(p_1), \bar{r}_{L_{e_1,e_2}}(p_2)) &= \|\bar{r}_{L_{e_1,e_2}}(p_1) - \bar{r}_{L_{e_1,e_2}}(p_2)\|^2 \\ &= \left\| p_1 - p_2 - \frac{2(p_1 - p_2)\cdot v}{\|v\|^2} v \right\|^2 \\ &= \|p_1 - p_2\|^2 + \frac{4((p_1 - p_2)\cdot v)^2}{\|v\|^4}\|v\|^2 \\ &\quad - \frac{4((p_1 - p_2)\cdot v)^2}{\|v\|^2} \\ &= \|p_1 - p_2\|^2 \\ &= d^{\mathrm{II}}(p_1, p_2). \end{aligned}$$

Therefore, $\bar{r}_{L_{e_1,e_2}}$ is an isometry. Note that

$$\bar{r}_{L_{e_1,e_2}}(e_1) = e_1 - \frac{2(e_1 - u)\cdot v}{\|v\|^2} v = e_1 - \frac{2v\cdot v}{\|v\|^2} v = e_1 - 2v = e_2,$$

and similarly, $\bar{r}_{L_{e_1,e_2}}(e_2) = e_1$, which is property (a) in Remark 5.4.

If e lies on the line L_{e_1,e_2}, then

$$d^{\mathrm{II}}(e, e_1) = d^{\mathrm{II}}(e, e_2),$$

i.e.,

$$\|e - e_1\|^2 = \|e - e_2\|^2,$$

which implies that

$$\frac{1}{4}(\|e_1\|^2 - \|e_2\|^2) = \frac{1}{2}(e_1 - e_2)\cdot e,$$

i.e., $u \cdot v = v \cdot e$; thus, $(e - u)\cdot v = 0$. Hence,

$$\bar{r}_{L_{e_1,e_2}}(e) = e - \frac{2(e - u)\cdot v}{\|v\|^2} v = e,$$

showing property (b) in Remark 5.4.

Choose two distinct events p_1 and p_2 from L; then, the events e_1, p_1, and p_2 are non-collinear. Let $\phi = \bar{r}^{-1}_{L_{e_1,e_2}}$. It can be readily verified that

$$\phi(e_1) = \bar{r}_{L_{e_1,e_2}}(e_1), \phi(p_1) = \bar{r}_{L_{e_1,e_2}}(p_1), \phi(p_2) = \bar{r}_{L_{e_1,e_2}}(p_2).$$

According to Lemma 5.2, $\bar{r}^{-1}_{L_{e_1,e_2}} = \phi = \bar{r}_{L_{e_1,e_2}}$. This is property (c) in Remark 5.4. □

The set of fixed points of $\bar{r}_{L_{e_1,e_2}}$ is the following:

$$\{e \in \mathbb{R}^{1,1} \mid (e-u) \cdot v = 0\},$$

which is a line. Note that the line L_{e_1,e_2} is a subset of the set of fixed points of $\bar{r}_{L_{e_1,e_2}}$. Therefore (see also Exercise 5.2),

$$L_{e_1,e_2} = \{e \in \mathbb{R}^{1,1} \mid (e-u) \cdot v = 0\}.$$

Since $e_1 = u + v$ and $e_2 = u - v$, one can think that the events u and v determine the line. It is obvious that scaling v by a non-zero real number a does not change the line. Additionally, replacing u by any event in L does not change the line. It can be readily seen that the two lines L_1 and L_2, determined by u_1, v_1, and u_2, v_2, respectively, are parallel if and only if $v_1 = av_2$ for some non-zero real number a.

We summarize the properties of a lightlike line as follows.

Lemma 5.6. *The following statements are all equivalent for a line $L = L_{e_1,e_2}$.*

1. *L is a lightlike line.*
2. *$e_1 \in L$ or $e_2 \in L$.*
3. *$e_1 - e_2$ is a lightlike event.*
4. *For any two events $\alpha_1, \alpha_2 \in L$, $\alpha_1 - \alpha_2$ is lightlike.*
5. *There are some distinct events $\alpha_1, \alpha_2 \in L$ such that $\alpha_1 - \alpha_2$ is lightlike.*

Proof. Let

$$u = \frac{1}{2}(e_1 + e_2), \quad v = \frac{1}{2}(e_1 - e_2).$$

Then, by Exercise 5.2,

$$L_{e_1,e_2} = \{e \in \mathbb{R}^{1,1} \mid (e-u) \cdot v = 0\}.$$

$1 \Rightarrow 2$: It is trivial.
$2 \Rightarrow 3$: Assume that $e_1 \in L = L_{e_1,e_2}$. Then,

$$0 = d^{\mathrm{II}}(e_1, e_1) = d^{\mathrm{II}}(e_1, e_2) = \|e_1 - e_2\|^2,$$

and thus, $e_1 - e_2$ is lightlike.
$3 \Rightarrow 4$: Note that

$$(\alpha_i - u) \cdot v = 0,$$

where $u = \frac{e_1+e_2}{2}$ and $v = \frac{e_1-e_2}{2}$. Hence,

$$0 = (\alpha_1 - u) \cdot v - (\alpha_2 - u) \cdot v = (\alpha_1 - \alpha_2) \cdot v.$$

Since v is lightlike, $v = (a, \pm a)$ for some non-zero $a \in \mathbb{R}$. Let $\alpha_1 - \alpha_2 = (b, c)$; then, $ab - (\pm)ac = 0$. Therefore, $b = \pm c$, and

$$\alpha_1 - \alpha_2 = (\pm c, c)$$

is a lightlike event.

$4 \Rightarrow 5$: It is clear.

$5 \Rightarrow 1$: As in "$3 \Rightarrow 4$", we have

$$(\alpha_1 - \alpha_2) \cdot v = 0.$$

Since $\alpha_1 - \alpha_2$ is lightlike, $v = \frac{e_1-e_2}{2}$ is also lightlike. Hence,

$$d^{\mathrm{II}}(e_1, e_2) = \|e_1 - e_2\|^2 = 0.$$

Since

$$d^{\mathrm{II}}(e_1, e_1) = 0 = d^{\mathrm{II}}(e_1, e_2),$$

$e_1 \in L = L_{e_1,e_2}$. Similarly, $e_2 \in L$.

□

We prove a reflection theorem for the Lorentz–Minkowski plane.

Theorem 5.7 (Four reflections theorem for the Lorentz–Minkowski plane). *An isometry of the Lorentz–Minkowski plane is a composition of at most four relativistic reflections.*

Proof. Let $\phi : \mathbb{R}^{1,1} \to \mathbb{R}^{1,1}$ be an isometry.

Case 1. First, assume that there are non-collinear events e_1, e_2, and e_3 such that

$$\phi(e_1) = e_1, \phi(e_2) = e_2 \text{ and } \phi(e_3) = e_3.$$

According to Lemma 5.2, ϕ is the identity map, which is a composition of two relativistic reflections.

Case 2. Assume that there are distinct events e_1 and e_2 such that

$$\phi(e_1) = e_1, \phi(e_2) = e_2.$$

If $e_1 - e_2$ is lightlike, then choose an event e_3 such that the events e_1, e_2, and e_3 are not collinear. Suppose that $\phi(e_3) \neq e_3$, then we let $L = L_{e_3,\phi(e_3)}$. Since

$$d^{\mathrm{II}}(e_1, e_3) = d^{\mathrm{II}}(\phi(e_1), \phi(e_3)) = d^{\mathrm{II}}(e_1, \phi(e_3)),$$

$e_1 \in L$ and similarly $e_2 \in L$. Since $e_1 - e_2$ is lightlike, by Lemma 5.6, L is a lightlike line, and the event e_3 lies on L. In summary, the events e_1, e_2, and e_3 lie on the line L, which is a contradiction. Therefore,

$$\phi(e_3) = e_3,$$

resorting to Case 1. If $e_1 - e_2$ is not lightlike, consider a line L that contains both the events e_1 and e_2. It is elementary to verify that L is not a lightlike line (Lemma 5.6). Choose an event e_3 such that the events e_1, e_2, and e_3 are not collinear. If $\phi(e_3) = e_3$, we resort to Case 1. Otherwise, consider a line $L_{\phi(e_3),e_3}$. Since

$$e_1, e_2 \in L_{e_3,\phi(e_3)},$$

$$L_{e_3,\phi(e_3)} = L.$$

Therefore, $\bar{r}_L \circ \phi$ leaves the events e_1, e_2, and e_3 fixed, $\bar{r}_L \circ \phi$ is the identity, and ϕ is a relativistic reflection.

Case 3. Assume that there is an event e_1 such that $\phi(e_1) = e_1$. Choose another event e_2 such that $e_1 - e_2$ is not lightlike. If $\phi(e_2) = e_2$, we resort to Case 2. Otherwise, let

$$L = L_{\phi(e_2),e_2}.$$

Note that $e_1 \in L$. Suppose that L is a lightlike line, then $e_2, \phi(e_2) \in L$. However, this means that $e_1 - e_2$ is lightlike (Lemma 5.6), which is a contradiction. Therefore, L is not a lightlike line and we can consider the relativistic reflection $\bar{r}_L$ in the line L. Note that the map $\bar{r}_L \circ \phi$ leaves the events e_1 and e_2 fixed. Therefore, we can apply Case 2 and conclude that ϕ is a composition of at most two relativistic reflections.

Case 4. Assume that ϕ has no fixed events. Let $e_1 = \phi(\mathbf{0})$. Note that $e_1 \neq \mathbf{0}$. If e_1 is non-lightlike, let $L = L_{\mathbf{0},e_1}$. Let $\psi = \bar{r}_L \circ \phi$; then,

$$\psi(\mathbf{0}) = \bar{r}_L(e_1) = \mathbf{0}.$$

Applying Case 3, we conclude that ψ is a composition of at most two relativistic reflections. Therefore, $\phi = \bar{r}_L \circ \psi$ is a composition of at most three relativistic reflections. If e_1 is lightlike, we can let $e_1 = (a, \pm a)$ with $a \neq 0$. Let $e_2 = (a, 0)$; then, $e_2 - \mathbf{0}$ and $e_1 - e_2$ are non-lightlike. Let

$$L_1 = L_{e_1,e_2}, \quad L_2 = L_{\mathbf{0},e_2}$$

and $\psi' = \bar{r}_{L_2} \circ \bar{r}_{L_1} \circ \phi$. Then,

$$\psi'(\mathbf{0}) = (\bar{r}_{L_2} \circ \bar{r}_{L_1} \circ \phi)(\mathbf{0}) = (\bar{r}_{L_2} \circ \bar{r}_{L_1})(e_1) = \bar{r}_{L_2}(e_2) = \mathbf{0}.$$

Applying Case 3, we conclude that ψ' is a composition of at most two relativistic reflections. Therefore, $\phi = \bar{r}_{L_1} \circ \bar{r}_{L_2} \circ \psi'$ is a composition of at most four relativistic reflections. □

Consider a composition of two relativistic reflections $\bar{r}_{L_1}$ and $\bar{r}_{L_2}$. A property similar to that used in the Euclidean plane holds as follows.

Theorem 5.8. *The composition of two relativistic reflections in two non-lightlike parallel lines is a translation.*

Proof. Let L_1 and L_2 be two non-lightlike parallel lines and

$$\bar{r}_{L_i}(e) = e - \frac{2(e - u_i) \cdot v_i}{\|v_i\|^2} v_i$$

for $i = 1, 2$. Since the lines L_1 and L_2 are parallel, we can assume that $v_1 = v_2 = v$ and $u_2 = u_1 + av$ for some event v and real number a. Hence,

$$\begin{aligned}
\left(\bar{r}_{L_2} \circ \bar{r}_{L_1}\right)(e) &= \bar{r}_{L_2}\left(e - \frac{2(e - u_1) \cdot v_1}{\|v_1\|^2} v_1\right) \\
&= \left(e - \frac{2(e - u_1) \cdot v_1}{\|v_1\|^2} v_1\right) - \frac{2\left(\left(e - \frac{2(e-u_1)\cdot v_1}{\|v_1\|^2} v_1\right) - u_2\right) \cdot v_2}{\|v_2\|^2} v_2 \\
&= e - \frac{2(e - u_1) \cdot v}{\|v\|^2} v - \frac{2\left(\left(e - \frac{2(e-u_1)\cdot v}{\|v\|^2} v\right) - (u_1 + av)\right) \cdot v}{\|v\|^2} v \\
&= e - \frac{2(e - u_1) \cdot v}{\|v\|^2} v - \frac{2\left(e \cdot v - 2(e - u_1) \cdot v - (u_1 + av) \cdot v\right)}{\|v\|^2} v \\
&= e - \frac{2(e - u_1) \cdot v}{\|v\|^2} v - \frac{2\left(-e \cdot v + u_1 \cdot v - a\|v\|^2\right)}{\|v\|^2} v \\
&= e + 2av = t_{2av}(e).
\end{aligned}$$

Therefore, $\bar{r}_{L_2} \circ \bar{r}_{L_1} = t_{2av}$, which is a translation. □

For a non-zero event α, the set $\alpha^\perp$ is defined as follows:

$$\alpha^\perp = \{e \in \mathbb{R}^{1,1} \mid \alpha \cdot e = 0\}.$$

For a non-lightlike event v, let $L = v^{\perp}$ and $L' = t_{v/2}(L)$. Then, $t_v = \bar{r}_{L'} \circ \bar{r}_L$, which is a composition of two relativistic reflections. In contrast to the case of the Euclidean plane, some translations cannot be expressed as a composition of two relativistic reflections.

Theorem 5.9. *For a non-zero, lightlike event α, the translation t_α cannot be expressed as a composition of two relativistic reflections. It can be expressed as a composition of four relativistic reflections.*

Proof. Suppose that $t_\alpha = \bar{r}_{L'} \circ \bar{r}_L$ for some non-lightlike lines L and L'. If the lines L and L' are parallel with the non-lightlike normal event v, then $\bar{r}_{L'} \circ \bar{r}_L = t_{kv}$ for some real number k. Note that $\alpha = kv$. However, then, the lines L and L' are lightlike, which is impossible. If the lines L and L' are not parallel, then

$$L \cap L' = \{e\}$$

for some event e. Then,

$$e \neq e + \alpha = t_\alpha(e) = (\bar{r}_{L'} \circ \bar{r}_L)(e) = e,$$

which is a contradiction. Note that $\alpha = (a, \pm a)$ for some $a \neq 0$. Let $\alpha_1 = (a, 0)$ and $\alpha_2 = \alpha - \alpha_1$. Note that the events α_1 and α_2 are not lightlike. Hence, t_{α_i} can be expressed as a composition of two relativistic reflections. Since $t_\alpha = t_{\alpha_2} \circ t_{\alpha_1}$, the proof is complete. □

A composition of two Euclidean reflections of $\mathbb{R}^2$ in two crossing lines is a Euclidean rotation. We will show that a similar property holds in the Lorentz–Minkowski plane. We need to develop a concept of "angles" in the Lorentz–Minkowski plane.

Exercises

5.7. Show that the relativistic reflection $\bar{r}_L$ depends on the line L, not on the individual events e_1 and e_2. In concrete words, show that

$$e - \frac{2(e-u)\cdot v}{\|v\|^2} v = e - \frac{2(e-u')\cdot v'}{\|v'\|^2} v'$$

for every event e if the events e_1' and e_2' satisfy $L_{e_1,e_2} = L_{e_1',e_2'}$, where

$$u' = \frac{1}{2}(e_1' + e_2') \text{ and } v' = \frac{1}{2}(e_1' - e_2').$$

5.8. For a relativistic reflection $\bar{r}_L$ in a non-lightlike line L, show that

$$\bar{r}_{L+\alpha} = t_\alpha \circ \bar{r}_L \circ t_{-\alpha},$$

where $L + \alpha = t_\alpha(L)$.

5.9. Given the isometries ϕ and ψ of $\mathbb{R}^{1,1}$, the conjugation of ψ by ϕ is the isometry

$$^{\phi}\psi = \phi \circ \psi \circ \phi^{-1}.$$

(a) For relativistic reflections $\bar{r}_L$ and $\bar{r}_M$, show that

$$^{\bar{r}_M}\bar{r}_L = \bar{r}_{L'},$$

where $L' = \bar{r}_M(L)$.

(b) For an isometry ϕ, show that

$$^{\phi}\bar{r}_L = \bar{r}_{L'},$$

where $L' = \phi(L)$.

5.10. Let L be a timelike line and ϕ be an isometry. Show that the line $\psi(L)$ is also timelike.

5.11. Recall that two isometries ϕ, ϕ' are said to be conjugate if there is an isometry ψ such that $\phi = {}^{\psi}\phi'$ (Exercise 1.13). Show that there are some relativistic reflections $\bar{r}_L, \bar{r}_M$ that are *not* conjugate.

5.3 Hyperbolic Angle

Note that the Lorentz boost b_λ has a very similar property to rotation in the Euclidean plane (Exercise 5.5),

$$b_{\lambda_2} \circ b_{\lambda_1} = b_{\lambda_1+\lambda_2},$$

and $b_\lambda(\mathbf{0}) = \mathbf{0}$. Thus, it may be worth trying to develop concepts of angles and rotations in the Lorentz–Minkowski plane.

Definition 5.10. Let $e_1 = (x_1, \tau_1)$ and $e_2 = (x_2, \tau_2)$ be spacelike events with $x_1 > 0$ and $x_2 > 0$, and let $e_i' = \frac{e_i}{\sqrt{\|e_i\|^2}}$. Then, e_i' is on the unit hyperbola

$$x^2 - \tau^2 = 1.$$

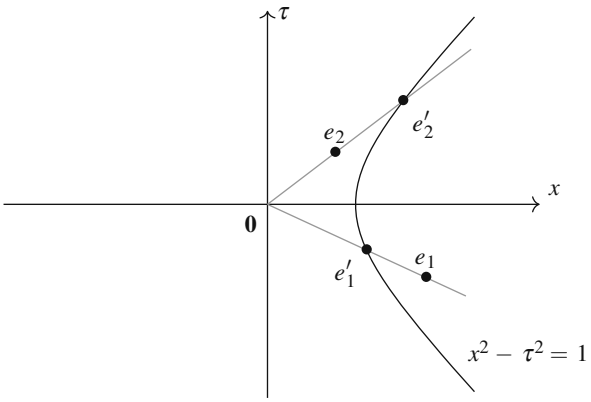


Fig. 5.2 Hyperbolic angle $\angle e_1\mathbf{0}e_2$ from e_1 to e_2 with respect to the origin $\mathbf{0}$

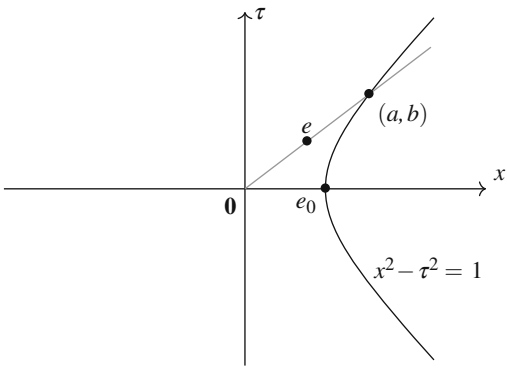


Fig. 5.3 $e = \sqrt{\|e\|^2}(a, b)$

Then, the *hyperbolic angle* $\angle e_1\mathbf{0}e_2$ from e_1 to e_2 with respect to the origin $\mathbf{0}$ is twice the signed Euclidean area of the sector on the unit hyperbola cut off from e_1' to e_2' (Figure 5.2). (Take a plus sign if $\tau_2' > \tau_1'$, and take a minus sign otherwise.)

Proposition 5.11. *For a spacelike event e with a positive x-coordinate,*

$$e = \sqrt{\|e\|^2}(\cosh\lambda, \sinh\lambda),$$

where $\lambda = \angle e_0\mathbf{0}e$ *with* $e_0 = (1, 0)$.

Proof. Let $e = \sqrt{\|e\|^2}(a, b)$; then, (a, b) lies on the curve $x^2 - \tau^2 = 1$ as shown in Figure 5.3. Let $\xi = \sinh^{-1}(b)$. Then, $b = \sinh(\xi)$ and

$$a = \sqrt{1 + b^2} = \sqrt{1 + \sinh^2\xi} = \cosh\xi.$$

For now, we assume $b \geq 0$. The hyperbolic angle $\lambda = \angle e_0\mathbf{0}e$ is twice the area surrounded by the curve $x^2 - \tau^2 = 1$ with $x > 0$, the x-axis and the line through the origin and the event e. Hence,

$$\begin{aligned}\lambda &= \angle e_0 \mathbf{0} e \\ &= 2\left(\frac{1}{2}ab - \int_1^a \sqrt{x^2-1}dx\right) \\ &= \cosh\xi \sinh\xi - 2\int_0^\xi \sqrt{\cosh^2 t - 1}\sinh t dt \\ &= \cosh\xi \sinh\xi + \xi - \frac{1}{2}\sinh 2\xi \\ &= \xi,\end{aligned}$$

i.e., $\xi = \lambda$. Therefore, $a = \cosh\lambda$, and $b = \sinh\lambda$. If $b < 0$, then the hyperbolic angle is simply the negative of the area; the other calculation is the same. □

Since the hyperbolic angle is defined by area, which is additive, the hyperbolic angle is also additive:

$$\angle e_1 \mathbf{0} e_3 = \angle e_1 \mathbf{0} e_2 + \angle e_2 \mathbf{0} e_3 \tag{5.1}$$

for spacelike events e_1, e_2, and e_3 with positive x-coordinates. Hyperbolic angles are closely related to hyperbolic functions. We summarize some of their properties, which can be readily shown by direct calculation.

$$\cosh\lambda = \frac{1}{2}(e^\lambda + e^{-\lambda}), \qquad \sinh\lambda = \frac{1}{2}(e^\lambda - e^{-\lambda}),$$

$$\tanh\lambda = \frac{\sinh\lambda}{\cosh\lambda} = \frac{\mathrm{e}^\lambda - \mathrm{e}^{-\lambda}}{\mathrm{e}^\lambda + \mathrm{e}^{-\lambda}},$$

$$\cosh^2\lambda - \sinh^2\lambda = 1,$$

$$|\sinh\lambda| < \cosh\lambda, \ \ |\tanh\lambda| < 1,$$

$$\sinh(\lambda \pm \lambda') = \sinh\lambda\cosh\lambda' \pm \cosh\lambda\sinh\lambda',$$

$$\cosh(\lambda \pm \lambda') = \cosh\lambda\cosh\lambda' \pm \sinh\lambda\sinh\lambda',$$

$$\tanh(\lambda \pm \lambda') = \frac{\tanh\lambda \pm \tanh\lambda'}{1 \pm \tanh\lambda\tanh\lambda'},$$

$$\sinh 2\lambda = 2\sinh\lambda\cosh\lambda,$$

$$\cosh 2\lambda = \cosh^2\lambda + \sinh^2\lambda,$$

$$\cosh^2 \lambda = \frac{\cosh 2\lambda + 1}{2},$$

$$\sinh^2 \lambda = \frac{\cosh 2\lambda - 1}{2}.$$

In general, a spacelike event e can be written as

$$e = \check{e}(\cosh \lambda, \sinh \lambda)$$

for some real number $\check{e}$ and λ. It is readily seen that these numbers, $\check{e}$ and λ, are unique (Exercise 5.12). For two spacelike events e_1 and e_2 with

$$e_1 = \check{e}_1(\cosh \lambda_1, \sinh \lambda_1), \quad e_2 = \check{e}_2(\cosh \lambda_2, \sinh \lambda_2),$$

the hyperbolic angle $\angle e_1 \mathbf{0} e_2$ is defined as follows:

$$\angle e_1 \mathbf{0} e_2 = \lambda_2 - \lambda_1.$$

For spacelike events e_1, e_2, and e_3, it is now readily seen that

$$\angle e_1 \mathbf{0} e_3 = \angle e_1 \mathbf{0} e_2 + \angle e_2 \mathbf{0} e_3.$$

We can similarly define the hyperbolic angle for two timelike events as follows:

$$e_1 = (x_1, \tau_1), e_2 = (x_2, \tau_2)$$

with $\tau_1 > 0$ and $\tau_2 > 0$ by using the hyperbola

$$x^2 - \tau^2 = -1.$$

We can verify

$$\angle e_1 \mathbf{0} e_2 = \angle \tilde{e}_1 \mathbf{0} \tilde{e}_2,$$

where $\tilde{e}_i = (\tau_i, x_i)$. The above formula can then be used as a definition of a hyperbolic angle for timelike events. Then, for a timelike event e,

$$e = \check{e}(\sinh \lambda, \cosh \lambda),$$

where $\lambda = \angle e_0 \mathbf{0} e$ with $e_0 = (0, 1)$. For a non-lightlike event e, the number $\check{e}$ is called the *signed relativistic norm* of e. If e is a lightlike event, we define $\check{e}$ to be zero. Note that

$$\check{e}^2 = \left| \|e\|^2 \right|.$$

For a timelike event e_1 and a spacelike event e_2, the hyperbolic angle between e_1 and e_2 with respect to the origin $\mathbf{0}$ is defined as follows:

$$\angle e_1\mathbf{0}e_2 = \angle \tilde{e}_1\mathbf{0}e_2 (= \angle e_1\mathbf{0}\tilde{e}_2).$$

For events

$$\begin{aligned}
e_1 &= (\cosh\lambda_1, \sinh\lambda_1),\\
e_2 &= (\cosh\lambda_2, -\sinh\lambda_2) = (\cosh(-\lambda_2), \sinh(-\lambda_2)),\\
e_3 &= (\sinh\lambda_3, \cosh\lambda_3),\\
e_4 &= (\sinh\lambda_4, -\cosh\lambda_4) = -(\sinh(-\lambda_4), \cosh(-\lambda_4)),
\end{aligned}$$

the hyperbolic angles between them are as follows,

$$\begin{aligned}
\angle e_1\mathbf{0}e_2 &= -\lambda_2 - \lambda_1,\\
\angle e_1\mathbf{0}e_3 &= \angle e_1\mathbf{0}\tilde{e}_3 = \lambda_3 - \lambda_1,\\
\angle e_1\mathbf{0}e_4 &= \angle e_1\mathbf{0}\tilde{e}_4 = -\lambda_4 - \lambda_1,\\
\angle e_2\mathbf{0}e_3 &= \angle e_2\mathbf{0}\tilde{e}_3 = \lambda_3 - (-\lambda_2) = \lambda_3 + \lambda_2,\\
\angle e_2\mathbf{0}e_4 &= \angle e_1\mathbf{0}\tilde{e}_4 = -\lambda_4 - (-\lambda_2) = -\lambda_4 + \lambda_2,\\
\angle e_3\mathbf{0}e_4 &= \angle \tilde{e}_1\mathbf{0}\tilde{e}_4 = -\lambda_4 - \lambda_3.
\end{aligned}$$

We propose a formula for the Minkowski inner product, similar to that of the Euclidean inner product.

Theorem 5.12. *For two non-lightlike events e_1 and e_2, let $\lambda = \angle e_1\mathbf{0}e_2$.*

1. *If both the events e_1 and e_2 are spacelike, then*

$$e_1 \cdot e_2 = \check{e}_1\check{e}_2 \cosh\lambda.$$

2. *If both the events e_1 and e_2 are timelike, then*

$$e_1 \cdot e_2 = -\check{e}_1\check{e}_2 \cosh\lambda.$$

3. *If the event e_1 is spacelike and the event e_2 is timelike, then*

$$e_1 \cdot e_2 = \check{e}_1\check{e}_2 \sinh\lambda.$$

Proof. Assume that both events e_1 and e_2 are spacelike. (The proof for the timelike events is very similar, and we will omit it.) Note that

$$e_i = \check{e}_i(\cosh\lambda_i, \sinh\lambda_i)$$

for some $\lambda_i \in \mathbb{R}$. Then $\lambda_2 - \lambda_1 = \lambda$.

Therefore,

$$\begin{aligned} e_1 \cdot e_2 &= \check{e}_1\check{e}_2(\cosh\lambda_1 \cosh\lambda_2 - \sinh\lambda_1 \sinh\lambda_2) \\ &= \check{e}_1\check{e}_2 \cosh(\lambda_2 - \lambda_1) \\ &= \check{e}_1\check{e}_2 \cosh\lambda. \end{aligned}$$

Suppose that e_1 is spacelike and e_2 is timelike. Then,

$$e_1 = \check{e}_1(\cosh\lambda_1, \sinh\lambda_1), \quad e_2 = \check{e}_2(\sinh\lambda_2, \cosh\lambda_2)$$

for some $\lambda_i \in \mathbb{R}$ and $\lambda_2 - \lambda_1 = \lambda$.

$$\begin{aligned} e_1 \cdot e_2 &= \check{e}_1\check{e}_2(\cosh\lambda_1 \sin\lambda_2 - \sinh\lambda_1 \cosh\lambda_2) \\ &= \check{e}_1\check{e}_2 \sinh(\lambda_2 - \lambda_1) \\ &= \check{e}_1\check{e}_2 \sinh\lambda. \end{aligned}$$

□

We note a direct consequence of this theorem.

Corollary 5.13. *If both events e_1 and e_2 are spacelike or timelike, then*

$$|e_1 \cdot e_2| \geq |\check{e}_1\check{e}_2|.$$

The following formula holds for all non-lightlike events e_1, e_2, and e_3 (Exercise 5.15):

$$\angle e_1\mathbf{0}e_3 = \angle e_1\mathbf{0}e_2 + \angle e_2\mathbf{0}e_3. \tag{5.2}$$

Suppose that both the events e_1 and e_2 are spacelike or timelike. We note that $\angle e_1\mathbf{0}e_2 = 0$ if and only if the events e_1 and e_2 lie on a line through the origin (Exercise 5.16).

For two events e_1 and e_2 that are different from an event e_0, the hyperbolic angle $\angle e_1e_0e_2$ from e_1 to e_2 with respect to e_0 is defined as follows:

$$\angle e_1e_0e_2 = \angle e_1'\mathbf{0}e_2',$$

where $e_1' = e_1 - e_0$ and $e_2' = e_2 - e_0$ (Figure 5.4).

Consider a line

$$L = \{e \in \mathbb{R}^{1,1} \mid (e - u) \cdot v = 0\}$$

with a normal event v. It is not difficult to show that

$$L = \{u + t\tilde{v} \mid t \in \mathbb{R}\}.$$

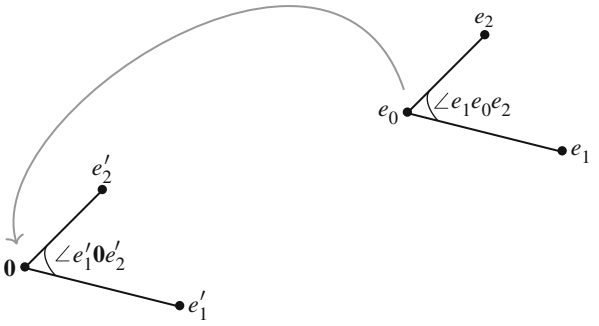


Fig. 5.4 Hyperbolic angle $\angle e_1e_0e_2$ from e_1 to e_2 with respect to e_0

In general, for a line $L = \{u + tw \mid t \in \mathbb{R}\}$, where w is a non-zero event, w is called a *directional event* of the line L. We call a line L *spacelike* (or *timelike*) if it has a spacelike directional event (or timelike directional event). Note that a line is spacelike if and only if it has a timelike normal event.

Let L and L' be lines with directional events w and w', respectively. The hyperbolic angle between L and L' is defined as the angle between w and w' as follows:

$$\angle LL' = \angle w\mathbf{0}w'.$$

It is not difficult to see that

$$\angle LL' = \angle v\mathbf{0}v',$$

where v and v' are normal events of L and L', respectively.

Exercises

5.12. For real numbers a, a', λ and λ' with $a, a' \neq 0$, assume that

$$a(\cosh\lambda, \sinh\lambda) = a'(\cosh\lambda', \sinh\lambda').$$

Show that $a = a'$ and $\lambda = \lambda'$.

5.13. For a spacelike event

$$e_1 = \check{e}_1(\cosh\lambda_1, \sinh\lambda_1)$$

and a timelike event

$$e_2 = \check{e}_2(\sinh\lambda_2, \cosh\lambda_2),$$

show that

$$b_\lambda(e_1) = \check{e}_1(\cosh(\lambda_1 + \lambda), \sinh(\lambda_1 + \lambda))$$

and

$$b_\lambda(e_2) = \check{e}_2(\sinh(\lambda_2 + \lambda), \cosh(\lambda_2 + \lambda)).$$

5.14. For two spacelike events e_1 and e_2 with $\check{e}_1\check{e}_2 > 0$, show that

$$\sqrt{\|e_1 + e_2\|^2} \geq \sqrt{\|e_1\|^2} + \sqrt{\|e_2\|^2}.$$

This is referred to as *the reverse triangle inequality*.

5.15. For non-lightlike events e_1, e_2, and e_3, show that

$$\angle e_1\mathbf{0}e_3 = \angle e_1\mathbf{0}e_2 + \angle e_2\mathbf{0}e_3.$$

5.16. Suppose that both events e_1 and e_2 are spacelike or timelike. Show that $\angle e_1\mathbf{0}e_2 = 0$ if and only if the events e_1 and e_2 lie on the same line through the origin.

5.17. Let e_1 and e_2 be non-lightlike orthogonal events. Prove that one of them is spacelike and that the other is timelike.

5.4 Relativistic Rotations

Figure 5.5 shows how the Lorentz boost b_λ acts on the Lorentz–Minkowski plane.

Lemma 5.14. *Let $\phi : \mathbb{R}^{1,1} \to \mathbb{R}^{1,1}$ be an isometry with $\phi(\mathbf{0}) = \mathbf{0}$. Suppose*

$$\angle e\mathbf{0}e' = 0$$

for every non-lightlike event e, where $e' = \phi(e)$. Then, $\phi = \mathrm{id}_{\mathbb{R}^{1,1}}$ or $-\mathrm{id}_{\mathbb{R}^{1,1}}$.

Proof. According to Theorem 5.3, ϕ preserves the Minkowski inner product. Let e be a non-lightlike event. Note that $\phi(e) = ae$ for some real number a. From

$$\|e\|^2 = e \cdot e = \phi(e) \cdot \phi(e) = a^2 e \cdot e = a^2\|e\|^2,$$

we conclude that $a^2 = 1$, i.e., $a = \pm 1$. Assume that $\phi \neq \mathrm{id}_{\mathbb{R}^{1,1}}$. Note that the three events $\mathbf{0}$, $(1, 0)$, and $(2, 1)$ are non-collinear. Hence, $\phi(1, 0) \neq (1, 0)$ or $\phi(2, 1) \neq (2, 1)$. If $\phi(1, 0) \neq (1, 0)$, then $\phi(1, 0) = -(1, 0)$. Let $\phi(2, 1) = a(2, 1)$. Since

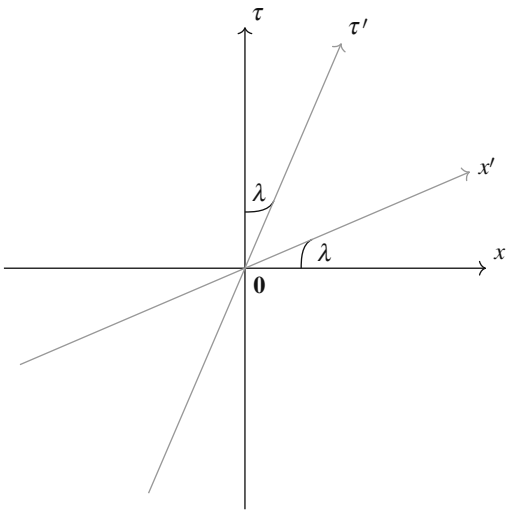


Fig. 5.5 Lorentz boost b_λ rotates the Lorentz–Minkowski plane "relativistically"

$$2 = (1, 0) \cdot (2, 1) = \phi(1, 0) \cdot \phi(2, 1) = (-1, 0) \cdot a(2, 1) = -2a,$$

$a = -1$. Since the images of the three events under ϕ and $-\mathrm{id}_{\mathbb{R}^{1,1}}$ match, we conclude that $\phi = -\mathrm{id}_{\mathbb{R}^{1,1}}$ according to Lemma 5.2. Additionally, in the case where $\phi(2, 1) \neq (2, 1)$, we can show the same result using similar arguments. □

Note that

$$\angle e\mathbf{0}e' = \lambda = \angle e\mathbf{0}e''$$

for every non-lightlike event e (Exercise 5.13), where $e' = b_\lambda(e)$ and $e'' = -b_\lambda(e)$.

Theorem 5.15. *Let $\phi : \mathbb{R}^{1,1} \to \mathbb{R}^{1,1}$ be an isometry with $\phi(\mathbf{0}) = \mathbf{0}$ and λ be a real number. Suppose that*

$$\angle e\mathbf{0}e' = \lambda$$

for every non-lightlike event e, where $e' = \phi(e)$. Then, $\phi = b_\lambda$ or $-b_\lambda$.

Proof. Let $\psi = b_{-\lambda} \circ \phi$. Using (5.2) and the result from Exercise 5.13,

$$\angle e\mathbf{0}\psi(e) = \angle e\mathbf{0}b_{-\lambda}(\phi(e)) = \angle e\mathbf{0}\phi(e) + \angle \phi(e)\mathbf{0}b_{-\lambda}(\phi(e)) = \lambda - \lambda = 0.$$

Hence, ψ satisfies the condition in Lemma 5.14, and

$$\psi = \mathrm{id}_{\mathbb{R}^{1,1}} \text{ or } -\mathrm{id}_{\mathbb{R}^{1,1}},$$

which implies that $\phi = b_\lambda$ or $-b_\lambda$. □

Let $\bar{b}_\lambda = -b_\lambda$. We call $\bar{b}_\lambda$ an *antipodal Lorentz boost*. Let

$$b_{\alpha,\lambda} = t_\alpha \circ b_\lambda \circ t_{-\alpha}, \quad \bar{b}_{\alpha,\lambda} = t_\alpha \circ \bar{b}_\lambda \circ t_{-\alpha}.$$

Let $\phi : \mathbb{R}^{1,1} \to \mathbb{R}^{1,1}$ be an isometry that fixes an event α and λ be a real number. According to Theorem 5.15,

$$\angle e\alpha e' = \lambda$$

for every event e with $e - \alpha$ being non-lightlike if and only if $\phi = b_{\alpha,\lambda}$ or $\bar{b}_{\alpha,\lambda}$, where $e' = \phi(e)$ (Exercise 5.18). In other words, $b_{\alpha,\lambda}$ and $\bar{b}_{\alpha,\lambda}$ "rotate" events around the event α by a hyperbolic angle λ. This is the motivation for the following definition.

Definition 5.16. The isometries $b_{\alpha,\lambda}$ and $\bar{b}_{\alpha,\lambda}$ of $\mathbb{R}^{1,1}$ are called *relativistic rotations* around the event α by the hyperbolic angle λ.

With some calculations, we can find the concrete form of the relativistic reflection in a non-lightlike line L that passes through the origin $\mathbf{0}$. Note that $L = v^\perp$, where v is a normal event of L. We choose the normal event v such that $\check{v} = 1$.

First, assume that L is a timelike line. Then, v is spacelike, and

$$v = (\cosh\lambda, \sinh\lambda),$$

where $\lambda = \angle e_0 \mathbf{0} v$ with $e_0 = (1, 0)$. Hence,

$$\begin{aligned}
\bar{r}_L(x, \tau) &= (x, \tau) - \frac{2(x, \tau) \cdot v}{\|v\|^2} v \\
&= (x, \tau) - 2(xa - \tau b)(a, b) \\
&= ((1 - 2a^2)x + 2ab\tau, -2abx + (1 + 2b^2)\tau \\
&= (x(1 - 2\cosh^2\lambda) + 2\tau\cosh\lambda\sinh\lambda, -2x\cosh\lambda\sinh\lambda \\
&\quad + \tau(1 + 2\sinh^2\lambda)) \\
&= (-x\cosh 2\lambda + \tau\sinh 2\lambda, -x\sinh 2\lambda + \tau\cosh 2\lambda),
\end{aligned}$$

where $(a, b) = v$.

Second, if L is spacelike, then v is timelike, and

$$v = (\sinh\lambda, \cosh\lambda),$$

where $\lambda = \angle e_1 \mathbf{0} v$ with $e_1 = (0, 1)$. Hence,

$$\begin{aligned}
\bar{r}_L(x, \tau) &= (x, \tau) - \frac{2(x, \tau) \cdot v}{\|v\|^2} v \\
&= (x, \tau) + 2(xa - \tau b)(a, b)
\end{aligned}$$

$$= ((1+2a^2)x - 2ab\tau, 2abx + (1-2b^2)\tau)$$
$$= (x(1+2\sinh^2\lambda) - 2\tau\sinh\lambda\cosh\lambda, 2x\sinh\lambda\cosh\lambda + \tau(1-2\cosh^2\lambda))$$
$$= (x\cosh 2\lambda - \tau\sinh 2\lambda, x\sinh 2\lambda - \tau\cosh 2\lambda).$$

For a line $L = v^\perp$ with a non-lightlike event v, let $\widetilde{L} = \tilde{v}^\perp$. Note also that $\bar{r}_L = -\bar{r}_{\widetilde{L}}$ and that the lines L and $\widetilde{L}$ are orthogonal, i.e., their normal events are orthogonal.

Now consider a composition $\phi = \bar{r}_{L'} \circ \bar{r}_L$ of relativistic reflections in two lines L and L' that pass through the origin $\mathbf{0}$.

Case 1. First, assume that the lines L and L' are timelike. Then,

$$L = (\cosh\lambda, \sinh\lambda)^\perp, L' = (\cosh\lambda', \sinh\lambda')^\perp$$

for some λ and λ'. Hence,

$$\bar{r}_L(x,\tau) = (-x\cosh 2\lambda + \tau\sinh 2\lambda, -x\sinh 2\lambda + \tau\cosh 2\lambda),$$

$$\bar{r}_{L'}(x,\tau) = (-x\cosh 2\lambda' + \tau\sinh 2\lambda', -x\sinh 2\lambda' + \tau\cosh 2\lambda').$$

Accordingly,

$$\begin{aligned}
&\phi(x,\tau)\\
&\quad = (\bar{r}_{L'} \circ \bar{r}_L)(x,\tau)\\
&\quad = (x(\cosh 2\lambda\cosh 2\lambda' - \sinh 2\lambda\sinh 2\lambda') + \tau(-\cosh 2\lambda'\sinh 2\lambda + \cosh 2\lambda\sinh 2\lambda'),\\
&\qquad\quad x(-\cosh 2\lambda'\sinh 2\lambda + \cosh 2\lambda\sinh 2\lambda') + \tau(\cosh 2\lambda\cosh 2\lambda' - \sinh 2\lambda\sinh 2\lambda'))\\
&\quad = (x\cosh(2\lambda'-2\lambda) + \tau\sinh(2\lambda'-2\lambda), x\sinh(2\lambda'-2\lambda) + \tau\cosh(2\lambda'-2\lambda))\\
&\quad = b_{(2\lambda'-2\lambda)}(x,\tau).
\end{aligned}$$

Therefore, $\bar{r}_{L'} \circ \bar{r}_L = b_{2(\lambda'-\lambda)} = b_{2\angle LL'}$.

Case 2. If the lines L and L' are spacelike, then $\bar{r}_L = -\bar{r}_{\widetilde{L}}$ and $\bar{r}_{L'} = -\bar{r}_{\widetilde{L}'}$. Since $\widetilde{L}$ and $\widetilde{L}'$ are orthogonal to L and L', respectively,

$$\angle LL' = \angle\widetilde{L}\widetilde{L}'.$$

Therefore,

$$\phi = \bar{r}_{L'} \circ \bar{r}_L = (-1)^2\bar{r}_{\widetilde{L}'} \circ \bar{r}_{\widetilde{L}} = b_{2\angle\widetilde{L}\widetilde{L}'} = b_{2\angle LL'},$$

which is a Lorentz boost.

Case 3. Finally, if one of the lines L or L' is spacelike and the other is timelike, e.g., L is spacelike and L' is timelike, then from $\bar{r}_L = -\bar{r}_{\widetilde{L}}$,

$$\phi = \bar{r}_{L'} \circ \bar{r}_L = (-1)\bar{r}_{L'} \circ \bar{r}_{\widetilde{L}} = -b_{2\angle\widetilde{L}L'} = \bar{b}_{2\angle\widetilde{L}L'},$$

which is an antipodal Lorentz boost.

We obtain results for relativistic rotations similar to those in the Euclidean plane.

Theorem 5.17. *An isometry of $\mathbb{R}^{1,1}$ is a relativistic rotation around an event α if and only if it is a composition of two relativistic reflections $\bar{r}_{L_1}$ and $\bar{r}_{L_2}$ in lines L_1 and L_2, respectively, which pass through the event α.*

Proof. The previous discussion gives a proof for the case in which α is the origin. The proof for the general case is similar. Let M_1 and M_2 be spacelike lines through the origin such that $\angle M_1 M_2 = \frac{\lambda}{2}$. Noting that

$$t_\alpha \circ \bar{r}_{M_i} \circ t_{-\alpha} = \bar{r}_{t_\alpha(M_i)},$$

$$\begin{aligned} b_{\alpha,\lambda} &= t_\alpha \circ b_\lambda \circ t_{-\alpha} \\ &= t_\alpha \circ \bar{r}_{M_2} \circ \bar{r}_{M_1} \circ t_{-\alpha} \\ &= t_\alpha \circ \bar{r}_{M_2} \circ t_{-\alpha} \circ t_\alpha \circ \bar{r}_{M_1} \circ t_{-\alpha} \\ &= \bar{r}_{t_\alpha(M_2)} \circ \bar{r}_{t_\alpha(M_1)}, \end{aligned}$$

which is a composition of two relativistic reflections in lines $t_\alpha(M_1)$ and $t_\alpha(M_2)$ through the event α. Similarly,

$$\bar{b}_{\alpha,\lambda} = \bar{r}_{t_\alpha(M_2)} \circ \bar{r}_{t_\alpha(\widetilde{M}_1)}.$$

Conversely, for non-lightlike lines L_1 and L_2 that pass through an event α, each line $t_{-\alpha}(L_i)$ passes through the origin. Hence,

$$\begin{aligned} t_{-\alpha} \circ \bar{r}_{L_2} \circ \bar{r}_{L_1} \circ t_\alpha &= t_{-\alpha} \circ \bar{r}_{L_2} \circ t_\alpha \circ t_{-\alpha} \circ \bar{r}_{L_1} \circ t_\alpha \\ &= \bar{r}_{t_{-\alpha}(L_2)} \circ \bar{r}_{t_{-\alpha}(L_1)} \end{aligned}$$

is a relativistic rotation around the origin by the angle

$$\lambda = 2\angle t_{-\alpha}(L_1)\, t_{-\alpha}(L_2) = 2\angle L_1 L_2,$$

which is b_λ or $\bar{b}_\lambda$. Hence, $\bar{r}_{L_2} \circ \bar{r}_{L_1}$ is

$$t_\alpha \circ b_\lambda \circ t_{-\alpha} = b_{\alpha,\lambda}$$

or

$$t_\alpha \circ \bar{b}_\lambda \circ t_{-\alpha} = \bar{b}_{\alpha,\lambda},$$

which is a relativistic rotation around the event α. □

Exercises

5.18. Let $\phi : \mathbb{R}^{1,1} \to \mathbb{R}^{1,1}$ be an isometry that fixes an event α and λ be a real number. Show that

$$\angle e\alpha e' = \lambda$$

for every event e with $e - \alpha$ being non-lightlike if and only if $\phi = b_{\alpha,\lambda}$ or $\bar{b}_{\alpha,\lambda}$, where $e' = \phi(e)$.

5.19. Let ϕ be a relativistic rotation such that $\phi^n = \mathrm{id}_{\mathbb{R}^{1,1}}$ for a positive integer n. Show:

1. $\phi = \mathrm{id}_{\mathbb{R}^{1,1}}$ when n is odd and
2. $\phi = \mathrm{id}_{\mathbb{R}^{1,1}}$ or $\bar{b}_{\alpha,0}$ for some event α when n is even.

5.5 Matrix and Isometry

Let $\mathrm{Iso}^+(\mathbb{R}^{1,1})$ be the set of all compositions of even numbers of relativistic reflections, and let $\mathrm{Iso}^-(\mathbb{R}^{1,1})$ be the set of all compositions of odd numbers of relativistic reflections. An isometry in $\mathrm{Iso}^+(\mathbb{R}^{1,1})$ is said to be orientation-preserving, and an isometry in $\mathrm{Iso}^-(\mathbb{R}^{1,1})$ is said to be orientation-reversing. Note that $\mathrm{Iso}(\mathbb{R}^{1,1}) = \mathrm{Iso}^+(\mathbb{R}^{1,1}) \cup \mathrm{Iso}^-(\mathbb{R}^{1,1})$.

We will show that the sets $\mathrm{Iso}^+(\mathbb{R}^{1,1})$ and $\mathrm{Iso}^-(\mathbb{R}^{1,1})$ are disjoint. This could be done by using purely geometric arguments, considering various configurations of reflection lines, as in Chapter 1, although the argument can be more complicated because of the existence of lightlike lines, in which relativistic reflections cannot be defined. Instead, we will take a more algebraic approach that uses matrix computation.

For a 2×2 matrix

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix},$$

we define a map $T_A : \mathbb{R}^{1,1} \to \mathbb{R}^{1,1}$ as follows:

$$T_A(x, \tau) = (ax + b\tau, cx + d\tau).$$

Remark 5.18. We consider an event $e = (x, \tau)$ a 1×2 matrix. For a matrix M, denote its transpose by M^t. Then,

$$(T_A(x, \tau))^t = A \begin{pmatrix} x \\ \tau \end{pmatrix}$$

or

$$T_A(x, \tau) = \left(A \begin{pmatrix} x \\ \tau \end{pmatrix} \right)^t = (x, \tau) A^t.$$

Theorem 5.19. *For two 2×2 matrices A and B,*

$$T_A \circ T_B = T_{AB}.$$

Proof. For $(x, \tau) \in \mathbb{R}^{1,1}$,

$$\begin{aligned} (T_A \circ T_B)(x, \tau) &= T_A((x, \tau) B^t) \\ &= (x, \tau) B^t A^t \\ &= (x, \tau)(AB)^t \\ &= T_{AB}(x, \tau), \end{aligned}$$

and the proof is complete. □

The Lorentz boost is $b_\lambda = T_{R(\lambda)}$, where

$$R(\lambda) = \begin{pmatrix} \cosh\lambda & \sinh\lambda \\ \sinh\lambda & \cosh\lambda \end{pmatrix}.$$

The relation

$$b_{\lambda_1 + \lambda_2} = b_{\lambda_1} \circ b_{\lambda_2}$$

is from the relation $R(\lambda_1 + \lambda_2) = R(\lambda_1) R(\lambda_2)$. Note that $\bar{b}_\lambda = T_{-R(\lambda)}$. For a timelike line $L = v^\perp$ with a spacelike normal event

$$v = (\cosh\lambda, \sinh\lambda),$$

the relativistic reflection in L is $\bar{r}_L = T_{\Lambda(2\lambda)}$, where

$$\Lambda(\lambda) = \begin{pmatrix} -\cosh\lambda & \sinh\lambda \\ -\sinh\lambda & \cosh\lambda \end{pmatrix}.$$

Similarly, $\bar{r}_{\widetilde{L}} = T_{-\Lambda(2\lambda)}$. Some simple calculations lead to

$$\Lambda(\lambda_2)\Lambda(\lambda_1) = R(\lambda_2 - \lambda_1).$$

Let $J = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$. Note that the Minkowski inner product $e_1 \cdot e_2$ can be expressed by multiplications of matrices:

$$e_1 \cdot e_2 = e_1 J e_2^t.$$

A 2×2 matrix A is said to be *J-orthogonal* if $A^t J A = J$.

Theorem 5.20. *For a 2×2 matrix A, the map $T_A : \mathbb{R}^{1,1} \to \mathbb{R}^{1,1}$ is an isometry if and only if A is J-orthogonal.*

Proof. Note that $e_1 \cdot e_2 = e_1 J e_2^t$ for $e_1, e_2 \in \mathbb{R}^{1,1}$. Then,

$$\begin{aligned} T_A(e_1) \cdot T(e_2) &= T_A(e_1) J T_A(e_2)^t \\ &= e_1 A^t J A e_2^t. \end{aligned}$$

Hence, $T_A(e_1) \cdot T_A(e_2) = e_1 \cdot e_2$ if and only if $A^t J A = J$. According to Theorem 5.3, the proof is complete. □

For the equation $A^t J A = J$, taking the determinant on both sides yields

$$\det(A^t J A) = \det(A^t)\det(J)\det(A) = \det(A)\det(J)\det(A) = \det(J).$$

Therefore, $\det(A)^2 = 1$, and thus, $\det(A) = \pm 1$. Note that $\det(\pm R(\lambda)) = 1$ and $\det(\pm\Lambda(\lambda)) = -1$. The following lemma provides a complete description of J-orthogonal matrices.

Lemma 5.21. *If A is a J-orthogonal matrix, then $A = \pm R(\lambda)$ or $A = \pm\Lambda(\lambda)$ for some real number λ.*

Proof. Let

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

Since

$$A^t J A = \begin{pmatrix} a^2 - c^2 & ab - cd \\ ab - cd & b^2 - d^2 \end{pmatrix} = J,$$

$$a^2 - c^2 = 1, \quad b^2 - d^2 = -1, \quad ab - cd = 0.$$

First, assume that the (1, 1) entry of matrix A is non-negative, i.e., $a \geq 0$. Let $\lambda = \sinh^{-1} c$; then, $c = \sinh\lambda$, and

$$a = \sqrt{c^2 + 1} = \sqrt{\sinh^2\lambda + 1} = \cosh\lambda.$$

Let $\lambda' = \sinh^{-1} b$, i.e., $b = \sinh\lambda'$.

Case 1. Suppose that $d \geq 0$.

$$d = \sqrt{b^2 + 1} = \sqrt{\sinh^2\lambda' + 1} = \sinh\lambda'.$$

Hence,

$$\begin{aligned} 0 &= ab - cd \\ &= \cosh\lambda \sinh\lambda' - \cosh\lambda' \sinh\lambda \\ &= \sinh(\lambda' - \lambda). \end{aligned}$$

Therefore, $\lambda' = \lambda$, and $A = R(\lambda)$.

Case 2. Suppose that $d < 0$. Then,

$$d = -\sqrt{b^2 + 1} = -\sqrt{\sinh^2\lambda' + 1} = -\cosh\lambda'.$$

We know that

$$\begin{aligned} 0 &= ab - cd \\ &= \cosh\lambda \sinh\lambda' + \cosh\lambda' \sinh\lambda \\ &= \sinh(\lambda' + \lambda), \end{aligned}$$

and therefore, $\lambda' = -\lambda$. Hence,

$$\begin{aligned} A &= \begin{pmatrix} \cosh\lambda & \sinh(-\lambda) \\ \sinh\lambda & -\cosh(-\lambda) \end{pmatrix} \\ &= -\begin{pmatrix} -\cosh\lambda & \sinh\lambda \\ -\sinh\lambda & \cosh\lambda \end{pmatrix} \\ &= -\Lambda(\lambda). \end{aligned}$$

Second, assume $a < 0$. Let $B = -A$; then, B is a J-orthogonal matrix with a positive (1, 1) entry. Applying the previous arguments, we conclude that $B = R(\lambda)$ or $-\Lambda(\lambda)$ for some real number λ. Therefore, $A = -B = -R(\lambda)$ or $\Lambda(\lambda)$. □

Consider a relativistic reflection $\bar{r}_L$. If L passes through the origin, $\bar{r}_L = T_{\Lambda(\lambda)}$ or $T_{-\Lambda(\lambda)}$ for some λ, as previously shown. Note that $\Lambda(\lambda)$ and $-\Lambda(\lambda)$ are J-orthogonal matrices with determinant -1.

If L does not pass through the origin, consider $L' = t_{-\beta}(L)$, where β is an event on L. Note that $L = t_\beta(L')$ and (Exercise 5.8)

$$\bar{r}_L = t_\beta \circ \bar{r}_{L'} \circ t_{-\beta}.$$

Note that $\bar{r}_{L'} = T_A$ for some J-orthogonal matrix with determinant -1. Hence,

$$\begin{aligned}\bar{r}_L(e) &= (t_\beta \circ \bar{r}_{L'} \circ t_{-\beta})(e)\\ &= \beta + (e-\beta)A^t\\ &= \beta - \beta A^t + eA^t\\ &= \alpha + eA^t\\ &= (t_\alpha \circ T_A)(e),\end{aligned}$$

where $\alpha = \beta - \beta A^t$. Therefore, $\bar{r}_L = t_\alpha \circ T_A$ for event α and some J-orthogonal matrix A with $\det(A) = -1$.

Theorem 5.22. *A composition of n relativistic reflections can be expressed as follows:*

$$t_\alpha \circ T_A$$

for some event α and a J-orthogonal matrix A with $\det(A) = (-1)^n$. Such α and A are unique.

Proof. We will use an induction on n. When $n = 1$, it is already showed.

Assume that this statement holds when $n = k$, and consider the case $n = k + 1$. Note that

$$\bar{r}_{L_{k+1}} \circ \bar{r}_{L_k} \circ \cdots \circ \bar{r}_{L_1} = \bar{r}_{L_{k+1}} \circ t_\alpha \circ T_A$$

for some event α and some J-orthogonal matrix A with determinant $(-1)^k$. Note also that

$$\bar{r}_{L_{k+1}} = t_\beta \circ T_B$$

for some event β and some J-orthogonal matrix B with determinant -1. Hence, for every event e,

$$\begin{aligned}(\bar{r}_{L_{k+1}} \circ \bar{r}_{L_k} \circ \cdots \circ \bar{r}_{L_1})(e) &= (\bar{r}_{L_{k+1}} \circ t_\alpha \circ T_A)(e)\\ &= (t_\beta \circ T_B \circ t_\alpha \circ T_A)(e)\end{aligned}$$

$$
\begin{aligned}
&= (t_\beta \circ T_B)(\alpha + eA^t) \\
&= \beta + (\alpha + eA^t)B^t \\
&= \beta + \alpha B^t + eA^t B^t \\
&= \alpha' + e(BA)^t \\
&= \alpha' + eA'^t \\
&= (t_{\alpha'} \circ T_{A'})(e),
\end{aligned}
$$

where $A' = BA$ and $\alpha' = \beta + \alpha B^t$. Hence,

$$\bar{r}_{L_{k+1}} \circ \bar{r}_{L_k} \circ \cdots \circ \bar{r}_{L_1} = t_{\alpha'} \circ T_{A'}.$$

Note that

$$\det(A') = \det(B)\det(A) = (-1)^{k+1}.$$

Therefore, we showed that the statement holds for the case $n = k + 1$.

The uniqueness of α and A can be readily obtained (Exercise 5.20). □

Corollary 5.23. *An isometry of the Lorentz–Minkowski plane has the following form:*

$$t_\alpha \circ T_A$$

for some event α and a J-orthogonal matrix A. It is orientation-preserving (or orientation-reversing) if and only if $\det(A) = 1$ *(or* -1*).*

Proof. According to Theorem 5.7, the isometry is a composition of relativistic reflections. Theorem 5.22 implies that it has a form of

$$t_\alpha \circ T_A.$$

The second statement is also a result of Theorem 5.22. □

At this point, the fact that the sets $\mathrm{Iso}^+(\mathbb{R}^{1,1})$ and $\mathrm{Iso}^-(\mathbb{R}^{1,1})$ are disjoint is obvious.

Corollary 5.24. *Let A be a J-orthogonal matrix and α be an event. Let $\phi = t_\alpha \circ T_A$.*

1. *If* $\det(A) = 1$*, then ϕ is either a translation or a relativistic rotation.*
2. *If* $\det(A) = -1$*, then ϕ is either a relativistic reflection or a composition of three relativistic reflections.*

Proof. ϕ is an isometry of $\mathbb{R}^{1,1}$ and so it is a composition of at most four relativistic reflections.

If $\det(A) = -1$, then ϕ is a relativistic reflection or a composition of three relativistic reflections by Theorem 5.22.

Assume that $\det(A) = 1$. If A is the 2×2 identity matrix I_2, then $\phi = t_\alpha$ is a translation. Assume that $A \neq I_2$. Let us check whether there is an event β that satisfies $\phi(\beta) = \beta$, i.e.,

$$\beta A^t + \alpha = \beta$$

i.e.,

$$\beta(I_2 - A^t) = \alpha.$$

Noting that $A = \pm R(\lambda)$ for some $\lambda \in \mathbb{R}$ by Lemma 5.21 and $A \neq I_2$, it can be readily shown that

$$\det(I_2 - A^t) = \det(I_2 \mp R(\lambda)^t) = 2(1 \mp \cosh(\lambda)) \neq 0.$$

Thus, it has the inverse matrix $(I_2 - A^t)^{-1}$, and $\beta = \alpha(I_2 - A^t)^{-1}$; such an event β does therefore exist.

Let

$$\psi = t_{-\beta} \circ \phi \circ t_\beta = t_{-\beta} \circ t_\alpha \circ T_A \circ t_\beta.$$

For every event e,

$$\begin{aligned} \psi(e) &= (t_{-\beta} \circ t_\alpha \circ T_A \circ t_\beta)(e) \\ &= (e + \beta)A^t + \alpha - \beta \\ &= eA^t + \alpha - \beta(I_2 - A^t) \\ &= eA^t + \alpha - \alpha \\ &= T_A(e). \end{aligned}$$

Hence, $\psi = T_A$. Since $A = \pm R(\lambda)$,

$$\psi = \begin{cases} b_\lambda, & \text{if } A = R(\lambda); \\ \bar{b}_\lambda, & \text{if } A = -R(\lambda). \end{cases}$$

Finally,

$$\phi = t_\beta \circ \psi \circ t_{-\beta} = \begin{cases} t_\beta \circ b_\lambda \circ t_{-\beta} = b_{\beta,\lambda}, & \text{if } A = R(\lambda); \\ t_\beta \circ \bar{b}_\lambda \circ t_{-\beta} = \bar{b}_{\beta,\lambda}, & \text{if } A = -R(\lambda), \end{cases}$$

which is a relativistic rotation around the event β. □

Corollary 5.25. *For a non-zero, lightlike event α, the translation t_α cannot be expressed as a composition of less than four relativistic reflections.*

Proof. According to Theorem 5.9, t_α cannot be expressed as a composition of two relativistic reflections. According to Corollary 5.23, t_α is orientation-preserving. Hence, t_α cannot be expressed as a composition of three relativistic reflections, and it is not a relativistic reflection. This completes the proof. □

Exercises

5.20. For two events α, β and two 2×2 matrices A, B, suppose that

$$t_\alpha \circ T_A = t_\beta \circ T_B.$$

Show that

$$\alpha = \beta \text{ and } A = B.$$

5.21. Suppose that an isometry of $\mathbb{R}^{1,1}$ cannot be expressed as a composition of less than four relativistic reflections. Show that it is a translation in a lightlike direction.

5.22. Show that the set of translations and relativistic rotations is closed under composition.

5.23. For an orientation-reversing isometry ϕ of $\mathbb{R}^{1,1}$, show that ϕ^2 is a translation.

5.24. Classify all the isometries such that $\phi^2 = \phi \circ \phi = \mathrm{id}_{\mathbb{R}^{1,1}}$.

5.25. Suppose that an isometry $\phi : \mathbb{R}^{1,1} \to \mathbb{R}^{1,1}$ satisfies $\phi^6 = \mathrm{id}_{\mathbb{R}^{1,1}}$. Show that $\phi^2 = \mathrm{id}_{\mathbb{R}^{1,1}}$.

5.6 Relativistic Lengths of Curves

For curves in $\mathbb{R}^{1,1}$, we define the notion of "length." Before proceeding, we must be more precise about the definition of a curve in $\mathbb{R}^{1,1}$. By a curve, we mean a smooth curve, i.e., a curve with a smooth, regular parametrization. A parametrization $\gamma(t) = (x(t), \tau(t))$ is called *smooth* if all the functions $x(t)$ and $\tau(t)$ of t have derivatives of all finite orders; it is called *regular* if $\frac{d\gamma(t)}{dt} \neq \mathbf{0}$ for any t. Henceforth, a parametrization will be assumed to be smooth and regular. For a parameterized curve

$$\gamma : [a, b] \to \mathbb{R}^{1,1}, \quad \gamma(t) = (x(t), \tau(t)),$$

γ is called *spacelike* (*timelike*) if $\|\gamma'(t)\|^2 > 0$ (< 0) for each t. It might seem natural to use the following as the definition of length for curves:

$$\int_a^b d^{\mathrm{II}}\left(\frac{d\gamma(t)}{dt}, \mathbf{0}\right) dt.$$

Note that

$$\int_a^b d^{\mathrm{II}}\left(\frac{d\gamma(t)}{dt}, \mathbf{0}\right) dt = \int_a^b \left\|\frac{d\gamma(t)}{dt}\right\|^2 dt = \int_a^b \left(\left(\frac{dx(t)}{dt}\right)^2 - \left(\frac{d\tau(t)}{dt}\right)^2\right) dt.$$

Consider a spacelike curve

$$\gamma : [0, 1] \to \mathbb{R}^{1,1}, \quad \gamma(t) = (2t, t).$$

Then, its length would be

$$\int_0^1 \left\|\frac{d\gamma(t)}{dt}\right\|^2 dt = \int_0^1 3dt = 3.$$

However, if we use another parametrization

$$\delta : [0, 1] \to \mathbb{R}^{1,1}, \quad \delta(t) = (2t^2, t^2)$$

for the same curve, the length is different:

$$\int_0^1 \left\|\frac{d\delta(t)}{dt}\right\|^2 dt = \int_0^1 12t^2 dt = 4.$$

The correct definition requires some modification.

Definition 5.26. For a spacelike curve,

$$\gamma : [a, b] \to \mathbb{R}^{1,1}, \quad \gamma(t) = (x(t), \tau(t)),$$

its *relativistic length* is as follows:

$$l_R(\gamma) = \int_a^b \sqrt{\left\|\frac{d\gamma(t)}{dt}\right\|^2} dt = \int_a^b \sqrt{\left(\frac{dx(t)}{dt}\right)^2 - \left(\frac{d\tau(t)}{dt}\right)^2} dt.$$

For a timelike curve,

$$l_R(\gamma) = \int_a^b \sqrt{-\left\|\frac{d\gamma(t)}{dt}\right\|^2}\,dt = \int_a^b \sqrt{-\left(\frac{dx(t)}{dt}\right)^2 + \left(\frac{d\tau(t)}{dt}\right)^2}\,dt.$$

It can be readily verified that this definition does not depend on the parametrization (Exercise 5.27). The following theorem justifies Definition 5.26 further.

Theorem 5.27. *If Γ is a spacelike (or a timelike) curve of finite relativistic length, then*

$$l_R(\phi(\Gamma)) = l_R(\Gamma)$$

for every isometry ϕ of $\mathbb{R}^{1,1}$.

Proof. According to Corollary 5.23, $\phi = t_\alpha \circ T_A$ for some event α and a J-orthogonal matrix A. It is obvious that the translation t_α preserves the relativistic length. Hence, we must only show that $l_R(T_A(\Gamma)) = l_R(\Gamma)$. Let

$$\gamma : [a, b] \to \mathbb{R}^{1,1}, \quad \gamma(t) = (x(t), \tau(t))$$

be a parametrization of Γ and $\delta = T_A \circ \gamma$. Note that

$$\delta(t) = (A\gamma(t)^t)^t$$

and

$$\frac{d\delta(t)}{dt} = \left(A\frac{d\gamma(t)^t}{dt}\right)^t = \frac{d\gamma(t)}{dt}A^t.$$

Hence,

$$\begin{aligned}\left\|\frac{d\delta(t)}{dt}\right\|^2 &= \frac{d\delta(t)}{dt}\cdot\frac{d\delta(t)}{dt}\\ &= \frac{d\delta(t)}{dt}J\frac{d\delta(t)^t}{dt}\\ &= \frac{d\gamma(t)}{dt}A^tJA\frac{d\gamma(t)^t}{dt}\\ &= \frac{d\gamma(t)}{dt}J\frac{d\gamma(t)^t}{dt}\\ &= \frac{d\gamma(t)}{dt}\cdot\frac{d\gamma(t)}{dt} = \left\|\frac{d\gamma(t)}{dt}\right\|^2,\end{aligned}$$

and therefore,

$$l_R(\phi(\Gamma)) = l_R(\delta) = \int_a^b \sqrt{\left\| \frac{d\delta(t)}{dt} \right\|^2} dt = \int_a^b \sqrt{\left\| \frac{d\gamma(t)}{dt} \right\|^2} dt = l_R(\gamma) = l_R(\Gamma).$$

A similar argument works for the case where Γ is a timelike curve. □

Let us consider a hyperbolic curve:

$$\mathbb{U}^1 = \{(x, \tau) \in \mathbb{R}^{1,1} \mid \tau = \sqrt{1 + x^2}\}.$$

It is one of two components of the curve defined by

$$\|e\|^2 = -1.$$

Hence, $\mathbb{U}^1$ may be thought of as a relativistic "circle" of squared radius -1 or of radius $\sqrt{-1}$.

$\mathbb{U}^1$ can be parameterized by

$$\gamma : \mathbb{R} \to \mathbb{U}^1, \quad \gamma(t) = (\sinh t, \cosh t). \tag{5.3}$$

Note that $\mathbb{U}^1$ is a spacelike curve and that each event in $\mathbb{U}^1$ is timelike. For two events e_1 and e_2 on $\mathbb{U}^1$, we define the Lorentz–Minkowski distance $d_{\mathbb{U}}(e_1, e_2)$ between e_1 and e_2 by the relativistic length along $\mathbb{U}^1$ from e_1 to e_2. Let $e_i = (\sinh \lambda_i, \cosh \lambda_i)$ for some λ_i; then, it is readily seen (Figure 5.6) that

$$d_{\mathbb{U}}(e_1, e_2) = |\lambda_1 - \lambda_2| = |\angle e_1 \mathbf{0} e_2|.$$

Using $-e_1 \cdot e_2 = \cosh \lambda$ with $\lambda = \angle e_1 \mathbf{0} e_2$ (Theorem 5.12) yields the following proposition.

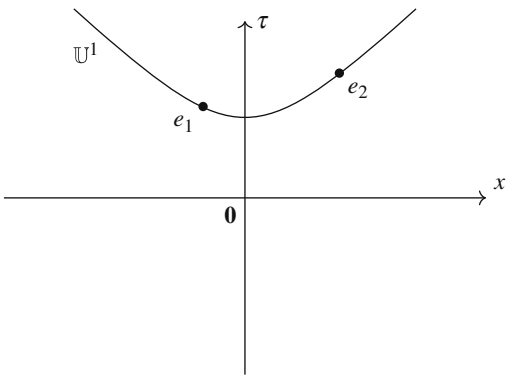


Fig. 5.6 $d_{\mathbb{U}}(e_1, e_2) = |\lambda_1 - \lambda_2| = |\angle e_1 \mathbf{0} e_2|$

Proposition 5.28. $\cosh d_{\mathbb{U}}(e_1, e_2) = -e_1 \cdot e_2$.

Note that the parametrization γ is an isometry from $\mathbb{R}$ to $\mathbb{U}^1$. (An isometry of $\mathbb{R}$ is defined in Exercise 1.27.) An isometry of $\mathbb{R}^{1,1}$ is called a *Lorentz transformation* if it fixes the origin; the set of them is denoted by $\mathrm{O}(1, 1)$. If a Lorentz transformation maps $\mathbb{U}^1$ onto $\mathbb{U}^1$, we call it *orthochronous* and denote the set of these by $\mathrm{O}^+(1, 1)$. Note that the restriction of an element ϕ of $\mathrm{O}^+(1, 1)$

$$\phi|_{\mathbb{U}^1} : \mathbb{U}^1 \to \mathbb{U}^1$$

is an isometry of $\mathbb{U}^1$. The following is a relativistic version of Theorem 2.18.

Theorem 5.29. *The restriction map*

$$\mathrm{O}^+(1, 1) \to \mathrm{Iso}(\mathbb{U}^1),$$

given by

$$\phi \mapsto \phi|_{\mathbb{U}^1},$$

is bijective.

Proof. We will build the inverse map of the map in the statement. For each isometry ϕ of $\mathbb{U}^1$, we must find a map from $\mathrm{O}^+(1, 1)$ whose restriction to $\mathbb{U}^1$ is ϕ. Since $\mathbb{R}$ and $\mathbb{U}^1$ are isometric and an isometry of $\mathbb{R}$ is a composition of reflections (Exercise 1.27), we can assume that ϕ is a reflection of $\mathbb{U}^1$ (fixing one single event α and moving other events to different events). Let L be a line that passes through $\mathbf{0}$ and α; then, L is timelike, and it is not difficult to see that $\bar{r}_L|_{\mathbb{U}^1} = \phi$ (Exercise 5.28). Note that $\bar{r}_L$ belongs to $\mathrm{O}^+(1, 1)$. Hence, we found such a map. □

Exercises

5.26. Let $\gamma : [a, b] \to \mathbb{R}^{1,1}$ be a spacelike (timelike) curve and $f : [c, d] \to [a, b]$ be a bijective differentiable function. Then,

$$\delta = \gamma \circ f : [c, d] \to \mathbb{R}^{1,1}$$

is a reparametrization of the curve. Show that the parametrization δ is spacelike (timelike).

5.27. Let $\gamma : [a, b] \to \mathbb{R}^{1,1}$ be a spacelike or timelike curve and $f : [c, d] \to [a, b]$ be a bijective increasing differentiable function. Then,

$$\delta = \gamma \circ f : [c, d] \to \mathbb{R}^{1,1}$$

is a reparametrization of the curve. Show that their relativistic lengths are the same, i.e.,

$$l_R(\gamma) = l_R(\delta).$$

5.28. Consider the map $\gamma : \mathbb{R} \to \mathbb{U}^1$ in (5.3) with an event $\alpha \in \mathbb{U}^1$. Assume that $\gamma(a) = \alpha$. Define a reflection $\bar{r} : \mathbb{R} \to \mathbb{R}$ by $\bar{r}(t) = 2a - t$. Consider a relativistic reflection $\bar{r}_L$ of $\mathbb{R}^{1,1}$, where L is a line on $\mathbb{R}^{1,1}$ that goes through $\mathbf{0}$ and α. Show that

$$\bar{r}_L(e) = (\gamma \circ \bar{r} \circ \gamma^{-1})(e)$$

for each event $e \in \mathbb{U}^1$.

5.7 Hyperboloid in $\mathbb{R}^{2,1}$

In this section, we will see that hyperbolic geometry and relativistic geometry are intrinsically related. Let us consider three-dimensional space with the Minkowski inner product

$$e_1 \cdot e_2 = x_1x_2 + y_1y_2 - \tau_1\tau_2$$

for events $e_i = (x_i, y_i, \tau_i) \in \mathbb{R}^3$. Then, the Lorentz–Minkowski distance $d^{\mathrm{II}}(e_1, e_2)$ is defined by

$$d^{\mathrm{II}}(e_1, e_2) = \|e_1 - e_2\|^2 = (e_1 - e_2) \cdot (e_1 - e_2),$$

and we denote the set $\mathbb{R}^3$ with the Lorentz–Minkowski distance by $\mathbb{R}^{2,1}$, which represents a three-dimensional Lorentz–Minkowski space. The elements of $\mathbb{R}^{2,1}$ are also called events. A bijective map $\phi : \mathbb{R}^{2,1} \to \mathbb{R}^{2,1}$ is called an isometry of $\mathbb{R}^{2,1}$ if it preserves the Lorentz–Minkowski distance. We denote by $\mathrm{Iso}(\mathbb{R}^{2,1})$ the set of all such isometries and by $\mathrm{O}(2, 1)$ the set of all isometries that fix the origin. All other notations can be similarly defined. For each event $e \in \mathbb{R}^{2,1}$, if $\|e\|^2 > 0$ (resp. $\|e\|^2 = 0$, $\|e\|^2 < 0$), then the event e is said to be *spacelike* (resp. *lightlike*, *timelike*). If two events e_1 and e_2 satisfy $e_1 \cdot e_2 = 0$, then they are said to be orthogonal to each other.

Lemma 5.30. *Suppose two non-zero events e_1 and e_2 are orthogonal to each other. If e_1 is timelike, then e_2 is spacelike.*

Proof. Let $e_i = (x_i, y_i, \tau_i)$. Since e_1 is timelike, $\|e_1\|^2 = x_1^2 + y_1^2 - \tau_1^2 < 0$. Hence, $\tau_1 \neq 0$, and

$$\frac{1}{\tau_1^2}(x_1^2 + y_1^2) < 1.$$

Because $e_1 \cdot e_2 = 0$, $\tau_1\tau_2 = x_1x_2 + y_1y_2$. Therefore,

$$\tau_2 = \frac{1}{\tau_1}(x_1x_2 + y_1y_2).$$

We note that

$$\begin{aligned}
\|e_2\|^2 &= x_2^2 + y_2^2 - \tau_2^2 \\
&= x_2^2 + y_2^2 - \left(\frac{1}{\tau_1}(x_1x_2 + y_1y_2)\right)^2 \\
&\geq x_2^2 + y_2^2 - \frac{1}{\tau_1^2}(x_1^2 + y_1^2)(x_2^2 + y_2^2) \\
&\geq x_2^2 + y_2^2 - (x_2^2 + y_2^2) \\
&= 0,
\end{aligned}$$

where the equalities hold only if $x_2^2+y_2^2 = 0$, i.e., $x_2 = y_2 = 0$. Hence, if $\|e_2\|^2 = 0$, then $x_2 = y_2 = 0$. However, then $0 = \|e_2\|^2 = \tau_2^2$ and $\tau_2 = 0$, which is impossible because e_2 is a non-zero event. Therefore, $\|e_2\|^2 > 0$. □

For a parameterized curve

$$\gamma : [a, b] \to \mathbb{R}^{2,1}, \quad \gamma(t) = (x(t), y(t), \tau(t)),$$

γ is timelike (spacelike) if $\|\gamma'(t)\|^2 < 0$ (> 0) for each t. For a timelike or a spacelike curve $\gamma : [a, b] \to \mathbb{R}^{2,1}$, we define its relativistic length as follows:

$$l_R(\gamma) = \int_a^b \sqrt{-\left\|\frac{d\gamma(t)}{dt}\right\|^2}\, dt$$

for a timelike curve and

$$l_R(\gamma) = \int_a^b \sqrt{\left\|\frac{d\gamma(t)}{dt}\right\|^2}\, dt$$

for a spacelike curve.

Let us consider a hyperbolic surface, which is called a *hyperboloid*:

$$\mathbb{U}^2 := \{(x, y, \tau) \in \mathbb{R}^{2,1} \mid \tau = \sqrt{1 + x^2 + y^2}\}.$$

Note that

$$\mathbb{U}^2 = \{(x, y, \tau) \in \mathbb{R}^{2,1} \mid x^2 + y^2 - \tau^2 = -1, \tau > 0\}.$$

Hence, $\mathbb{U}^2$ is one of two components of the surface defined by

$$\|e\|^2 = -1.$$

Thus, $\mathbb{U}^2$ may be considered a relativistic "sphere" of squared radius -1 or of radius $\sqrt{-1}$.

Proposition 5.31. *A curve on $\mathbb{U}^2$ is spacelike.*

Proof. Let

$$\gamma : [a, b] \to \mathbb{U}^2, \quad \gamma(t) = (x(t), y(t), \tau(t))$$

be a parametrization of a curve on $\mathbb{U}^2$. Note that

$$-1 = \|\gamma(t)\|^2 = x(t)^2 + y(t)^2 - \tau(t)^2.$$

The event $\gamma(t)$ is timelike. Differentiating both sides of the equation with respect to t yields

$$\begin{aligned} 0 =& 2x(t)\frac{dx(t)}{dt} + 2y(t)\frac{dy(t)}{dt} - 2\tau(t)\frac{d\tau(t)}{dt} \\ =& 2(x(t), y(t), \tau(t)) \cdot (\frac{dx(t)}{dt}, \frac{dy(t)}{dt}, \frac{d\tau(t)}{dt}) \\ =& 2\gamma(t) \cdot \frac{d\gamma(t)}{dt}. \end{aligned}$$

Hence, the events $\gamma(t)$ and $\frac{d\gamma(t)}{dt}$ are orthogonal. According to Lemma 5.30, $\frac{d\gamma(t)}{dt}$ is spacelike, so the curve γ is spacelike. □

We will use the relativistic length $l_R(\gamma)$ as the $\mathbb{U}^2$-length of a curve γ on $\mathbb{U}^2$:

$$l_{\mathbb{U}^2}(\gamma) = l_R(\gamma).$$

If we project the hyperboloid $\mathbb{U}^2$ through the origin to the plane $\tau = 1$, the image is a unit disk, which is identified as a Klein disk by dropping the τ-coordinate (Figure 5.7): $(x, y, 1) \mapsto (x, y)$.

This results in a bijective map $\xi : \mathbb{U}^2 \to \mathbb{K}^2$ from the hyperbolic surface to the Klein disk as follows:

$$(x, y, \tau) \overset{\xi}{\longmapsto} \left(\frac{x}{\tau}, \frac{y}{\tau}\right).$$

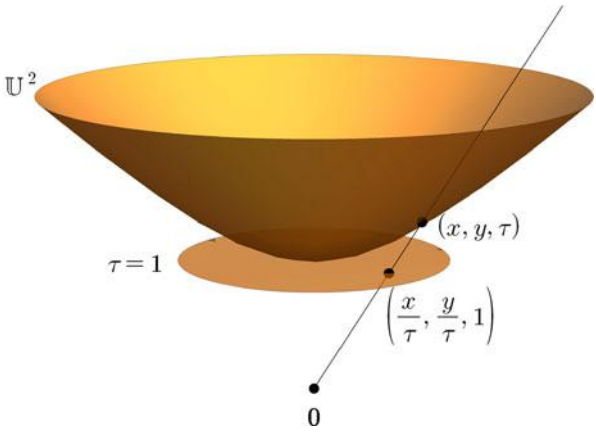


Fig. 5.7 Projection of the hyperboloid $\mathbb{U}^2$ through the origin to the plane $\tau = 1$

The following theorem states that $\mathbb{U}^2$ and $\mathbb{K}^2$ (and $\mathbb{H}^2$ and $\mathbb{D}^2$) are essentially the same space.

Theorem 5.32. *For a curve Γ on $\mathbb{U}^2$ of finite relativistic length,*

$$l_{\mathbb{U}^2}(\Gamma) = l_{\mathbb{K}^2}(\xi(\Gamma)).$$

Proof. One can show that the map $\zeta : \mathbb{U}^2 \to \mathbb{D}^2$ is as follows:

$$(x, y, \tau) \mapsto \left(\frac{x}{1+\tau}, \frac{y}{1+\tau}\right).$$

Then, $l_{\mathbb{D}^2}(\zeta(\Gamma)) = l_{\mathbb{K}^2}(\xi(\Gamma))$, and thus, it is sufficient to show that

$$l_{\mathbb{U}^2}(\Gamma) = l_{\mathbb{D}^2}(\zeta(\Gamma)).$$

Let $\gamma : [a, b] \to \mathbb{U}^2$, $\gamma(t) = (x(t), y(t), \tau(t))$ be a parametrization of Γ. Then,

$$l_{\mathbb{U}^2}(\gamma) = \int_a^b \sqrt{\left\|\frac{d\gamma}{dt}\right\|^2}\, dt = \int_a^b \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2 - \left(\frac{d\tau}{dt}\right)^2}\, dt.$$

Now note that

$$\delta(t) = \zeta(\gamma(t)) = \left(\frac{x(t)}{1+\tau(t)}, \frac{y(t)}{1+\tau(t)}\right)$$

is a parametrization of $\zeta(\Gamma)$. Let

$$X = \frac{x}{1+\tau},\ Y = \frac{y}{1+\tau}.$$

Using the relations

$$x^2 + y^2 + 1 = \tau^2$$

and

$$x\frac{dx}{dt} + y\frac{dy}{dt} = \tau\frac{d\tau}{dt},$$

$$\begin{aligned}
&\left(\frac{2}{1-X^2-Y^2}\right)^2\left(\left(\frac{dX}{dt}\right)^2+\left(\frac{dY}{dt}\right)^2\right)\\
&= \frac{4\left(\left(-\frac{x\frac{d\tau}{dt}}{(1+\tau)^2}+\frac{\frac{dx}{dt}}{1+\tau}\right)^2+\left(-\frac{y\frac{d\tau}{dt}}{(1+\tau)^2}+\frac{\frac{dy}{dt}}{1+\tau}\right)^2\right)}{\left(1-\frac{x^2}{(1+\tau)^2}-\frac{y^2}{(1+\tau)^2}\right)^2}\\
&= \frac{4\left(\left(x\frac{d\tau}{dt}-(1+\tau)\frac{dx}{dt}\right)^2+\left(y\frac{d\tau}{dt}-(1+\tau)\frac{dy}{dt}\right)^2\right)}{\left(-(1+\tau)^2+x^2+y^2\right)^2}\\
&= \frac{\left(\left(x\frac{d\tau}{dt}-(1+\tau)\frac{dx}{dt}\right)^2+\left(y\frac{d\tau}{dt}-(1+\tau)\frac{dy}{dt}\right)^2\right)}{(1+\tau)^2}\\
&= \left(\frac{dx}{dt}\right)^2+\left(\frac{dy}{dt}\right)^2+\frac{(\tau-1)\left(\frac{d\tau}{dt}\right)^2-2\frac{d\tau}{dt}\left(x\frac{dx}{dt}+y\frac{dy}{dt}\right)}{(1+\tau)}\\
&= \left(\frac{dx}{dt}\right)^2+\left(\frac{dy}{dt}\right)^2-\left(\frac{d\tau}{dt}\right)^2.
\end{aligned}$$

Therefore,

$$\begin{aligned}
l_{\mathbb{D}^2}(\zeta(\Gamma)) &= l_{\mathbb{D}^2}(\delta)\\
&= \int_a^b \frac{2}{1-X^2-Y^2}\sqrt{\left(\frac{dX}{dt}\right)^2+\left(\frac{dY}{dt}\right)^2}\,dt\\
&= \int_a^b \sqrt{\left(\frac{dx}{dt}\right)^2+\left(\frac{dy}{dt}\right)^2-\left(\frac{d\tau}{dt}\right)^2}\,dt
\end{aligned}$$

$$= l_{\mathbb{U}^2}(\gamma)$$

$$= l_{\mathbb{U}^2}(\Gamma).$$

□

For two distinct events p_1 and p_2 on $\mathbb{U}^2$, a path on $\mathbb{U}^2$ from p_1 to p_2 is a smooth curve $\gamma : [a, b] \to \mathbb{U}^2$ such that

$$\gamma(a) = p_1, \gamma(b) = p_2.$$

Definition 5.33. A path γ on $\mathbb{U}^2$ from an event p_1 to another event p_2 is called a $\mathbb{U}^2$-*shortest* path from p_1 to p_2 if

$$l_{\mathbb{U}^2}(\gamma) \leq l_{\mathbb{U}^2}(\gamma')$$

for any path γ' on $\mathbb{U}^2$ from p_1 to p_2.

Theorem 5.32 implies that a curve γ on $\mathbb{U}^2$ is a $\mathbb{U}^2$-shortest path if and only if its image on $\mathbb{K}^2$ under ξ is a $\mathbb{K}^2$-shortest path on $\mathbb{K}^2$. As before, we define the $\mathbb{U}^2$ distance $d_{\mathbb{U}^2}$ on $\mathbb{U}^2$ as follows:

$$d_{\mathbb{U}^2}(p, q) = l_{\mathbb{U}^2}(\gamma)$$

for $p, q \in \mathbb{U}^2$, where γ is a $\mathbb{U}^2$-shortest path from p to q. An isometry of $\mathbb{U}^2$ is a bijective map from $\mathbb{U}^2$ to itself that preserves the $\mathbb{U}^2$ distance.

Note that

$$d_{\mathbb{U}^2}(p, q) = l_{\mathbb{U}^2}(\gamma) = l_{\mathbb{K}^2}(\xi(\gamma)) = d_{\mathbb{K}^2}(\xi(p), \xi(q)).$$

Hence, the map $\xi : \mathbb{U}^2 \to \mathbb{K}^2$ is an isometry, and the two spaces $\mathbb{U}^2$ and $\mathbb{K}^2$ are isometric. We now define a $\mathbb{U}^2$-line as the set of all the events on $\mathbb{U}^2$ that have the same $\mathbb{U}^2$ distance from some two distinct events on $\mathbb{U}^2$. Then, it is not difficult to see that a curve γ on $\mathbb{U}^2$ is a $\mathbb{U}^2$-line if and only if $\xi(\gamma)$ is a $\mathbb{K}^2$-line. Hence, we can pose the following theorem.

Theorem 5.34. *A $\mathbb{U}^2$-line on $\mathbb{U}^2$ is an intersection of $\mathbb{U}^2$ with a plane in $\mathbb{R}^{2,1}$ containing the origin.*

Proof. There is a one-to-one correspondence between the set of $\mathbb{U}^2$-lines on $\mathbb{U}^2$ and the set of $\mathbb{K}^2$-lines on $\mathbb{K}^2$. A $\mathbb{K}^2$-line L is a Euclidean line that is the intersection of the plane $\tau = 1$ and a plane P passing through the origin. Hence, the $\mathbb{K}^2$-line L yields a $\mathbb{U}^2$-line M on $\mathbb{U}^2$ that is the intersection of the plane P and $\mathbb{U}^2$. Conversely, for a given $\mathbb{K}^2$-line M on the plane $\tau = 1$, there is a unique plane P, passing through the origin, whose intersection with the plane $\tau = 1$ is a $\mathbb{K}^2$-line L. See Figure 5.8.

□

Note that the plane, corresponding to a given $\mathbb{U}^2$-line in the proof of Theorem 5.34, is unique.

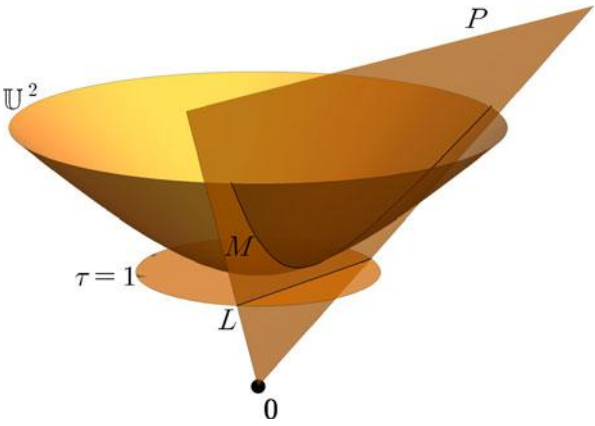


Fig. 5.8 $\mathbb{K}^2$-line L and $\mathbb{U}^2$-line M

It is now clear that the spaces $\mathbb{H}^2$, $\mathbb{D}^2$, $\mathbb{K}^2$ and $\mathbb{U}^2$ are all isometric (Figure 5.9).

A cross-section in $\mathbb{R}^{2,1}$ is shown in Figure 5.9, which shows how the models $\mathbb{H}^2$, $\mathbb{D}^2$, $\mathbb{K}^2$,$\mathbb{J}^2$, and $\mathbb{U}^2$ are related. The five bullet points η, δ, κ, ζ, and υ represent the same point in the hyperbolic geometry in two dimensions. The following is how the points on the figure correspond to each of the models of hyperbolic surfaces in Figure 5.9.

- $\mathbb{H}^2$: $(x, y) \leftrightarrow (1, x, y)$
- $\mathbb{D}^2$: $(x, y) \leftrightarrow (x, y, 0)$
- $\mathbb{K}^2$: $(x, y) \leftrightarrow (x, y, 1)$

See also Figure 5.10, where a Poincaré disk on the plane $\tau = 0$ is shown, explaining the correspondence between $\mathbb{D}^2$-lines (L) and $\mathbb{U}^2$-lines (M).

Exercises

5.29. Consider a $\mathbb{U}^2$ circle Γ of $\mathbb{U}^2$-radius ρ with center $\alpha \in \mathbb{U}^2$:

$$\Gamma = \{e \in \mathbb{U}^2 \mid d_{\mathbb{U}^2}(\alpha, e) = \rho\}.$$

Find its $\mathbb{U}^2$-circumference, $l_{\mathbb{U}^2}(\Gamma)$.

5.8 Isometries of $\mathbb{R}^{2,1}$

In Section 2.4, we saw that we can understand some aspects of isometries of $\mathbb{R}^3$ based on the geometry of $\mathbb{S}^2$. Since the hyperboloid $\mathbb{U}^2$ is isometric with the hyperbolic plane, we now have a good knowledge of its isometries. In this section,

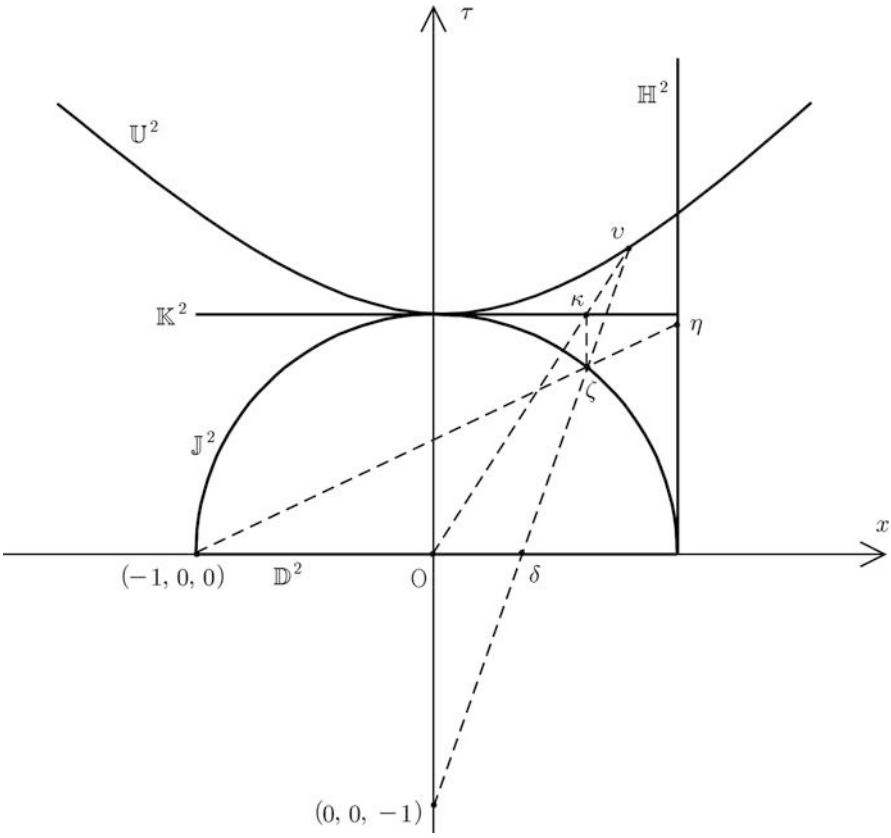


Fig. 5.9 $\mathbb{H}^2$, $\mathbb{D}^2$, $\mathbb{K}^2$, $\mathbb{J}^2$ and $\mathbb{U}^2$

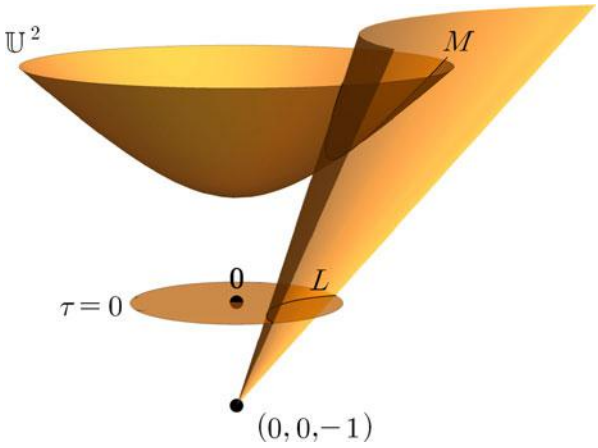


Fig. 5.10 $\mathbb{D}^2$-line L and $\mathbb{U}^2$-line M

we investigate the structure of the isometries of $\mathbb{R}^{2,1}$ using our knowledge of the isometries of $\mathbb{U}^2$.

Let P be a plane in $\mathbb{R}^{2,1}$ that contains the origin. Then,

$$P = \{te_1 + se_2 \mid t, s \in \mathbb{R}\}$$

for some events e_1 and e_2. Conversely, if e_1 and e_2 are non-zero events in P such that there is no real number λ such that $e_2 = \lambda e_1$, then

$$P = \{te_1 + se_2 \mid t, s \in \mathbb{R}\}.$$

Lemma 5.35. *Every $\mathbb{U}^2$-line through an event $e_0 \in \mathbb{U}^2$ can be parameterized by*

$$t \mapsto e_0 \cosh t + \alpha \sinh t,$$

where α is an event such that

$$\|\alpha\|^2 = 1, \quad e_0 \cdot \alpha = 0.$$

Proof. Let Γ be a $\mathbb{U}^2$-line through the event e_0. Then, by Theorem 5.34, $\Gamma = \mathbb{U}^2 \cap P$ for some plane P through the origin. We can choose an event β from P such that $\beta = (x, y, 0)$ with $\|\beta\|^2 = 1$. Since e_0 and β are non-zero events and they are not proportional to one another,

$$P = \{te_0 + s\beta \mid t, s \in \mathbb{R}\}.$$

We seek an event α in P such that

$$\|\alpha\|^2 = 1, \quad e_0 \cdot \alpha = 0.$$

Let $\alpha = ae_0 + b\beta$. Then,

$$1 = \|\alpha\|^2 = -a^2 + b^2 + 2ab(e_0 \cdot \beta),$$

$$0 = e_0 \cdot \alpha = -a + b(e_0 \cdot \beta).$$

Hence,

$$a = \frac{\pm e_0 \cdot \beta}{\sqrt{1 + (e_0 \cdot \beta)^2}}, \quad b = \frac{\pm 1}{\sqrt{1 + (e_0 \cdot \beta)^2}},$$

and we have found such an event α. Since $\alpha = ae_0 + b\beta$,

$$P = \{te_0 + s\beta \mid t, s \in \mathbb{R}\} = \{te_0 + s\alpha \mid t, s \in \mathbb{R}\}.$$

Let $\gamma(t) = e_0 \cosh t + \alpha \sinh t$. Note that $\gamma(0) = e_0$ and $\gamma(t) \in P$. Since

$$\|\gamma(t)\|^2 = \|e_0\|^2 \cosh^2 t + \|\alpha\|^2 \sinh^2 t = -\cosh^2 t + \sinh^2 t = -1,$$

$\gamma(t) \in \mathbb{U}^2$. Hence, $\gamma(t) \in \mathbb{U}^2 \cap P = \Gamma$.

Conversely, for any event $e \in \Gamma = \mathbb{U}^2 \cap P$, $e = te_0 + s\alpha$ for some $s, t \in \mathbb{R}$. Note that

$$-1 = \|e\|^2 = t^2\|e_0\|^2 + 2ts(e_0 \cdot \alpha) + s^2\|\alpha\|^2 = -t^2 + s^2.$$

Note also that $t = \cosh\lambda$ and $s = \sinh\lambda$ for some real number λ. Therefore, $e = \gamma(\lambda)$. □

We present a two-dimensional version of Proposition 5.28.

Corollary 5.36. $\cosh\left(d_{\mathbb{U}^2}(e_1, e_2)\right) = -e_1 \cdot e_2$ *for any* $e_1, e_2 \in \mathbb{U}^2$.

Proof. According to Lemma 5.35, there is a parameterized curve

$$\gamma(t) = e_1 \cosh t + \alpha \sinh t$$

for some α with $\|\alpha\|^2 = 1$ and $e_1 \cdot \alpha = 0$ such that $\gamma(a) = e_2$ for some $a \in \mathbb{R}$.

$$\left\| \frac{d\gamma(t)}{dt} \right\|^2 = \cosh^2 t - \sinh^2 t = 1.$$

Hence,

$$d_{\mathbb{U}^2}(e_1, e_2) = \int_0^a \sqrt{\left\| \frac{d\gamma(t)}{dt} \right\|^2} dt = a.$$

Note that

$$e_1 \cdot e_2 = e_1 \cdot \gamma(a) = e_1 \cdot (e_1 \cosh a + \alpha \sinh a) = -\cosh a.$$

Therefore,

$$\cosh\left(d_{\mathbb{U}^2}(e_1, e_2)\right) = \cosh a = -e_1 \cdot e_2.$$

□

Compare it with the formula for the case of $\mathbb{S}^2$ and $\mathbb{R}^3$:

$$\cos\left(d_{\mathbb{S}^2}(p_1, p_2)\right) = p_1 \cdot p_2$$

for any $p_1, p_2 \in \mathbb{S}^2$, where the inner product "$\cdot$" is the ordinary one for the Euclidean space $\mathbb{R}^3$.

The formula for the $\mathbb{U}^2$-distance in Corollary 5.36 is simpler and easier to use than that for $\mathbb{H}^2$-distance (Exercise 4.4). This is one of the advantages that are enjoyed when one uses $\mathbb{U}^2$ for a model of hyperbolic geometry. One can use the formula in Corollary 5.36 in deducing a formula for the distance in $\mathbb{D}^2$ as follows:

Note that

$$(x, y) \overset{\zeta^{-1}}{\longmapsto} \left(\frac{2x}{1-(x^2+y^2)}, \frac{2y}{1-(x^2+y^2)}, \frac{1+x^2+y^2}{1-(x^2+y^2)} \right)$$

or

$$p \overset{\zeta^{-1}}{\longmapsto} \left(\frac{2p}{1-\|p\|^2}, \frac{1+\|p\|^2}{1-\|p\|^2} \right) = \frac{1}{1-\|p\|^2} \left(2p, 1+\|p\|^2 \right).$$

Hence,

$$\begin{aligned}
\cosh(d_{\mathbb{D}^2}(p_1, p_2)) &= \cosh(d_{\mathbb{U}^2}(\zeta^{-1}(p_1), \zeta^{-1}(p_2))) \\
&= -\zeta^{-1}(p_1) \cdot \zeta^{-1}(p_2) \\
&= -\frac{1}{(1-\|p_1\|^2)(1-\|p_2\|^2)} (4p_1 \cdot p_2 - (1+\|p_1\|^2)(1+\|p_2\|^2)) \\
&= 1 - \frac{1}{(1-\|p_1\|^2)(1-\|p_2\|^2)} (4p_1 \cdot p_2 - 2\|p_1\|^2 - 2\|p_2\|^2) \\
&= 1 - \frac{2\|p_1 - p_2\|^2}{(1-\|p_1\|^2)(1-\|p_2\|^2)} \\
&= 1 + \frac{2(d(p_1, p_2))^2}{(1-\|p_1\|^2)(1-\|p_2\|^2)},
\end{aligned}$$

i.e.,

$$\cosh(d_{\mathbb{D}^2}(p_1, p_2)) = 1 + \frac{2(d(p_1, p_2))^2}{(1-\|p_1\|^2)(1-\|p_2\|^2)}.$$

A map $\phi \in \mathrm{O}(2, 1)$ is called *orthochronous* if it maps $\mathbb{U}^2$ onto $\mathbb{U}^2$, and we denote the set of these maps by $\mathrm{O}^+(2, 1)$.

Proposition 5.37. *The restriction*

$$\phi|_{\mathbb{U}^2} : \mathbb{U}^2 \to \mathbb{U}^2$$

of an element ϕ *of* $\mathrm{O}^+(2, 1)$ *is an isometry of* $\mathbb{U}^2$.

Proof. It is not difficult to prove that a map from $\mathbb{R}^{2,1}$ to $\mathbb{R}^{2,1}$, fixing the origin, is an isometry of $\mathbb{R}^{2,1}$ if and only if it preserves the Minkowski inner product (regard

events as elements of $\mathbb{R}^{2,1}$ in the proof of Theorem 5.3). For any events $e_1, e_2 \in \mathbb{U}^2$, according to Corollary 5.36,

$$\cosh d_{\mathbb{U}^2}(\phi(e_1), \phi(e_2)) = -\phi(e_1) \cdot \phi(e_2) = -e_1 \cdot e_2 = \cosh d_{\mathbb{U}^2}(e_1, e_2),$$

which implies that $d_{\mathbb{U}^2}(\phi(e_1), \phi(e_2)) = d_{\mathbb{U}^2}(e_1, e_2)$. □

For two distinct events $e_1, e_2 \in \mathbb{R}^{2,1}$, we can define a plane P_{e_1,e_2} in $\mathbb{R}^{2,1}$ as in $\mathbb{R}^{1,1}$:

$$P_{e_1,e_2} := \{e \in \mathbb{R}^{2,1} \mid d^{\mathrm{II}}(e_1, e) = d^{\mathrm{II}}(e_2, e)\}.$$

Note that

$$P_{e_1,e_2} = \{e \in \mathbb{R}^{2,1} \mid (e - u) \cdot v = 0\}, \tag{5.4}$$

where $u = \frac{1}{2}(e_1 + e_2)$ and $v = \frac{1}{2}(e_1 - e_2)$ (Exercise 5.2). Furthermore, if $e_1 - e_2$ is not lightlike, we define the relativistic reflection $\bar{r}_{P_{e_1,e_2}}$ in the plane P_{e_1,e_2} as follows:

$$\bar{r}_{P_{e_1,e_2}}(e) = e - \frac{2(e - u) \cdot v}{\|v\|^2} v$$

It can be readily shown that the relativistic reflection is an isometry of $\mathbb{R}^{2,1}$ (Theorem 5.5) and that every translation can be expressed as a composition of two (for the non-lightlike direction, Theorem 5.8) or four (for the lightlike direction, Theorem 5.9) relativistic reflections, as shown in Section 5.2—just regard events as elements of $\mathbb{R}^{2,1}$ in the proofs.

Lemma 5.38. *For two distinct events $\alpha, \beta \in \mathbb{U}^2$, consider a $\mathbb{U}^2$-line*

$$L_{\alpha,\beta} := \{e \in \mathbb{U}^2 \mid d_{\mathbb{U}^2}(e, \alpha) = d_{\mathbb{U}^2}(e, \beta)\}.$$

Then

$$L_{\alpha,\beta} = P_{\alpha,\beta} \cap \mathbb{U}^2.$$

Furthermore, the event $\alpha - \beta$ is not lightlike. Hence, the relativistic reflection $\bar{r}_{P_{\alpha,\beta}}$ of $\mathbb{R}^{2,1}$ in the plane $P_{\alpha,\beta}$ is defined.

Proof. For each $e \in L_{\alpha,\beta}$,

$$\begin{aligned} d^{\mathrm{II}}(\alpha, e) &= \|\alpha - e\|^2 \\ &= \|\alpha\|^2 - 2\alpha \cdot e + \|e\|^2 \\ &= -2 - 2\alpha \cdot e \end{aligned}$$

$$
\begin{aligned}
&= -2 + 2\cosh\left(d_{\mathbb{U}^2}(\alpha, e)\right) && (\because \text{Corollary 5.36}) \\
&= -2 + 2\cosh\left(d_{\mathbb{U}^2}(\beta, e)\right) \\
&= -2 - 2\beta \cdot e && (\because \text{Corollary 5.36}) \\
&= \|\beta\|^2 - 2\phi(\alpha) \cdot e + \|e\|^2 \\
&= \|\beta - e\|^2 \\
&= d^{\mathrm{II}}(\beta, e).
\end{aligned}
$$

Hence, $e \in P_{\alpha,\beta}$ and so $L_{\alpha,\beta} \subset P_{\alpha,\beta}$. By Theorem 5.34, $L_{\alpha,\beta} = P_{\alpha,\beta} \cap \mathbb{U}^2$.

$$
\begin{aligned}
\|\alpha - \beta\|^2 &= \|\alpha\|^2 - 2\alpha \cdot \beta + \|\beta\|^2 \\
&= -2 - 2\alpha \cdot \beta \\
&= -2 + 2\cosh\left(d_{\mathbb{U}^2}(\alpha, \beta)\right) && (\because \text{Corollary 5.36}) \\
&\neq 0. && (\because d_{\mathbb{U}^2}(\alpha, \beta) \neq 0)
\end{aligned}
$$

Hence, the event $\alpha - \beta$ is not lightlike and so the relativistic reflection $\bar{r}_P$ in $P = P_{\alpha,\beta}$ is defined. □

An isometry ϕ of $\mathbb{U}^2$ is called a reflection if the corresponding isometry $\zeta^{-1} \circ \phi \circ \zeta$ of $\mathbb{D}^2$ is an inversion in a $\mathbb{D}^2$-line. In the following theorem, it is proved that a reflection of $\mathbb{U}^2$ is the restriction to $\mathbb{U}^2$ of a relativistic reflection in a plane that passes through the origin in $\mathbb{R}^{2,1}$. This property was discussed for the case of $\mathbb{R}^3$ and $\mathbb{S}^2$ in Section 2.2.

Theorem 5.39. *An isometry ϕ of $\mathbb{U}^2$ is a reflection in a $\mathbb{U}^2$-line L if and only if*

$$\phi(e) = \bar{r}_P(e)$$

for each event $e \in \mathbb{U}^2$, where P is the plane in $\mathbb{R}^{2,1}$ through the origin such that $P \cap \mathbb{U}^2 = L$.

Proof. By Theorem 5.34, $L = \mathbb{U}^2 \cap P$ for some plane P through the origin. By Lemma 5.8, the relativistic reflection $\bar{r}_P$ in the plane P is defined.

Let ϕ be the reflection of $\mathbb{U}^2$ in a $\mathbb{U}^2$-line L. If $e \in L$, then $e \in P$ and so

$$\phi(e) = e = \bar{r}_P(e).$$

If an event $e \in \mathbb{U}^2$ does not lie on L, then $e \neq \phi(e)$ and

$$L = L_{e,\phi(e)}.$$

By Lemma 5.8,

$$L_{e,\phi(e)} = P_{e,\phi(e)} \cap \mathbb{U}^2.$$

Hence, $P = P_{e,\phi(e)}$ and so $\bar{r}_P(e) = \phi(e)$.

Conversely, assume that an isometry ϕ of $\mathbb{U}^2$ satisfies

$$\phi(e) = \bar{r}_P(e)$$

for each $e \in \mathbb{U}^2$. If $e \in L$, then $e \in P$ and so

$$\phi(e) = \bar{r}_P(e) = e.$$

If $e \notin L$, then $e \notin P$ and so $\bar{r}_P(e) \neq e$. Hence, $P = P_{e,\bar{r}_P(e)}$ and, by Lemma 5.8, $L = L_{e,\bar{r}_P(e)}$.

$$\phi(e) = \bar{r}_P(e) = \bar{r}_P(e).$$

Therefore, ϕ is the reflection of $\mathbb{U}^2$ in L. □

Since $\mathbb{U}^2$ and $\mathbb{D}^2$ are isometric, every isometry of $\mathbb{U}^2$ is a composition of at most three reflections.

If a plane P contains the origin, then we can let $u = \mathbf{0}$ in (5.4) (Exercise 5.2),

$$P = \{e \in \mathbb{R}^{1,1} \mid e \cdot v = 0\} = v^{\perp}.$$

We are interested in the case that P meets with $\mathbb{U}^2$ along a $\mathbb{U}^2$-line. Hence, assume that $P \cap \mathbb{U}^2 \neq \emptyset$ and choose an event $\alpha \in P \cap \mathbb{U}^2$.

$$\alpha \cdot v = 0, \ \ \|\alpha\|^2 = -1.$$

By Lemma 5.30, $\|v\|^2 > 0$. Thus, we can assume that $\|v\|^2 = 1$ and then

$$\bar{r}_P(e) = e - 2(e \cdot v)v.$$

The following is a higher-dimensional version of Theorem 5.29 and very similar to Theorem 2.18.

Theorem 5.40. *The map*

$$\Psi : \mathrm{O}^+(2, 1) \to \mathrm{Iso}(\mathbb{U}^2)$$

defined by the restriction is bijective.

Proof. In this proof, we assume some knowledge of linear algebra. We will build an inverse map of the map in the statement. For each isometry ϕ of $\mathbb{U}^2$, we must find a map ψ from $\mathrm{O}^+(2, 1)$ whose restriction to $\mathbb{U}^2$ is ϕ. Let

$$e_1 = (0, 0, 1),\ \ e_2 = (1, 0, \sqrt{2}),\ \ e_3 = (0, 1, \sqrt{2}),$$

which are elements of $\mathbb{U}^2$. Let $e_i' = \phi(e_i)$. According to Corollary 5.36,

$$\phi(e_i) \cdot \phi(e_j) = -\cosh d_{\mathbb{U}^2}(\phi(e_i), \phi(e_j)) = -\cosh d_{\mathbb{U}^2}(e_i, e_j) = e_i \cdot e_j$$

for each i, j. Note that any event $e \in \mathbb{R}^{2,1}$ can be expressed as

$$e = a_1 e_1 + a_2 e_2 + a_3 e_3$$

with coefficients $a_1, a_2, a_3 \in \mathbb{R}$ and those coefficients are unique. We define a linear map $\psi : \mathbb{R}^{2,1} \to \mathbb{R}^{2,1}$ as follows:

$$a_1 e_1 + a_2 e_2 + a_3 e_3 \mapsto a_1 \phi(e_1) + a_2 \phi(e_2) + a_3 \phi(e_3),$$

where $a_i \in \mathbb{R}$. We can denote

$$a_1 e_1 + a_2 e_2 + a_3 e_3$$

by

$$\sum_i a_i e_i.$$

Then

$$\psi\left(\sum_i a_i e_i\right) = \sum_i a_i \psi(e_i)).$$

Hence,

$$\begin{aligned}
\psi\left(\sum_i a_i e_i\right) \cdot \psi\left(\sum_j b_j e_j\right) &= \sum_i a_i \phi(e_i) \cdot \sum_j b_j \phi(e_j) \\
&= \sum_{i,j} a_i b_j \left(\phi(e_i) \cdot \phi(e_j)\right) \\
&= \sum_{i,j} a_i b_j (e_i \cdot e_j) \\
&= \sum_i a_i e_i \cdot \sum_j b_j e_j.
\end{aligned}$$

Therefore, ψ preserves the Minkowski inner product, and accordingly, it is an isometry of $\mathbb{R}^{2,1}$.

Note that $e_1 \in \mathbb{U}^2$ and $\psi(e_1) = \phi(e_1) \in \mathbb{U}^2$. Hence, $e_1 \in \mathbb{U}^2 \cap \psi(\mathbb{U}^2)$ and so $\mathbb{U}^2 \cap \psi(\mathbb{U}^2) \neq \emptyset$. Using the result from Exercise 5.30, we conclude that $\phi(\mathbb{U}^2) = \mathbb{U}^2$ and so $\psi \in \mathrm{O}^+(2,1)$.

For $e, e' \in \mathbb{U}^2$,

$$\begin{aligned} d_{\mathbb{U}^2}(\psi(e), \psi(e')) &= \cosh^{-1}(-\psi(e) \cdot \psi(e')) && (\because \text{ Corollary 5.36}x) \\ &= \cosh^{-1}(-e \cdot e') \\ &= d_{\mathbb{U}^2}(e, e') && (\because \text{ Corollary 5.36}). \end{aligned}$$

So $\psi|_{\mathbb{U}^2}$ is an isometry of $\mathbb{U}^2$. Since there are no planes in $\mathbb{R}^{2,1}$ that contain all the points e_1, e_2, e_3 and the origin, there are no $\mathbb{U}^2$-lines that pass through all the points e_1, e_2, and e_3. Note that $\psi(e_i) = \phi(e_i)$ for $i = 1, 2, 3$. By Theorem 4.19,

$$\psi|_{\mathbb{U}^2} = \phi,$$

where we note that $\mathbb{H}^2$ and $\mathbb{U}^2$ are isometric with each other. For a given isometry ϕ of $\mathbb{U}^2$, we have built an isometry ψ in $\mathrm{O}^+(2,1)$ such that $\Psi(\psi) = \phi$. Hence, the map Ψ is bijective. □

By Theorem 5.39, now it is clear that a relativistic reflection in $\mathrm{O}^+(2,1)$ corresponds to a reflection of $\mathbb{U}^2$ via the map Ψ in Theorem 5.40.

Theorem 5.41. *An isometry of $\mathbb{R}^{2,1}$ is a composition of at most eight relativistic reflections.*

Proof *(Sketch).* Let ϕ be an isometry of $\mathbb{R}^{2,1}$. Suppose that $\phi \in \mathrm{O}^+(2,1)$. Then, from the proof of Theorem 5.40, we can show that $\phi|_{\mathbb{U}^2}$ is an isometry of $\mathbb{U}^2$, which is a composition of at most three reflections. This implies that ϕ is a composition of at most three relativistic reflections. If $\phi \in \mathrm{O}(2,1)$ but $\phi \notin \mathrm{O}^+(2,1)$, then $\phi(e) \notin \mathbb{U}^2$ for some $e \in \mathbb{U}^2$. Since $(\bar{r} \circ \phi)(e) \in \mathbb{U}^2$,

$$(\bar{r} \circ \phi)(\mathbb{U}^2) \cap \mathbb{U}^2 \neq \emptyset,$$

where $\bar{r}$ is the relativistic reflection in the xy-plane. Using the result from Exercise 5.30, we conclude that

$$(\bar{r} \circ \phi)(\mathbb{U}^2) = \mathbb{U}^2,$$

and that $\bar{r} \circ \phi$ lies in $\mathrm{O}^+(2,1)$; thus, it is a composition of at most three relativistic reflections. Therefore, ϕ is a composition of at most four relativistic reflections in this case. If $\phi \notin \mathrm{O}(2,1)$, then

$$t_\alpha \circ \phi \in \mathrm{O}(2,1),$$

where $\alpha = -\phi(0)$. Note that $t_\alpha \circ \phi$ is a composition of at most four relativistic reflections and that the translation t_α is a composition of two or four relativistic reflections. Therefore, ϕ is a composition of at most eight relativistic reflections in the most general case. □

It is known that an isometry of $\mathbb{R}^{2,1}$ can be expressed as a composition of at most four (not five!) relativistic reflections.

Corollary 5.42. *An isometry of $\mathbb{R}^{2,1}$ is a composition of relativistic reflections.*

Using this corollary, one can show that an isometry of $\mathbb{R}^{2,1}$ can be expressed with a matrix multiplication and addition by an event, as in Corollary 5.23.

Exercises

5.30. Suppose that the set $\phi(\mathbb{U}^2) \cap \mathbb{U}^2$ is non-empty for a map $\phi \in \mathrm{O}(2,1)$. Prove that ϕ maps $\mathbb{U}^2$ onto $\mathbb{U}^2$.

5.31. Let e_1 and e_2 be timelike events in $\mathbb{R}^{2,1}$. Show that

$$(e_1 \cdot e_2)^2 \geq \|e_1\|^2 \|e_2\|^2$$

with equality if and only if $e_1 = \lambda e_2$ for some real number λ.

5.32. Two $\mathbb{U}^2$-lines are said to intersect orthogonally with each other if the two corresponding $\mathbb{D}^2$-lines on $\mathbb{D}^2$ intersect orthogonally with each other.

For two $\mathbb{U}^2$-lines $L = \alpha^\perp \cap \mathbb{U}^2$ and $M = \beta^\perp \cap \mathbb{U}^2$, show that the $\mathbb{U}^2$-lines L and M intersect orthogonally with each other if and only if

$$\alpha \cdot \beta = 0.$$

(*Hint.* Let ϕ be the reflection of $\mathbb{U}^2$ in the $\mathbb{U}^2$-line L. Then the $\mathbb{U}^2$-line M intersect orthogonally with L if and only if $\phi(M) = M$ and $L \neq M$.)

Chapter 6
Geometry of Special Relativity

"The views of space and time which I wish to lay before you have sprung from the soil of experimental physics, and therein lies their strength. They are radical. Henceforth space by itself, and time by itself, are doomed to fade away into mere shadows, and only a kind of union of the two will preserve an independent reality."

Hermann Minkowski (1864–1909)

"In all affairs it's a healthy thing now and then to hang a question mark on the things you have long taken for granted."

Bertrand Russell (1872–1970)

6.1 $\mathbb{R}^{3,1}$ and the Special Relativity of Einstein

Spacetime is the arena in which all physical events take place, i.e., an event is a point in spacetime specified by its position and time. The basic elements of spacetime are events. In spacetime, an event is a unique position (x, y, z) with a unique time t. Thus, it is specified by quadruples of real numbers, i.e., (x, y, z, t). For example, in the year 2014, an exploding star (supernova) was spotted, later named SN 2014J, in a nearby galaxy, which is at a distance of approximately 12 million light years. This explosion is an example of an event, wherein its position and time are described

N.-H. Lee, *Geometry: from Isometries to Special Relativity*, Undergraduate Texts in Mathematics, https://doi.org/10.1007/978-3-030-42101-4_6

from our point of view, i.e., the viewpoint of the human beings on Earth. We can set Earth as the origin of a coordinate system of spacetime and assign the coordinates (x, y, z, t).

Spacetime itself can be viewed as the set of all events in the same way that the Euclidean plane is the set of all of its points. Hence, it can be identified with the set $\mathbb{R}^4$. The trajectories of elementary point particles such as electrons through space and time are thus a continuum of events that may be regarded as a "curve" in $\mathbb{R}^4$. The trajectory of a compound object (consisting of a huge number of elementary point particles) is a union of many curves twisted together by virtue of interactions between the particles in it through spacetime.

Physicists use the term "observer" as a synonym for a specific reference frame from which a set of events is being recorded. Referring to an observer in special relativity is not specifically considering an individual creature who is witnessing events, but rather, it is a particular coordinate system by which events are to be assigned coordinates. The effects of special relativity occur regardless of whether there is a human being within the inertial reference frame to observe them.

For an observer $\mathbf{O}$, the set of events has a one-to-one correspondence to $\mathbb{R}^4$ by specifying each event by quadruples of real numbers (x, y, z, t). If there is another observer $\mathbf{O}'$, she may record each event in her own way. In the example of the supernova SN 2014J above, imagine that there are other creatures in the center of the Andromeda Galaxy, which is a spiral galaxy approximately 2.5 million light-years (2.4×1019 km) from Earth, moving at 300 kilometers per second in the direction of the earth. They also observed the explosion, and they would record it from their point of view, assigning it the coordinates (x', y', z', t').

The transformations in coordinates of all events between these two observers $\mathbf{O}$ and $\mathbf{O}'$ can be expressed by a map $\phi : \mathbb{R}^4 \to \mathbb{R}^4$, $(x, y, z, t) \mapsto (x', y', z', t')$. Before Einstein and his contemporaries, the common belief was that the spatial distance and the time interval are independently invariant under this transformation:

$$\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2} = \sqrt{(x_1' - x_2')^2 + (y_1' - y_2')^2 + (z_1' - z_2')^2}, \tag{6.1}$$

$$t_1 - t_2 = t_1' - t_2'. \tag{6.2}$$

In the nineteenth century, this belief started to lose ground. In classical electromagnetism, the electromagnetic field obeys a set of equations known as Maxwell's equations. It was noted that light is a fluctuating wave of electromagnetic fields. According to Maxwell's equations, the speed of light is a universal constant ($\approx 3.00 \times 10^8$ m/s), independent of the observers. However, (6.1) and (6.2) (and some intuition) imply that the speed must vary with respect to the observers unless it is infinite.

One suggested solution to this contradiction was to assume the existence of a luminiferous background substance (called "ether") through which the light

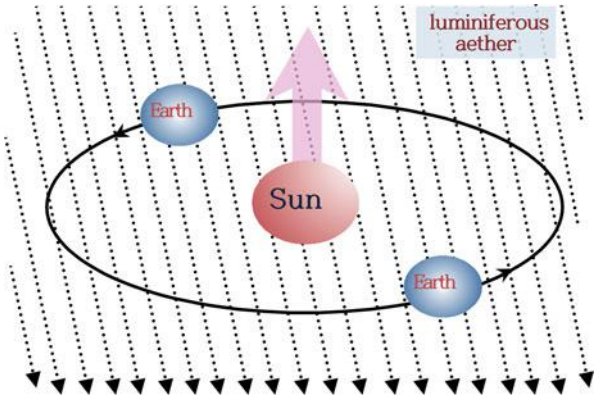


Fig. 6.1 Earth in the ether wind

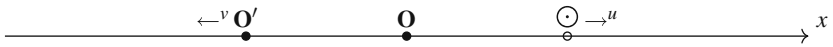


Fig. 6.2 Observers **O** and **O**′ in relative motion and a moving object $\odot$

propagates. Just as waves on the surface of water must have a supporting substance, a "medium," to be propagated (in this case, water) and audible sound needs a medium to transmit its wave motions (such as gas or liquid), light, which is an electromagnetic wave, must also propagate in a medium, the so-called "luminiferous ether." Since light can be propagated through a vacuum, it was assumed that even a vacuum must be filled with luminiferous ether.

The Earth orbits around the Sun at a speed of approximately 30 km/s. The Earth is in motion, through the ocean of ether. According to this hypothesis, Earth and the ether are in relative motion, which implies that there should be a so-called ether wind (Figure 6.1). In experiments conducted by Michelson and Morley, it was shown that such wind does not exist and that the speed of light is constant in fall and summer and in any direction.

Einstein and his contemporaries suggested another relation, instead of (6.1) and (6.2):

$$\begin{aligned}(x_1-x_2)^2+(y_1-y_2)^2+(z_1-z_2)^2-c^2(t_1-t_2)^2\\ =(x_1'-x_2')^2+(y_1'-y_2')^2+(z_1'-z_2')^2-c^2(t_1'-t_2')^2,\end{aligned}\tag{6.3}$$

where c is the speed of light.

Choosing a new unit for time by letting $\tau = ct$, (6.3) becomes

$$\begin{aligned}(x_1-x_2)^2+(y_1-y_2)^2+(z_1-z_2)^2-(\tau_1-\tau_2)^2\\ =(x_1'-x_2')^2+(y_1'-y_2')^2+(z_1'-z_2')^2-(\tau_1'-\tau_2')^2.\end{aligned}$$

Note that the speed of light is 1 in this time scale. Thus, the transformations of special relativity are isometries that preserve the Lorentz–Minkowski distance, defined as follows:

$$d^{\mathrm{II}}(e_1, e_2) = (x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2 - (\tau_1 - \tau_2)^2$$

for $e_1 = (x_1, y_1, z_1, \tau_1)$ and $e_2 = (x_2, y_2, z_2, \tau_2)$. Hence, we denote spacetime by $\mathbb{R}^{3,1}$. The physical theory resulting from (6.3) is called *the theory of special relativity*. Some consequences of (6.3) are the following:

1. Two events happening in two different locations that occur simultaneously for an observer may occur non-simultaneously for another observer (lack of absolute simultaneity). See Theorem 6.11.
2. The time lapse between two events is not independent of the observers and is dependent on the relative speeds of the observers. (The twin paradox describes a twin who departs from the Earth in a spaceship traveling near the speed of light and returns to find that his or her twin sibling has gotten much older.) See Theorem 6.16.
3. The dimensions (e.g., length) of an object measured by an observer may not be the same as those measured by another observer. (The ladder paradox involves a ladder, which is longer than a garage at rest, moving near the speed of light and being contained within the smaller garage.)
4. Speeds do not simply add. If the observer **O** measures an object $\odot$ as moving at speed u in the positive x-direction, then the observer $\mathbf{O}'$, moving at speed v in the negative x-direction with respect to **O**, will measure the object as moving with speed

$$\frac{u+v}{1+\frac{uv}{c^2}}. \tag{6.4}$$

See Section 6.5 for a derivation of this formula.

Exercises

6.1. Let $\mathbb{R}_1 = \{x \in \mathbb{R} \mid -1 < x < 1\}$. Motivated by (6.4), with the convention of $c = 1$, we define

$$u \oplus v = \frac{u+v}{1+uv}$$

for $u, v \in \mathbb{R}_1$. Show that

1.

$$u \oplus v \in \mathbb{R}_1$$

for $u, v \in \mathbb{R}_1$.

2.

$$(u \oplus v) \oplus w = u \oplus (v \oplus w)$$

for $u, v, w \in \mathbb{R}_1$.

3.

$$\underbrace{u \oplus u \oplus \cdots \oplus u}_{n \text{ times}} = \frac{(1+|u|)^n - (1-|u|)^n}{(1+|u|)^n + (1-|u|)^n} \frac{u}{|u|}$$

for a positive integer n and $u \in \mathbb{R}_1$.

6.2. For $r \in \mathbb{R}$ and $u \in \mathbb{R}_1$, let

$$r \otimes u = \frac{(1+|u|)^r - (1-|u|)^r}{(1+|u|)^r + (1-|u|)^r} \frac{u}{|u|}.$$

For any $r, r' \in \mathbb{R}$ and $u, v \in \mathbb{R}_1$, show that

1. $r \otimes u = \tanh\left(r \tanh^{-1}(u)\right)$.
2. $r \otimes u$ belongs to $\mathbb{R}_1$.
3. $(r + r') \otimes u = (r \otimes u) \oplus (r' \otimes u)$.
4. $(rr') \otimes u = r \otimes (r' \otimes u)$.
5. $r \otimes (u \oplus v) = (r \otimes u) \oplus (r \otimes v)$.

⁎ The operations $\oplus$ and $\otimes$ form the so-called Gyrogroups, which can be used to describe hyperbolic geometry in a different way.

6.2 Causality

Let us consider four-dimensional space with the Minkowski inner product

$$e_1 \cdot e_2 = x_1 x_2 + y_1 y_2 + z_1 z_2 - \tau_1 \tau_2$$

for events $e_i = (x_i, y_i, z_i, \tau_i) \in \mathbb{R}^4$. Then, the Lorentz–Minkowski distance $d^{\mathrm{II}}(e_1, e_2)$ is defined as follows:

$$d^{\mathrm{II}}(e_1, e_2) = \|e_1 - e_2\|^2 = (e_1 - e_2) \cdot (e_1 - e_2),$$

and we denote the set $\mathbb{R}^4$ with the Lorentz–Minkowski distance by $\mathbb{R}^{3,1}$ and call it four-dimensional Lorentz–Minkowski space, which is a mathematical model of spacetime. A bijective map $\phi : \mathbb{R}^{3,1} \to \mathbb{R}^{3,1}$ is called an isometry of $\mathbb{R}^{3,1}$ if it preserves the Lorentz–Minkowski distance. We denote the set of all such isometries

by $\mathrm{Iso}(\mathbb{R}^{3,1})$ and the set of all isometries fixing the origin $\mathbf{0}$ by $\mathrm{O}(3,1)$. All other notations can be similarly defined.

Recall that we are assuming that a curve is described by a smooth, regular parametrization. A parametrization $\gamma(t) = (x(t), y(t), z(t), \tau(t))$ is called smooth if all the functions $x(t)$, $y(t)$, $z(t)$, and $\tau(t)$ of t have derivatives of all finite orders, and it is called regular if

$$\frac{d\gamma(t)}{dt} \neq \mathbf{0}$$

for any t. A parametrization is always assumed to be smooth and regular.

For an object (approximated as a point in space, e.g., a particle or an observer), one can consider a curve composed of spacetime events that correspond to the history of the object. Each point of this curve is an event that can be labeled with a spatial position and the time of the object. We call this curve the *worldline* of the object. For example, the orbit of the Earth in the solar system is approximately a circle, a three-dimensional closed curve in space (Figure 6.3): the Earth returns every year to the same position in space. However, it comes back there at a different time. The worldline of the Earth is actually helical in spacetime (Figure 6.4) and does not return to the same point in the spacetime.

Roughly speaking, physics is the study of the worldlines of objects—how they look and how they can be determined. Before we proceed to the study of worldlines,

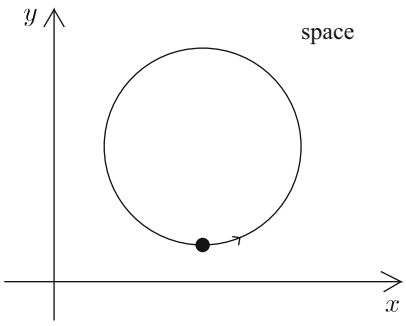


Fig. 6.3 Motion of the Earth in space

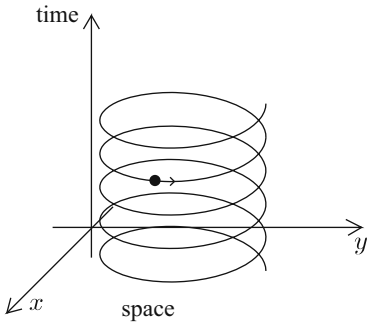


Fig. 6.4 Worldline of the Earth in spacetime

Fig. 6.5 Cause e_1 and effect e_2

we need to consider an important physical concept, *causality*. Causality is the relationship between causes and effects. An effect cannot occur before its cause.

Referring to Figure 6.5, let $e_1 = (x_1, y_1, z_1, \tau_1)$ be the event of the bullet being fired and $e_2 = (x_2, y_2, z_2, \tau_2)$ be the event of the bullet hitting the target. Since e_2 is an effect and e_1 is its cause, the time relation should be $\tau_1 < \tau_2$. Suppose that these events are recorded as $e_1' = (x_1', y_1', z_1', \tau_1')$ and $e_2' = (x_2', y_2', z_2', \tau_2')$ by another observer. We still should have $\tau_1' < \tau_2'$. In the previous section, we saw that simultaneity is not an absolute concept in relativity. Hence, comparing the time coordinates is not sufficient to address causality (see Exercise 6.7). For $e = (x, y, z, \tau) \in \mathbb{R}^{3,1}$, the event e is said to be *future-directed* (*past-directed*) if $\tau > 0$ ($\tau < 0$). We now can define a causality relation between events.

Definition 6.1. For two events $e_1, e_2 \in \mathbb{R}^{3,1}$, e_1 is said to *causally precede* e_2 if $e_2 - e_1$ is neither past-directed nor spacelike, i.e.,

$$\tau_2 - \tau_1 \geq 0 \text{ and } (x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2 - (\tau_1 - \tau_2)^2 \leq 0$$

for $e_i = (x_i, y_i, z_i, \tau_i)$, denoted as $e_1 \prec e_2$.

It is a simple but useful fact that

$$e_1 \prec e_2 \Leftrightarrow \mathbf{0} \prec e_2 - e_1.$$

Note that $e_1 \not\prec e_2$ does not imply $e_2 \prec e_1$. Two events e_1 and e_2 are said to be *causally related* if $e_1 \prec e_2$ or $e_2 \prec e_1$ and *causally unrelated* otherwise. If two events e_1, e_2 are causally unrelated, then

$$|\tau_1 - \tau_2| < \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2},$$

i.e., the spatial distance between e_1 and e_2 is greater than the time distance between them. One can say that they are too far from each other to be causally related within the given time elapse $|\tau_1 - \tau_2|$.

The causality relation is transitive. To prove this, we need a lemma.

Lemma 6.2. *For two events e_1, e_2,*

$$\mathbf{0} \prec e_1 + e_2$$

if

$$\mathbf{0} \prec e_1 \text{ and } \mathbf{0} \prec e_2.$$

Proof. First note that $e_1 + e_2$ is not past-directed. Therefore, we must show that

$$\|e_1 + e_2\|^2 \leq 0.$$

Let $e_i = (x_i, y_i, z_i, \tau_i)$ for $i = 1, 2$. Then, the given condition implies that

$$\tau_i \geq 0 \text{ and } x_i^2 + y_i^2 + z_i^2 \leq \tau_i^2.$$

Recall the Cauchy–Schwartz inequality,

$$x_1x_2 + y_1y_2 + z_1z_2 \leq \sqrt{x_1^2 + y_1^2 + z_1^2}\sqrt{x_2^2 + y_2^2 + z_2^2}.$$

Hence,

$$\begin{aligned} e_1 \cdot e_2 &= x_1x_2 + y_1y_2 + z_1z_2 - \tau_1\tau_2 \\ &\leq \sqrt{x_1^2 + y_1^2 + z_1^2}\sqrt{x_2^2 + y_2^2 + z_2^2} - \tau_1\tau_2 \\ &\leq \tau_1\tau_2 - \tau_1\tau_2 \\ &= 0. \end{aligned}$$

Finally,

$$\|e_1 + e_2\|^2 = \|e_1\|^2 + \|e_2\|^2 + 2e_1 \cdot e_2 \leq \|e_1\|^2 + \|e_2\|^2 \leq 0.$$

□

Theorem 6.3. *The causality relation $\prec$ is transitive, i.e., if $e_1 \prec e_2$ and $e_2 \prec e_3$, then $e_1 \prec e_3$.*

Proof. Note that

$$\mathbf{0} \prec e_2 - e_1, \mathbf{0} \prec e_3 - e_2.$$

According to Lemma 6.2,

$$e_3 - e_1 = (e_3 - e_2) + (e_2 - e_1) \succ \mathbf{0},$$

which means that $e_1 \prec e_3$. □

The causality relation is a *partial order*, satisfying:

- Reflexivity: $e_1 \prec e_1$
- Antisymmetry: $e_1 \prec e_2, e_2 \prec e_1 \Rightarrow e_1 = e_2$
- Transitivity: $e_1 \prec e_2, e_2 \prec e_3 \Rightarrow e_1 \prec e_3$.

The reflexivity is obvious and the transitivity is showed in the previous theorem. For the antisymmetry, assume that

$$e_1 \prec e_2, \quad e_2 \prec e_1$$

for $e_i = (x_i, y_i, z_i, \tau_i)$. Since neither $e_2 - e_1$ nor $e_1 - e_2$ is past-directed,

$$\tau_1 \leq \tau_2, \quad \tau_2 \leq \tau_1,$$

which implies $\tau_1 = \tau_2$. The event $e_2 - e_1$ is not spacelike. Hence,

$$\begin{aligned} 0 &\geq \|e_2 - e_1\|^2 \\ &= (x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2 - (\tau_1 - \tau_2)^2 \\ &= (x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2 \end{aligned}$$

and so

$$(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2 = 0.$$

We conclude that

$$x_1 = x_2, \quad y_1 = y_2, \quad z_1 = z_2$$

and accordingly $e_1 = e_2$.

In general, an isometry does not need to preserve the causality relation. For example, if $\phi(x, y, z, \tau) = (x, y, z, -\tau)$, then ϕ is an isometry but does not preserve the causality relation.

Definition 6.4. An isometry ϕ of $\mathbb{R}^{3,1}$ is said to be *causal* if it preserves the causality relation, i.e., $\phi(e_1) \prec \phi(e_2)$ for any two events $e_1, e_2 \in \mathbb{R}^{3,1}$ with $e_1 \prec e_2$.

It is known that if an isometry ϕ of $\mathbb{R}^{3,1}$ is not causal, then it is causality-reversing, i.e. $\phi(e_1) \succ \phi(e_2)$ for any $e_1, e_2 \in \mathbb{R}^{3,1}$ with $e_1 \prec e_2$ (see Corollary 6.8 for the case of $\mathbb{R}^{1,1}$).

Example 6.1. A four-dimensional Lorentz boost is a causal isometry that is defined as follows:

$$B_\lambda : \mathbb{R}^{3,1} \to \mathbb{R}^{3,1}$$

$$B_\lambda(x, y, z, \tau) = (x', y', z', \tau'),$$

$$\begin{cases} x' = x\cosh\lambda + \tau\sinh\lambda, \\ y' = y, \\ z' = z, \\ \tau' = x\sinh\lambda + \tau\cosh\lambda \end{cases}$$

for $\lambda \in \mathbb{R}$. Suppose $e_1 \prec e_2$ with $e_i = (x_i, y_i, z_i, \tau_i)$ for $i = 1, 2$. Then, $\tau_1 \leq \tau_2$, and

$$\|e_1 - e_2\|^2 = (x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2 - (\tau_1 - \tau_2)^2 \leq 0,$$

which implies that

$$\begin{aligned} |x_1 - x_2| &= \sqrt{(x_1 - x_2)^2} \\ &\leq \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2} \\ &\leq \sqrt{(\tau_1 - \tau_2)^2} \\ &= |\tau_1 - \tau_2| \\ &= \tau_2 - \tau_1. \end{aligned}$$

Let

$$e_i' = B_\lambda(e_i) = (x_i', y_i', z_i', \tau_i').$$

It is easy to check that B_λ is an isometry of $\mathbb{R}^{3,1}$ (Proposition 6.10),

$$\|e_1' - e_2'\|^2 = \|e_1 - e_2\|^2 \leq 0.$$

Moreover,

$$\begin{aligned} \tau_1' - \tau_2' &= x_1\sinh\lambda + \tau_1\cosh\lambda - x_2\sinh\lambda - \tau_2\cosh\lambda \\ &= (x_1 - x_2)\sinh\lambda + (\tau_1 - \tau_2)\cosh\lambda \\ &\leq |x_1 - x_2||\sinh\lambda| + (\tau_1 - \tau_2)\cosh\lambda \\ &\leq (\tau_2 - \tau_1)|\sinh\lambda| + (\tau_1 - \tau_2)\cosh\lambda \\ &= (\tau_2 - \tau_1)\,(|\sinh\lambda| - \cosh\lambda) \\ &\leq 0. \end{aligned}$$

Hence, $e_1' \prec e_2'$. Therefore, B_λ is a causal isometry.

It is also trivial to check that a translation t_α by an event $\alpha \in \mathbb{R}^{3,1}$ is a causal isometry.

A causal isometry has the following properties which seem natural.

Theorem 6.5.

a) A composition of two causal isometries is causal.
b) The inverse of a causal isometry is causal.

Proof.

a) Let ϕ_1 and ϕ_2 be causal isometries of $\mathbb{R}^{3,1}$. Then,

$$e_1 \prec e_2 \Rightarrow \phi_1(e_1) \prec \phi_1(e_2) \Rightarrow \phi_2(\phi_1(e_1)) \prec \phi_2(\phi_1(e_2)).$$

Therefore, the composition $\phi_2 \circ \phi_1$ is causal.

b) Let ϕ be a causal isometry of $\mathbb{R}^{3,1}$. Suppose that ϕ^{-1} is not causal. Then,

$$\phi^{-1}(e_1) \not\prec \phi^{-1}(e_2)$$

for some events $e_1, e_2 \in \mathbb{R}^{3,1}$ with $e_1 \prec e_2$. Since ϕ^{-1} is also an isometry,

$$\begin{aligned} \|\phi^{-1}(e_2) - \phi^{-1}(e_1)\|^2 &= d^{\mathrm{II}}(\phi^{-1}(e_2), \phi^{-1}(e_1)) \\ &= d^{\mathrm{II}}(e_2, e_1) \\ &= \|e_2 - e_1\|^2 \\ &\leq 0. \end{aligned}$$

Hence, $\phi^{-1}(e_2) - \phi^{-1}(e_1)$ is past-directed because otherwise, $\phi^{-1}(e_1) \prec \phi^{-1}(e_2)$. Now $\phi^{-1}(e_1) - \phi^{-1}(e_2)$ is future-directed, and therefore,

$$\phi^{-1}(e_2) \prec \phi^{-1}(e_1).$$

Note that

$$\phi(\phi^{-1}(e_2)) \prec \phi(\phi^{-1}(e_1)),$$

i.e.,

$$e_2 \prec e_1.$$

Then, $e_1 = e_2$ and hence, we have

$$\phi^{-1}(e_1) = \phi^{-1}(e_2),$$

which implies

$$\phi^{-1}(e_1) \prec \phi^{-1}(e_2).$$

Now we have a contradiction and so ϕ^{-1} is a causal isometry.

□

Mapping $\mathbb{R}^{3,1}$ by a causal isometry in geometry is equivalent to recording events in the spacetime by another observer in physics. Hence, the physical saying

"An event (x, y, z, τ) is recorded as (x', y', z', τ') by another observer."

is equivalent to the geometrical saying

"An event (x, y, z, τ) is mapped to (x', y', z', τ') by a causal isometry."

Exercises

6.3. Suppose that $e_1 \prec e_2$ and $e_3 \prec e_4$ for events e_1, e_2, e_3, and e_4. Show that

a) $e_1 + e_3 \prec e_2 + e_4$.
b) $e_1 - e_4 \prec e_2 - e_3$.

6.4. For $v \in \mathbb{R}$ with $|v| < 1$, the map

$$L_v : \mathbb{R}^{3,1} \to \mathbb{R}^{3,1},$$

defined by

$$L_v(x, y, z, \tau) = (x', y', z', \tau'),$$

$$\begin{cases} x' = \frac{1}{\sqrt{1-v^2}}(x - v\tau), \\ y' = y, \\ z' = z, \\ \tau' = \frac{1}{\sqrt{1-v^2}}(\tau - vx), \end{cases}$$

is called a *Lorentz transformation*. Show that

$$L_v = B_\lambda$$

for some $\lambda \in \mathbb{R}$.

6.3 Causal Isometry

One can define the notion of causality for $\mathbb{R}^{1,1}$ similarly. Using knowledge about isometries of $\mathbb{R}^{1,1}$ from Chapter 5, we will show numerous interesting facts about casualty in $\mathbb{R}^{1,1}$ and $\mathbb{R}^{3,1}$ in this section.

Lemma 6.6. *For any events $e_1, e_2 \in \mathbb{R}^{1,1}$ with $e_1 \prec e_2$,*

1. *$T_A(e_1) \prec T_A(e_2)$ if $A = R(\lambda)$ or $\Lambda(\lambda)$ for some λ,*
2. *$T_A(e_1) \succ T_A(e_2)$ if $A = -R(\lambda)$ or $-\Lambda(\lambda)$ for some λ.*

Proof. Let $e_i = (x_i, \tau_i)$ and $e_i' = (x_i', \tau_i') = T_A(e_i)$. Note that $\tau_1 \leq \tau_2$ and

$$(x_1 - x_2)^2 - (\tau_1 - \tau_2)^2 \leq 0.$$

Hence,

$$|x_1 - x_2| \leq \tau_2 - \tau_1.$$

Since T_A is an isometry of $\mathbb{R}^{1,1}$ for all the cases, we need to show that $\tau_1' \leq \tau_2'$ if $A = R(\lambda)$ or $\Lambda(\lambda)$ and $\tau_1' \geq \tau_2'$ if $A = -R(\lambda)$ or $-\Lambda(\lambda)$. These can be shown by direct calculation. For example, if $A = \Lambda(\lambda)$, then

$$\begin{aligned} x' &= -x \cosh\lambda + \tau \sinh\lambda, \\ \tau' &= -x \sinh\lambda + \tau \cosh\lambda. \end{aligned}$$

Hence,

$$\begin{aligned} \tau_1' - \tau_2' &= -x_1 \sinh\lambda + \tau_1 \cosh\lambda + x_2 \sinh\lambda - \tau_2 \cosh\lambda \\ &= -(x_1 - x_2) \sinh\lambda + (\tau_1 - \tau_2) \cosh\lambda \\ &\leq |x_1 - x_2| |\sinh\lambda| + (\tau_1 - \tau_2) \cosh\lambda \\ &\leq (\tau_2 - \tau_1) |\sinh\lambda| + (\tau_1 - \tau_2) \cosh\lambda \\ &= (\tau_2 - \tau_1) (|\sinh\lambda| - \cosh\lambda) \\ &\leq 0 \end{aligned}$$

and so $\tau_1' \leq \tau_2'$. We leave as an exercise (Exercise 6.5) the calculation for the rest of cases. □

The following theorem describes causal isometries of $\mathbb{R}^{1,1}$ completely.

Theorem 6.7. *An isometry of $\mathbb{R}^{1,1}$ is causal if and only if it has the form of*

$$t_\alpha \circ T_A,$$

where $A = R(\lambda)$ or $\Lambda(\lambda)$ for some $\lambda \in \mathbb{R}$.

Proof. According to Corollary 5.23, an isometry ϕ of $\mathbb{R}^{1,1}$ has the form

$$\phi = t_\alpha \circ T_A$$

for some J-orthogonal matrix A and event $\alpha \in \mathbb{R}^{1,1}$. By Lemma 5.21, $A = \pm R(\lambda)$ or $\pm \Lambda(\lambda)$ for some $\lambda \in \mathbb{R}$. Note that the translation t_α is causal. As seen in Lemma 6.6, T_A is causal only if $A = R(\lambda)$ or $\Lambda(\lambda)$. Hence, the proof is completed. □

An isometry of $\mathbb{R}^{1,1}$ is causal or causality-reversing as shown in the following corollary.

Corollary 6.8. *If an isometry ϕ of $\mathbb{R}^{1,1}$ is not causal, then $\phi(e_1) \succ \phi(e_2)$ for any $e_1, e_2 \in \mathbb{R}^{1,1}$ with $e_1 \prec e_2$.*

Proof. By Corollary 5.23 and Lemma 5.21, an isometry ϕ of $\mathbb{R}^{1,1}$ has the form

$$\phi = t_\alpha \circ T_A$$

for event $\alpha \in \mathbb{R}^{1,1}$, where $A = \pm R(\lambda)$ or $\pm \Lambda(\lambda)$ for some $\lambda \in \mathbb{R}$. By Theorem 6.7, $A = -R(\lambda)$ or $-\Lambda(\lambda)$. As seen in Lemma 6.6, $T_A(e_1) \succ T_A(e_2)$ for events e_1, e_2 with $e_1 \prec e_2$ in this case. Therefore,

$$\begin{aligned} e_1 \prec e_2 &\Rightarrow T_A(e_1) \succ T_A(e_2) \\ &\Rightarrow t_\alpha(T_A(e_1)) \succ t_\alpha(T_A(e_2)) \\ &\Rightarrow \phi(e_1) \succ \phi(e_2). \end{aligned}$$

□

The following lemma will be used in proving Theorem 6.11.

Lemma 6.9. *If two events $e_1, e_2 \in \mathbb{R}^{1,1}$ are causally unrelated, there are causal isometries ϕ_1, ϕ_2, and ϕ_3 of $\mathbb{R}^{1,1}$ such that*

1. *the event $\phi_1(e_1) - \phi_1(e_2)$ is past-directed,*
2. *the events $\phi_2(e_1)$ and $\phi_2(e_2)$ have the same time coordinate,*
3. *the event $\phi_3(e_1) - \phi_3(e_2)$ is future-directed.*

Proof. Note that

$$\|e_1 - e_2\|^2 > 0$$

and so $e_1 - e_2$ is spacelike. Hence,

$$e_1 - e_2 = a(\cosh\lambda, \sinh\lambda)$$

for some $a, \lambda \in \mathbb{R}$ with $a \neq 0$. By Theorem 6.7, the Lorentz boost

$$b_\eta = T_{R(\eta)}$$

is a causal isometry. Note that

$$b_\eta(e_1 - e_2) = a(\cosh(\lambda + \eta), \sinh(\lambda + \eta)).$$

We can find real numbers η_1, η_2, and η_3 such that the numbers

$$a\sinh(\lambda+\eta_1),\ \ a\sinh(\lambda+\eta_2),\ \ a\sinh(\lambda+\eta_3)$$

are negative, zero, positive, respectively.

Let $\phi_i = b_{\eta_i}\circ t_{-e_2}$ for $i=1,2,3$, then ϕ_1, ϕ_2, and ϕ_3 are causal isometries. Note that

$$\begin{aligned}\phi_i(e_1)-\phi_i(e_2) &= b_{\eta_i}(e_1-e_2)-b_{\eta_i}(\mathbf{0})\\ &= a(\cosh(\lambda+\eta_i),\sinh(\lambda+\eta_i))-\mathbf{0}\\ &= (a\cosh(\lambda+\eta_i),a\sinh(\lambda+\eta_i)).\end{aligned}$$

Hence, the event $\phi_1(e_1)-\phi_1(e_2)$ is past-directed, the events $\phi_2(e_1)$ and $\phi_2(e_2)$ have the same time coordinate and the event $\phi_3(e_1)-\phi_3(e_2)$ is future-directed. □

For a map

$$\psi:\mathbb{R}^{1,1}\to\mathbb{R}^{1,1}$$

defined by

$$\psi(x,\tau)=(f(x,\tau),g(x,\tau)),$$

define a map

$$\widetilde{\psi}:\mathbb{R}^{3,1}\to\mathbb{R}^{3,1}$$

by

$$\widetilde{\psi}(x,y,z,\tau)=(f(x,\tau),y,z,g(x,\tau))$$

respectively, where f and g are some functions of x and τ. The map $\widetilde{\psi}$ is said to be induced by the map ϕ. Note that the isometry B_λ of $\mathbb{R}^{3,1}$ in Example 6.1 is induced by the isometry b_λ of $\mathbb{R}^{1,1}$. The following hold:

Proposition 6.10.

1. *The map ψ is an isometry of $\mathbb{R}^{1,1}$ if and only if the map $\widetilde{\psi}$ is an isometry of $\mathbb{R}^{3,1}$.*
2. *The map ψ is a causal isometry of $\mathbb{R}^{1,1}$ if and only if the map $\widetilde{\psi}$ is a causal isometry of $\mathbb{R}^{3,1}$.*

Proof. For an event $e=(x,y,z,\tau)\in\mathbb{R}^{3,1}$, let

$$\widehat{e}=(x,\tau)\in\mathbb{R}^{1,1}.$$

Then

$$\widehat{\widetilde{\psi}(e)} = \psi\,(\widehat{e})\,.$$

First, assume that ψ is an isometry of $\mathbb{R}^{1,1}$. For events $e_1, e_2 \in \mathbb{R}^{3,1}$ with

$$e_i = (x_i, y_i, z_i, \tau_i),$$

we have

$$d^{\mathrm{II}}(\psi(\widehat{e}_1), \psi(\widehat{e}_2)) = d^{\mathrm{II}}(\widehat{e}_1, \widehat{e}_2).$$

Note also

$$\begin{aligned} d^{\mathrm{II}}(e_1, e_2) &= \|e_1 - e_2\|^2 \\ &= (x_1 - x_2)^2 - (\tau_1 - \tau_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2 \\ &= \|\widehat{e}_1 - \widehat{e}_2\|^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2 \\ &= d^{\mathrm{II}}(\widehat{e}_1, \widehat{e}_2) + (y_1 - y_2)^2 + (z_1 - z_2)^2. \end{aligned}$$

Hence,

$$\begin{aligned} d^{\mathrm{II}}(\widetilde{\psi}(e_1), \widetilde{\psi}(e_2)) &= d^{\mathrm{II}}\left(\widehat{\widetilde{\psi}(e_1)}, \widehat{\widetilde{\psi}(e_2)}\right) + (y_1 - y_2)^2 + (z_1 - z_2)^2 \\ &= d^{\mathrm{II}}(\psi(\widehat{e}_1), \psi(\widehat{e}_2)) + (y_1 - y_2)^2 + (z_1 - z_2)^2 \\ &= d^{\mathrm{II}}(\widehat{e}_1, \widehat{e}_2) + (y_1 - y_2)^2 + (z_1 - z_2)^2 \\ &= d^{\mathrm{II}}(e_1, e_2) \end{aligned}$$

and so $\widetilde{\psi}$ is an isometry of $\mathbb{R}^{3,1}$.

Second, assume that $\widetilde{\psi}$ is an isometry of $\mathbb{R}^{3,1}$. Then, for any events $e_1, e_2 \in \mathbb{R}^{3,1}$,

$$d^{\mathrm{II}}(\widetilde{\psi}(e_1), \widetilde{\psi}(e_2)) = d^{\mathrm{II}}(e_1, e_2),$$

i.e.,

$$d^{\mathrm{II}}\left(\widehat{\widetilde{\psi}(e_1)}, \widehat{\widetilde{\psi}(e_2)}\right) + (y_1 - y_2)^2 + (z_1 - z_2)^2 = d^{\mathrm{II}}\,(\widehat{e}_1, \widehat{e}_2) + (y_1 - y_2)^2 + (z_1 - z_2)^2,$$

which implies

$$d^{\mathrm{II}}\,(\psi(\widehat{e}_1), \psi(\widehat{e}_2)) = d^{\mathrm{II}}\,(\widehat{e}_1, \widehat{e}_2)\,.$$

Since we can choose $\widehat{e}_1$ and $\widehat{e}_2$ arbitrarily, we have shown that ψ is an isometry of $\mathbb{R}^{1,1}$.

It is trivial to check that the event $\psi(e_1) - \psi(e_2)$ is past-directed if and only if the event $\widetilde{\psi}(e_1) - \widetilde{\psi}(e_2)$ is past-directed. Hence, the second claim holds. □

For a map $\phi : \mathbb{R}^3 \to \mathbb{R}^3$, define a map

$$\widetilde{\phi} : \mathbb{R}^{3,1} \to \mathbb{R}^{3,1}$$

by

$$\widetilde{\phi}(x, y, z, \tau) = (\phi(x, y, z), \tau).$$

It is easy to see that ϕ is an isometry of $\mathbb{R}^3$ if and only if $\widetilde{\phi}$ is an isometry of $\mathbb{R}^{3,1}$ (Exercise 6.6).

Theorem 6.11. *If two events $e_1, e_2 \in \mathbb{R}^{3,1}$ are causally unrelated, there are causal isometries ϕ_1, ϕ_2, and ϕ_3 of $\mathbb{R}^{3,1}$ such that*

1. *the event $\phi_1(e_1) - \phi_1(e_2)$ is past-directed,*
2. *the events $\phi_2(e_1)$ and $\phi_2(e_2)$ have the same time coordinate,*
3. *the event $\phi_3(e_1) - \phi_3(e_2)$ is future-directed.*

Proof. Since the events e_1, e_2 are causally unrelated,

$$\|e_1 - e_2\|^2 > 0.$$

Let $e_i = (x_i, y_i, z_i, \tau_i)$ for $i = 1, 2$ and

$$e_i' = t_{-e_2}(e_i),$$

i.e., $e_1' = e_1 - e_2$ and $e_2' = \mathbf{0}$. Let

$$e_1' = (x_1', y_1', z_1', \tau_1') = (\alpha, \tau_1'),$$

where $\alpha = (x_1', y_1', z_1')$. Note that

$$\|\alpha\|^2 - {\tau_1'}^2 = \|e_1'\|^2 = \|e_1 - e_2\|^2 > 0.$$

Let $\beta = (\|\alpha\|, 0, 0,) \in \mathbb{R}^3$.

If $\alpha \neq \beta$, consider a plane $P = P_{\alpha,\beta}$ in $\mathbb{R}^3$. Then the plane P contains the origin of $\mathbb{R}^3$, $\bar{r}_P(\mathbf{0}_{\mathbb{R}^3}) = \mathbf{0}_{\mathbb{R}^3}$ and

$$\bar{r}_P(\alpha) = \beta,$$

where $\mathbf{0}_{\mathbb{R}^3}$ is the origin of $\mathbb{R}^3$. Note that $\bar{r}_P$ induces an isometry $\widetilde{\bar{r}}_P$ of $\mathbb{R}^{3,1}$. Let

$$e_i'' = \widetilde{\bar{r}}_P(e_i').$$

Then $e_2'' = \mathbf{0}$

$$e_1'' = (\bar{r}_P(\alpha), \tau_1') = (\beta, \tau_1') = (\|\alpha\|, 0, 0, \tau_1').$$

If $\alpha \neq \beta$, let $e_i'' = e_i'$.
Since

$$\left\| \widehat{e_1''} - \widehat{e_2''} \right\|^2 = \|\alpha\|^2 - {\tau_1'}^2 > 0,$$

the events $\widehat{e_1''}$, $\widehat{e_2''}$ of $\mathbb{R}^{1,1}$ are causally unrelated. According to Lemma 6.9, there are causal isometries ψ_1, ψ_2, and ψ_3 of $\mathbb{R}^{1,1}$ such that the event

$$\psi_1\left(\widehat{e_1''}\right) - \psi_1\left(\widehat{e_2''}\right)$$

is past-directed, the events $\psi_2\left(\widehat{e_1''}\right)$ and $\psi_2\left(\widehat{e_2''}\right)$ have the same time coordinate and the event

$$\psi_3\left(\widehat{e_1''}\right) - \psi_3\left(\widehat{e_2''}\right)$$

is future-directed.
If $\alpha \neq \beta$, let

$$\phi_i = \widetilde{\psi}_i \circ \widetilde{\bar{r}}_P \circ t_{-e_2}$$

for $i = 1, 2, 3$. If $\alpha = \beta$, let

$$\phi_i = \widetilde{\psi}_i \circ t_{-e_2}$$

for $i = 1, 2, 3$.
Then ϕ_1, ϕ_2, and ϕ_3 are causal isometries of $\mathbb{R}^{3,1}$. Noting that

$$\phi_i(e_1) - \phi_i(e_2) = \widetilde{\psi}_i(e_1'') - \widetilde{\psi}_i(e_2'').$$

we have

$$\widehat{\phi_i(e_1)} - \widehat{\phi_i(e_2)} = \psi_i\left(\widehat{e_1''}\right) - \psi_i\left(\widehat{e_2''}\right).$$

Therefore, the event $\phi_1(e_1) - \phi_1(e_2)$ is past-directed, the events $\phi_2(e_1)$ and $\phi_2(e_2)$ have the same time coordinate and the event $\phi_3(e_1) - \phi_3(e_2)$ is future-directed. □

Hence, if two events e_1, e_2 are causally unrelated, then their time order is not absolute—some observer may record that the event e_1 occurs before e_2 while some other observer may record that the event e_1 occurs after e_2. Therefore, we conclude

that two events e_1, e_2 should be causally related if one of them is a cause or an effect of another.

Suppose that events e_1 and e_2 are causally related. The following theorem guarantees that there is an observer for whom the two events e_1 and e_2 occupy the same spatial position at different moments.

Theorem 6.12. *If two events e_1 and e_2 of $\mathbb{R}^{3,1}$ are causally related, there is a causal isometry ϕ of $\mathbb{R}^{3,1}$ such that*

$$\phi(e_1) = (0, 0, 0, \tau_1), \ \ \phi(e_1) = (0, 0, 0, \tau_2)$$

for some τ_1, τ_2

Proof. We will use arguments similar to those in the proof of Theorem 6.11. Since the events e_1, e_2 are causally related,

$$\|e_1 - e_2\|^2 \leq 0.$$

Let $e_i = (x_i, y_i, z_i, \tau_i)$ for $i = 1, 2$ and

$$e_i' = t_{-e_2}(e_i),$$

then $e_2' = \mathbf{0}$. Let

$$e_1' = (x_1', y_1', z_1', \tau_1') = (\alpha, \tau_1'),$$

where $\alpha = (x_1', y_1', z_1')$. Note that

$$\|\alpha\|^2 - {\tau_1'}^2 = \|e_1'\|^2 = \|e_1 - e_2\|^2 \leq 0.$$

Let $\beta = (\|\alpha\|, 0, 0,) \in \mathbb{R}^3$.

If $\alpha \neq \beta$, consider a plane $P = P_{\alpha,\beta}$ in $\mathbb{R}^3$. Then the plane P contains the origin of $\mathbb{R}^3$, $\bar{r}_P(\mathbf{0}_{\mathbb{R}^3}) = \mathbf{0}_{\mathbb{R}^3}$ and

$$\bar{r}_P(\alpha) = \beta,$$

where $\mathbf{0}_{\mathbb{R}^3}$ is the origin of $\mathbb{R}^3$. Note that $\bar{r}_P$ induces an isometry $\widetilde{\bar{r}}_P$ of $\mathbb{R}^{3,1}$. Let

$$e_i'' = \widetilde{\bar{r}}_P(e_i').$$

Then $e_2'' = \mathbf{0}$ and

$$e_1'' = (\bar{r}_P(\alpha), \tau_1') = (\beta, \tau_1') = (\|\alpha\|, 0, 0, \tau_1').$$

If $\alpha = \beta$, let $e_i'' = e_i'$ for $i = 1, 2$.

Since

$$\left\|\widehat{e_1''}\right\|^2 = \|\alpha\|^2 - \tau_1'^2 \leq 0,$$

$$\widehat{e_1''} = (\|\alpha\|, \tau_1') = a(\sinh\lambda, \cosh\lambda)$$

for some a, λ. Then

$$b_{-\lambda}(\widehat{e_1''}) = a(\sinh(\lambda - \lambda), \cosh(\lambda - \lambda)) = (0, a)$$

(see Exercise 5.13).

If $\alpha \neq \beta$, let

$$\phi = \widetilde{b_{-\lambda}} \circ \widetilde{r}_P \circ t_{-e_2}.$$

If $\alpha = \beta$, let

$$\phi = \widetilde{b_{-\lambda}} \circ t_{-e_2}.$$

Then ϕ is a causal isometry of $\mathbb{R}^{3,1}$. Finally, we have

$$\phi(e_1) = \widetilde{b_{-\lambda}}(e_1'') = (0, 0, 0, a)$$

and

$$\phi(e_2) = \widetilde{b_{-\lambda}}(\mathbf{0}) = \mathbf{0} = (0, 0, 0, 0).$$

□

Exercises

6.5. Complete the proof of Lemma 6.6.

6.6. Let $\phi : \mathbb{R}^3 \to \mathbb{R}^3$ be an isometry of $\mathbb{R}^3$. Show that the map

$$\widetilde{\phi} : \mathbb{R}^{3,1} \to \mathbb{R}^{3,1},$$

defined as

$$\widetilde{\phi}(x, y, z, \tau) = (\phi(x, y, z), \tau),$$

is an isometry of $\mathbb{R}^{3,1}$.

6.7. Consider events $e_1 = \mathbf{0}$, $e_2 = (a, 0, 0, b)$ with $a > b > 0$. Note that $e_2 - e_1$ is future-directed. Show that there is a four-dimensional Lorentz boost B_λ such that $B_\lambda(e_2) - B_\lambda(e_1)$ is past-directed.

6.8. Show that a causal orientation-preserving isometry of $\mathbb{R}^{1,1}$ has the form of

$$t_\alpha \circ b_\lambda .$$

for some $\alpha \in \mathbb{R}^{1,1}$ and some $\lambda \in \mathbb{R}$.

6.4 Worldline

We now return to the discussion of the worldlines of objects in $\mathbb{R}^{3,1}$. A sequence of spacetime events on the worldline represents the history of the object, and these events should thus be causally related to one another. A curve in $\mathbb{R}^{3,1}$ is called *causal* if each pair of events on it is causally related. There are infinitely many ways of parameterizing a worldline, but there is a natural condition, called "future-directed." A parameterized curve $\gamma(t) = (x(t), y(t), z(t), \tau(t))$ is said to be *future-directed* (*past-directed*) if

$$\frac{d\tau(t)}{dt} > 0 \quad \left(\frac{d\tau(t)}{dt} < 0, \quad \text{resp.} \right)$$

for any t.

Proposition 6.13. *A parametrization of a causal curve is either future-directed or past-directed.*

Proof. Let

$$\gamma(t) = (x(t), y(t), z(t), \tau(t)) = (\alpha(t), \tau(t))$$

be a parametrization of a causal curve, where $\alpha(t) = (x(t), y(t), z(t)) \in \mathbb{R}^3$.

Suppose that

$$\left. \frac{d\tau(t)}{dt} \right|_{t=t_0} = 0$$

for some t_0. Note that a parametrization is assumed to be regular. Hence,

$$\left. \frac{d\gamma(t)}{dt} \right|_{t=t_0} \neq \mathbf{0},$$

$$\left.\frac{d\alpha(t)}{dt}\right|_{t=t_0} = \left.\left(\frac{dx(t)}{dt}, \frac{dy(t)}{dt}, \frac{dz(t)}{dt}\right)\right|_{t=t_0} \neq (0,0,0),$$

and therefore,

$$\left.\left\|\frac{d\alpha(t)}{dt}\right\|^2\right|_{t=t_0} = \left.\left(\left(\frac{dx(t)}{dt}\right)^2 + \left(\frac{dy(t)}{dt}\right)^2 + \left(\frac{dz(t)}{dt}\right)^2\right)\right|_{t=t_0} > 0.$$

Hence, at least one of

$$\left.\frac{dx(t)}{dt}\right|_{t=t_0}, \quad \left.\frac{dy(t)}{dt}\right|_{t=t_0}, \quad \left.\frac{dz(t)}{dt}\right|_{t=t_0}$$

is non-zero, say

$$\left.\frac{dx(t)}{dt}\right|_{t=t_0} \neq 0.$$

By continuity, there is some positive number $\epsilon > 0$ such that

$$\left|\frac{dx(t)}{dt}\right| > \frac{a}{2}, \qquad \left|\frac{d\tau(t)}{dt}\right| < \frac{a}{2}$$

for any $t \in [t_0 - \epsilon, t_0 + \epsilon]$, where

$$a = \left|\left.\frac{dx(t)}{dt}\right|_{t=t_0}\right| > 0.$$

By the mean value theorem,

$$\frac{x(t_0+\epsilon) - x(t_0)}{\epsilon} = \left.\frac{dx(t)}{dt}\right|_{t=c_1}, \qquad \frac{\tau(t_0+\epsilon) - \tau(t_0)}{\epsilon} = \left.\frac{d\tau(t)}{dt}\right|_{t=c_2}$$

for some $c_1, c_2 \in [t_0, t_0 + \epsilon]$. Hence,

$$\|\alpha(t_0+\epsilon) - \alpha(t_0)\|^2 \geq |x(t_0+\epsilon) - x(t_0)|^2 = \epsilon^2 \left|\left.\frac{dx(t)}{dt}\right|_{t=c_1}\right|^2 > \epsilon^2 \frac{a^2}{4}$$

and

$$(\tau(t_0+\epsilon) - \tau(t_0))^2 = \epsilon^2 \left(\left.\frac{d\tau(t)}{dt}\right|_{t=c_2}\right)^2 < \epsilon^2 \frac{a^2}{4}.$$

Note that $\gamma(t_0+\epsilon)$ and $\gamma(t_0)$ are causally related. Therefore,

$$\|\gamma(t_0+\epsilon)-\gamma(t_0)\|^2 \le 0.$$

However,

$$\|\gamma(t_0+\epsilon)-\gamma(t_0)\|^2 = \|\alpha(t_0+\epsilon)-\alpha(t_0)\|^2-(\tau(t_0+\epsilon)-\tau(t_0))^2 > \epsilon^2\frac{a^2}{4}-\epsilon^2\frac{a^2}{4}=0,$$

which is a contradiction. Therefore,

$$\frac{d\tau(t)}{dt} \neq 0$$

for any t. Therefore, by continuity,

$$\frac{d\tau(t)}{dt}$$

is always positive or negative, which implies the statement in the theorem. □

The following theorem characterizes the causality of curves.

Theorem 6.14. *A parameterized curve γ is causal if and only if*

$$\frac{d\gamma(t)}{dt}$$

is not spacelike.

Proof. First, assume that the curve γ is causal. Then, the events $\gamma(t+\Delta t)$ and $\gamma(t)$ are causally related for any t and Δt. Therefore,

$$\|\gamma(t+\Delta t)-\gamma(t)\|^2 \le 0$$

and

$$\left\|\frac{d\gamma(t)}{dt}\right\|^2 = \lim_{\Delta t\to 0}\left\|\frac{\gamma(t+\Delta t)-\gamma(t)}{\Delta t}\right\|^2 = \lim_{\Delta t\to 0}\frac{\|\gamma(t+\Delta t)-\gamma(t)\|^2}{|\Delta t|^2} \le 0.$$

Therefore, $\frac{d\gamma(t)}{dt}$ is not spacelike.

Conversely, assume $\frac{d\gamma(t)}{dt}$ is not spacelike, i.e.,

$$\left\|\frac{d\gamma(t)}{dt}\right\|^2 \le 0 \tag{6.5}$$

for any t. Let $\gamma(t)=(\alpha(t),\tau(t))$. If

$$\frac{d\tau(t)}{dt} = 0$$

for some t, then

$$0 \geq \left\| \frac{d\gamma(t)}{dt} \right\|^2 = \left\| \frac{d\alpha(t)}{dt} \right\|^2 - \left| \frac{d\tau(t)}{dt} \right|^2 = \left\| \frac{d\alpha(t)}{dt} \right\|^2 \geq 0.$$

Therefore,

$$\left\| \frac{d\alpha(t)}{dt} \right\|^2 = 0$$

and

$$\frac{d\gamma(t)}{dt} = \mathbf{0},$$

which is contradictory to the regularity of γ. Hence,

$$\frac{d\tau(t)}{dt} \neq 0$$

for any t. Therefore, by continuity,

$$\frac{d\tau(t)}{dt}$$

is always positive or negative. Suppose that the curve γ is not causal; then, there are some a and b ($a < b$) such that $\gamma(a)$ and $\gamma(b)$ are not causally related. Then,

$$\|\gamma(a) - \gamma(b)\|^2 > 0. \tag{6.6}$$

Note that $\alpha(t)$ is a parameterized curve in $\mathbb{R}^3$ and

$$\int_a^b \left\| \frac{d\alpha(t)}{dt} \right\| dt$$

is the Euclidean length of the curve from $t = a$ to $t = b$. Hence, the length should be greater than or equal to the Euclidean distance between $\alpha(a)$ and $\alpha(b)$:

$$\int_a^b \left\| \frac{d\alpha(t)}{dt} \right\| dt \geq \|\alpha(a) - \alpha(b)\|.$$

However, (6.5) implies that

$$\left\| \frac{d\alpha(t)}{dt} \right\|^2 - \left| \frac{d\tau(t)}{dt} \right|^2 \leq 0,$$

i.e.,

$$\left\| \frac{d\alpha(t)}{dt} \right\| \leq \left| \frac{d\tau(t)}{dt} \right|.$$

Hence,

$$\|\alpha(a)-\alpha(b)\| \leq \int_a^b \left\| \frac{d\alpha(t)}{dt} \right\| dt \leq \int_a^b \left| \frac{d\tau(t)}{dt} \right| dt = \left| \int_a^b \frac{d\tau(t)}{dt} dt \right| = |\tau(b)-\tau(a)|,$$

where we used the fact that $\frac{d\tau(t)}{dt}$ is always positive or negative. Then,

$$0 < \|\gamma(a) - \gamma(b)\|^2 = \|\alpha(a) - \alpha(b)\|^2 - (\tau(b) - \tau(a))^2 \leq 0,$$

which is a contradiction. Therefore, the curve γ is causal. □

According to Proposition 6.13 and Theorem 6.14, every worldline has a future-directed parametrization that is not spacelike.

A line l in $\mathbb{R}^{3,1}$ is a subset

$$l = \{\alpha + t\beta \mid t \in \mathbb{R}\}$$

of $\mathbb{R}^{3,1}$, where α, β are some events in $\mathbb{R}^{3,1}$ with $\beta \neq \mathbf{0}$. The line l is said to be timelike (spacelike, resp.) if $\|\beta\|^2 < 0$ ($\|\beta\|^2 > 0$, resp.). If $\|\beta\|^2 = 0$, then it is called a lightline.

Each observer has her own coordinate system of the spacetime. Hence, for an observer, there is a corresponding coordinate system of the spacetime, which physicists often call a reference frame. An observer is supposed to record her position as the spatial origin of her own coordinate system. In other words, the worldline of the observer is recorded as $(0, 0, 0, t)$ in her coordinate system for any t because she regards her position the center of spatial space $\mathbb{R}^3$ as we often put the Earth as the center when we observe our universe

Note that mapping $\mathbb{R}^{3,1}$ by a causal isometry is equivalent to recording the spacetime by an observer—for an observer, there is a causal isometry. Therefore, if a curve $\gamma(t)$ is the worldline of an observer, then

$$\phi(\gamma(t)) = (0, 0, 0, t)$$

i.e.,

$$\gamma(t) = \phi^{-1}(0, 0, 0, t)$$

for any t, where ϕ is a causal isometry ϕ of $\mathbb{R}^{3,1}$, corresponding to the observer.

The following holds in general.

Theorem 6.15. *A curve is the worldline of an observer if and only if it is a timelike line.*

Proof. We give a proof for the case of $\mathbb{R}^{1,1}$ and refer to [17] for the general case. Let $\gamma(t)$ be the worldline of an observer, then $\gamma(t) = \phi^{-1}(0,t)$ for some causal isometry of $\mathbb{R}^{1,1}$. Note that ϕ^{-1} is also a causal isometry. According to Corollary 5.23,ϕ has the form

$$\phi^{-1} = t_\alpha \circ T_A$$

for some J-orthogonal matrix A and event $\alpha \in \mathbb{R}^{1,1}$. Let

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix},$$

then $c^2 - d^2 = -1$. Let $\beta = (c,d)$, then $\|\beta\|^2 = -1$ and so β is a timelike event.

$$\gamma(t) = \phi^{-1}(0,t) = (t_\alpha \circ T_A)(0,t) = \alpha + t\beta,$$

which is a parametrization for a timelike line.

Conversely, let

$$l = \{\alpha + t\beta \mid t \in \mathbb{R}\}$$

be a timelike line. Then $\beta = (c,d)$ is a timelike event. We can assume that

$$\beta = (\sinh\lambda, \cosh\lambda)$$

for some $\lambda \in \mathbb{R}$ by replacing β by $a\beta$ for some $a \in \mathbb{R}$ if necessary. Let $\phi = b_{-\lambda} \circ t_{-\alpha}$, then ϕ is a causal isometry. Note that

$$\phi^{-1} = t_\alpha \circ b_\lambda.$$

Hence, the worldline of the observer, corresponding to the causal isometry ϕ, is

$$\gamma(t) = \phi^{-1}(0,t) = (t_\alpha \circ b_\lambda)(0,t) = \alpha + t\beta,$$

which is a parametrization of the line l. □

A timelike line describes a motion of constant velocity (see Definition 6.17 for an official definition of the velocity). The fact that any timelike line can be the worldline of an observer is closely related with the following physical principle of relativity.

"The laws of physics are identical in every inertial frame of reference (i.e., every non-accelerating frame of reference)"

For a timelike line l, there is an observer whose worldline is the line. Let $e_1, e_2 \in L$. Note that

$$\phi^{-1}(e_1) = (0, 0, 0, \tau_1), \ \ \phi^{-1}(e_2) = (0, 0, 0, \tau_2)$$

for some τ_1, τ_2, where ϕ is the causal isometry corresponding to the observer. Suppose that the observer is carrying a clock. The two events e_1, e_2 occur at τ_1, τ_2, respectively, according to the clock. Hence, the elapsed time on the clock of the observer between the event e_1 and the event e_2 is

$$|\tau_2 - \tau_1|.$$

Note that

$$\begin{aligned} -(\tau_2 - \tau_1)^2 &= \|\phi^{-1}(e_2) - \phi^{-1}(e_1)\|^2 \\ &= d^{\mathrm{II}}(\phi^{-1}(e_2), \phi^{-1}(e_1)) \\ &= d^{\mathrm{II}}(e_2, e_1) \\ &= \|e_2 - e_1\|^2. \end{aligned}$$

Since

$$\|e_2 - e_1\|^2 < 0,$$

$$|\tau_2 - \tau_1| = \sqrt{-\|e_2 - e_1\|^2}. \tag{6.7}$$

Let $\gamma(t)$ $(a \leq t \leq b)$ be the worldline of a traveler (not necessarily an observer) who is carrying a clock. Note that the traveler may accelerate during her journey and the worldline may not be a line. We want to find out the elapse of time from $\gamma(a)$ to $\gamma(b)$ the clock records. Let us approximate the curve $\gamma(t)$ by a series of timelike line segments (Figure 6.6) as

$$a = t_0 < t_1 < \cdots < t_n = b.$$

For each line segment, the time elapse is given by

$$\sqrt{-\|\gamma(t_i) - \gamma(t_{i-1})\|^2}$$

as in (6.7). Hence, the elapse of time from an event $\gamma(a)$ to an event $\gamma(b)$ can be found by taking the limit:

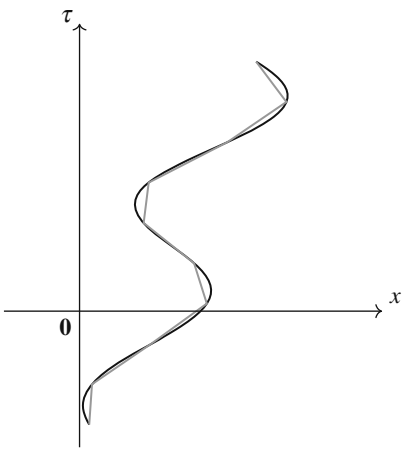


Fig. 6.6 Approximation of a curve by line segments

$$\lim_{n\to\infty}\sum_{i=0}^{n}\sqrt{-\|\gamma(t_i)-\gamma(t_{i-1})\|^2}.$$

For simplicity, assume that

$$t_i = a + \frac{b-a}{n}i = a + i\,\Delta t,$$

where $\Delta t = \frac{b-a}{n}$. Then

$$\lim_{n\to\infty}\sum_{i=0}^{n}\sqrt{-\|\gamma(t_i)-\gamma(t_{i-1})\|^2} = \lim_{n\to\infty}\sum_{i=0}^{n}\sqrt{-\left\|\frac{\gamma(t_i)-\gamma(t_{i-1})}{\Delta t}\right\|^2}\,\Delta t$$
$$= \int_a^b \sqrt{-\left\|\frac{d\gamma}{dt}\right\|^2}\,dt.$$

Note that it is the relativistic length $l_R(\gamma)$ of the curve γ, which is defined in Section 5.6. It is invariant under isometries (Theorem 5.27) and reparametrizations (Exercise 5.27)[1]. It is called the *proper time* along the worldline between those two events. The following theorem addresses the *twin paradox*.

Theorem 6.16. *Let $\gamma(t)$ be a worldline such that $\gamma(t_1) = e_1$, $\gamma(t_2) = e_2$ for two events e_1, e_2. then elapsed proper time along the worldline between e_1, e_2 satisfies the inequality*

[1] It was showed for $\mathbb{R}^{1,1}$ but the same argument also applies to the case of $\mathbb{R}^{3,1}$.

$$\int_{t_1}^{t_2} \sqrt{-\left\| \frac{d\gamma}{dt} \right\|^2} dt \leq \sqrt{-\|e_1 - e_2\|},$$

where the equality holds if and only if the worldline is a line segment between e_1, e_2.

Proof. Let l be a line through the events e_1, e_2. Then l is a timelike line and, by Theorem 6.15, there is an observer whose worldline is the line l. Let ϕ be the corresponding causal isometry. Then $\phi^{-1}(e_1) = (0, \tau_1)$ and $\phi^{-1}(e_2) = (0, \tau_2)$. Note that

$$l_R(\gamma) = l_R(\phi^{-1}(\gamma)),$$

where the length is measured between $t = t_1, t = t_2$. Let

$$\phi^{-1}(\gamma(t)) = (x(t), y(t), z(t), \tau(t)).$$

Then, $x(t) = 0$, $y(t) = 0$ for any t and so

$$\begin{aligned}
l_R(\gamma) &= l_R(\phi^{-1}(\gamma)) \\
&= \int_{t_1}^{t_2} \sqrt{-\left\| \frac{d\phi^{-1}(\gamma(t))}{dt} \right\|^2} dt \\
&= \int_{t_1}^{t_2} \sqrt{-\left(\frac{dx(t)}{dt}\right)^2 - \left(\frac{dy(t)}{dt}\right)^2 - \left(\frac{dz(t)}{dt}\right)^2 + \left(\frac{d\tau(t)}{dt}\right)^2} dt \\
&\leq \int_{t_1}^{t_2} \left| \frac{d\tau(t)}{dt} \right| dt \\
&= \left| \int_{t_1}^{t_2} \frac{d\tau(t)}{dt} dt \right| && (\because \text{Proposition 6.13}) \\
&= |\tau_2 - \tau_1| \\
&= \sqrt{-\|e_1 - e_2\|}. && (\because (6.7))
\end{aligned}$$

The equality holds if and only if $\frac{d\tau(t)}{dt} = 0$ for any t. Note that $\frac{d\tau(t)}{dt} = 0$ for any t if and only if $\phi^{-1}(\gamma)$ is a line, which implies the claim in the theorem. □

Hence, the clock, carried by an accelerated traveler, appears to run more slow than that of a non-accelerating traveler. This phenomenon is called the *time-dilation*. The time-dilation is very well-confirmed by many experiments in numerous situations. This effect should be carefully taken into account in correcting the time in global positioning system (GPS), in order to keep the system to work reliably.

Exercises

6.9. A traveler on a spaceship takes a journey to a remote star. The spaceship uniformly accelerates at a until it reaches the midpoint of the journey and its worldline of the traveler is

$$\gamma(t) = \frac{1}{a}(\cosh t - 1, 0, 0, \sinh t),$$

where $\gamma(0) = \mathbf{0}$ represents the launching event of the spaceship. Note that this parametrization of the worldline is made from the viewpoint of people on the Earth.

Let d be the spatial distance between the Earth and the star.

1. Suppose that the traveler reaches the midpoint of her journey to the star when $t = t_0$. Let T be the proper time elapse of the traveler between $t = 0$, $t = t_0$, and T' be time elapse of people on Earth between $t = 0$, $t = t_0$. Express T, T', and t_0 in terms of a and d.
2. We want the spaceship to accelerate at a comfortable rate that has the effect of mimicking weights of its crews on Earth. Assume that the spaceship accelerates at g, which is the gravitational acceleration on the surface of the Earth, so that its crew experiences the equivalent of a gravitational field with the same strength as that on Earth. The spaceship starts to decelerate at g just after it reaches the midpoint of the journey and eventually stops at the position of the star. It will take twice as long in terms of proper time T for the spaceship to arrive at the destination. Suppose that the spaceship returns to the Earth in the same way. Then the total proper time elapse of the traveler during her journey is $4T$. Suppose that the traveler starts her journey at her tenth birthday, which is January 2nd, 2021. She will come back to the Earth exactly when her age reaches 46, which is her father's age when she starts the journey. How many years pass on Earth when she comes back? Find out the distance d between the Earth and the star.

   Use the light-year for the unit for the distance and the year for the unit for time. g is approximately 1.03 light-year/year^2.

6.5 Kinetics in $\mathbb{R}^{3,1}$

We define a fundamental physical notion for a worldline.

Definition 6.17. Let Γ be the worldline of an object and e be an event on Γ. Let $\gamma(t) = (x(t), y(t), z(t), \tau(t)) = (\alpha(t), \tau(t))$ be a future-directed parametrization of Γ with $\gamma(t_0) = e$. The three-dimensional vector

$$v = \left.\frac{1}{\frac{d\tau(t)}{dt}} \frac{d\alpha(t)}{dt}\right|_{t=t_0}$$

is called the *velocity* of the object at e, and its Euclidean norm

$$\|v\| = \left. \frac{\| \frac{d\alpha(t)}{dt} \|}{\frac{d\tau(t)}{dt}} \right|_{t=t_0}$$

is called the *speed* of the object at e.

It is not difficult to see that the velocity and thus the speed do not depend on parametrization (Exercise 6.11).

Let us derive the formula (6.4) of the addition of speeds. In Figure 6.2, the observer $\mathbf{O}$ records events with coordinates (x, y, z, τ) and another observer $\mathbf{O}'$, moving with speed v relative to $\mathbf{O}$, records events using the coordinates (x', y', z', τ'). Let $\phi(x, y, z, \tau) = (x', y', z', \tau')$. Then the map

$$\phi : \mathbb{R}^{3,1} \to \mathbb{R}^{3,1}$$

is a causal isometry. We can assume that $\phi(\mathbf{0}) = \mathbf{0}$. The worldline of the object $\odot$ with respect to the observer $\mathbf{O}$ is given by

$$\gamma(t) = (ut, 0, 0, t).$$

Since $y' = y$ and $z' = z$, ϕ has form of

$$\phi(x, y, z, \tau) = (f(x, \tau), y, z, g(x, \tau)),$$

where f and g are some functions of x and τ.

From the results of Proposition 6.10 and Exercise 6.8 with $\phi(\mathbf{0}) = \mathbf{0}$, we can assume that ϕ is a Lorentz boost, hence

$$\phi(x, y, z, \tau) = (x \cosh\lambda + \tau \sinh\lambda, y, z, x \sinh\lambda + \tau \cosh\lambda)$$

for some λ.

The worldline of the observer $\mathbf{O}$ with respect to the observer $\mathbf{O}'$ is

$$\alpha(t) = \phi(0, 0, 0, t) = (t \sinh\lambda, 0, 0, t \cosh\lambda)$$

Its velocity is $(v, 0, 0)$ as indicated in Figure 6.2, therefore

$$v = \frac{\sinh\lambda}{\cosh\lambda} = \tanh\lambda.$$

The worldline of the object $\odot$ with respect to the observer $\mathbf{O}'$ is

$$\gamma'(t) = \phi(\gamma(t)) = (ut \cosh\lambda + t \sinh\lambda, 0, 0, ut \sinh\lambda + t \cosh\lambda)$$

and its velocity is

$$\left(\frac{u\cosh\lambda+\sinh\lambda}{u\sinh\lambda+\cosh\lambda},0,0\right)=\left(\frac{u+\tanh\lambda}{u\tanh\lambda+1},0,0\right)=\left(\frac{u+v}{uv+1},0,0\right).$$

Therefore, the speed of the object $\odot$ with respect to the observer $\mathbf{O}'$ is

$$\frac{u+v}{1+uv},$$

which is the formula (6.4) (note that we are using a timescale so that $c=1$).

If the object does not move, then its speed is zero. However, if its worldline is mapped to another one by a causal isometry of $\mathbb{R}^{3,1}$ (physically speaking, if the object is observed by another observer), its velocity and speed may change. Newton's first law states that if there is no force acting on the object, then the velocity of the object is constant. For a photon (the smallest discrete amount of light), its speed is 1 (we are using a timescale that causes the speed of light to be 1), so its worldline is lightlike, i.e.,

$$\left\|\frac{d\gamma(t)}{dt}\right\|^2=0,$$

which implies that

$$\|v\|=\frac{\|\frac{d\alpha(t)}{dt}\|}{\frac{d\tau(t)}{dt}}=1.$$

According to physics, all conventional matter and all known forms of information in the universe can travel at a speed that is less than or equal to the speed of light. We "prove" this fact mathematically.

Theorem 6.18. *The speed of an object is less than or equal to the speed of light at any event of its worldline.*

Proof. Let $\gamma(t)=(x(t),y(t),z(t),\tau(t))=(\alpha(t),\tau(t))$ be a future-directed parametrization of the worldline of the object. Since the worldline is a causal curve,

$$\frac{d\gamma(t)}{dt}=\left(\frac{dx(t)}{dt},\frac{dy(t)}{dt},\frac{dz(t)}{dt},\frac{d\tau(t)}{dt}\right)$$

is not spacelike according to Theorem 6.14. Hence,

$$0\geq\left\|\frac{d\gamma(t)}{dt}\right\|^2=\left\|\frac{d\alpha(t)}{dt}\right\|^2-\left(\frac{d\tau(t)}{dt}\right)^2,$$

i.e.,

$$\frac{d\tau(t)}{dt} \geq \left\| \frac{d\alpha(t)}{dt} \right\|.$$

Therefore,

$$1 \geq \frac{\| \frac{d\alpha(t)}{dt} \|}{\frac{d\tau(t)}{dt}}.$$

□

According to physics, the speed of a massive object is always less than that of light, which means that its worldline is a timelike curve (a photon is considered to be massless). There are many ways to parameterize a worldline; however, a special kind of parametrization is very useful to study the worldline of an object.

Definition 6.19. A parametrization γ of a worldline is said to be *natural* if it is future-directed and

$$\left\| \frac{d\gamma(t)}{dt} \right\|^2 = -1$$

for every t.

Theorem 6.20. *A timelike worldline admits a natural parametrization. Furthermore, if both γ and δ are natural parametrizations of the worldline, then*

$$\gamma(t + a) = \delta(t)$$

for some real number a.

Proof. Let $\gamma : [a, b] \to \mathbb{R}^{3,1}$, $\gamma(t) = (x(t), y(t), z(t), \tau(t))$ be a future-directed parametrization of a worldline. Then,

$$\left\| \frac{d\gamma(t)}{dt} \right\|^2 < 0, \quad \frac{d\tau(t)}{dt} > 0.$$

Define a function $s : [a, b] \to \mathbb{R}$ as

$$s(t) = \int_a^t \sqrt{-\left\| \frac{d\gamma(u)}{du} \right\|^2} du.$$

Then, $s(t)$ is a strictly increasing function, and its image is an interval. Therefore, the function $s(t)$ has an inverse function $t(s)$. Writing $\bar{\gamma}(s) = \gamma(t(s))$, we obtain a reparametrization $\bar{\gamma}$ of the worldline. Note that

$$\frac{ds(t)}{dt} = \sqrt{-\left\|\frac{d\gamma(t)}{dt}\right\|^2}.$$

Let $\bar{\gamma}(s) = (\bar{x}(s), \bar{y}(s), \bar{z}(s), \bar{\tau}(s))$. First,

$$\frac{d\bar{\tau}(s)}{ds} = \frac{\frac{d\tau(t)}{dt}}{\frac{ds(t)}{dt}} = \frac{\frac{d\tau(t)}{dt}}{\sqrt{-\left\|\frac{d\gamma(t)}{dt}\right\|^2}} > 0,$$

which implies that the parametrization $\bar{\gamma}$ is future-directed. In addition,

$$\frac{d\bar{\gamma}(s)}{ds} = \frac{\frac{d\gamma(t)}{dt}}{\frac{ds(t)}{dt}} = \frac{\frac{d\gamma(t)}{dt}}{\sqrt{-\left\|\frac{d\gamma(t)}{dt}\right\|^2}}.$$

Therefore,

$$\left\|\frac{d\bar{\gamma}(s)}{ds}\right\|^2 = \frac{\left\|\frac{d\gamma(t)}{dt}\right\|^2}{-\left\|\frac{d\gamma(t)}{dt}\right\|^2} = -1,$$

which shows that $\bar{\gamma}$ is a natural parametrization.

Let γ and δ be natural parametrizations of the worldline. Then, δ is a reparametrization of γ; therefore, there is a function $u(t)$ such that $\delta(t) = \gamma(u(t))$. Since

$$\begin{aligned} -1 &= \left\|\frac{d\delta(t)}{dt}\right\|^2 \\ &= \left\|\frac{d\gamma(u(t))}{dt}\right\|^2 \\ &= \left\|\frac{d\gamma(u)}{du}\frac{du(t)}{dt}\right\|^2 \\ &= \left\|\frac{d\gamma(u)}{du}\right\|^2 \left|\frac{du(t)}{dt}\right|^2 \\ &= -\left|\frac{du(t)}{dt}\right|^2, \end{aligned}$$

i.e., $\left|\frac{du(t)}{dt}\right|^2 = 1$. Hence, $\frac{du(t)}{dt} = \pm 1$.

Let $\gamma(t) = (x(t), y(t), z(t), \tau(t))$; then,

$$\delta(t) = \gamma(u(t)) = (x(u(t)), y(u(t)), z(u(t)), \tau(u(t))).$$

Since both γ and δ are future-directed,

$$0 < \frac{d\tau(t)}{dt}$$

and

$$0 < \frac{d\tau(u(t))}{dt} = \frac{d\tau(u)}{du}\frac{du(t)}{dt}.$$

Hence, $\frac{du(t)}{dt} > 0$ and

$$\frac{du(t)}{dt} = +1.$$

Finally, $u(t) = t + a$ for a real constant a, and hence,

$$\delta(t) = \gamma(u(t)) = \gamma(t + a).$$

□

The worldline of a massive object is timelike. Therefore, it admits a natural parametrization.

Definition 6.21. Let Γ be the worldline of an object and e be an event on Γ. Let $\gamma(t) = (x(t), y(t), z(t), \tau(t)) = (\alpha(t), \tau(t))$ be a natural parametrization of Γ with $\gamma(t_0) = e$.

$$\left.\frac{d\gamma(t)}{dt}\right|_{t=t_0} = \left.\left(\frac{dx(t)}{dt}, \frac{dy(t)}{dt}, \frac{dz(t)}{dt}, \frac{d\tau(t)}{dt}\right)\right|_{t=t_0}$$

is called the *four-velocity* of the object at e.

One can show that the four-velocity does not depend on the parametrization.

Proposition 6.22. *Let v be the velocity of a massive object at some event. The four-velocity of the object at that event is as follows:*

$$\frac{1}{\sqrt{1 - \|v\|^2}}(v, 1).$$

Proof. Let $\gamma(t) = (x(t), y(t), z(t), \tau(t)) = (\alpha(t), \tau(t))$ be a natural parametrization of the worldline of the object. Then,

$$-1 = \left\| \frac{d\gamma(t)}{dt} \right\|^2 = \left\| \frac{d\alpha(t)}{dt} \right\|^2 - \left(\frac{d\tau(t)}{dt} \right)^2 = \|v\|^2 \left(\frac{\tau(t)}{dt} \right)^2 - \left(\frac{\tau(t)}{dt} \right)^2.$$

Hence,

$$\frac{d\tau(t)}{dt} = \frac{1}{\sqrt{1 - \|v\|^2}}.$$

The four-velocity is

$$\frac{d\gamma(t)}{dt} = \left(\frac{d\alpha(t)}{dt}, \frac{d\tau(t)}{dt} \right) = \frac{d\tau(t)}{dt} \left(\frac{\frac{d\alpha(t)}{dt}}{\frac{d\tau(t)}{dt}}, 1 \right) = \frac{1}{\sqrt{1 - \|v\|^2}} (v, 1).$$

□

Let m be the mass of a massive object; then, the quantity mass times the four-velocity,

$$p = \frac{m}{\sqrt{1 - \|v\|^2}} (v, 1), \tag{6.8}$$

is called the *four-momentum* of the object. Note that $\|p\|^2 = -m^2$. When the speed of the object is much less than the speed of light, which is 1 in our time scale, the motion of the object can be described by classical mechanics. More concretely, if $\|v\|$ is much less than 1 ($\|v\| \ll 1$), then

$$\frac{1}{\sqrt{1 - \|v\|^2}} \approx 1 + \frac{1}{2} \|v\|^2.$$

Hence, if we approximate the four-momentum of the object up to the first term with respect to $\|v\|$,

$$\frac{m}{\sqrt{1 - \|v\|^2}} (v, 1) \approx \left(mv, m + \frac{1}{2} m \|v\|^2 \right).$$

Before proceeding, let us see how accurate this approximation is in the real world. The typical speed of a bullet is approximately 1000 m/s. In our time scale, its speed is

$$\|v\| = \frac{1000 \text{ m/s}}{299792458 \text{ m/s}} = 0.00000333564\ldots$$

(the speed of light is 299792458 m/s). Hence,

$$\frac{1}{\sqrt{1 - \|v\|^2}} = 1 + 5.56311\ldots \times 10^{-12},$$

$$1 + \frac{1}{2}\|v\|^2 = 1 + 5.56325\ldots \times 10^{-12}.$$

Note that mv and $\frac{1}{2}m\|v\|^2$ are the momentum and kinetic energy of the object, respectively, in classical mechanics. Let $p = (p_x, p_y, p_z, E)$. Since

$$E - \sqrt{-\|p\|^2} \approx \frac{1}{2}m\|v\|^2$$

for $\|v\| \ll 1$, which is the classical kinetic energy of the object, we call

$$E_k := E - \sqrt{-\|p\|^2} = \frac{m}{\sqrt{1 - \|v\|^2}} - m$$

the *relativistic kinetic energy* of the object. The quantity E is called the *total energy* of the object (this naming will be justified later). The total energy when the speed is zero is called the *rest energy* of the object, and we denote it by E_0. Note that

$$E = E_k + E_0.$$

For the bullet previously considered, the difference between its classical kinetic energy and its relativistic kinetic energy is less than 0.0025%, which implies that the relativistic effect is negligible when the speed of the object is much less than the speed of light. Electrons are very light and easy to accelerate. In a physics lab, one can easily find an electron whose speed is 99% of the speed of the light, i.e., its speed is 0.99 in our time scale. In this case, the ratio of its relativistic kinetic energy to its classical kinetic energy is greater than 12. Therefore, the relativistic effect is large and cannot be ignored when studying the motion of these electrons.

According to Newton's first law of motion, an object either remains at rest or continues to move at a constant velocity unless acted upon by a force. Therefore, the worldline for such an object is a line in $\mathbb{R}^{3,1}$ that has a constant four-momentum.

Imagine that there are two objects that initially do not interact with each other. If there are no outside forces, they have constant four-momentums p_1 and p_2 according to Newton's first law of motion. Suppose that these two objects collide with each other at some moment and then do not interact with each other afterward; hence, they have constant four-momentums p_1' and p_2' (Figure 6.7). The relativistic version of Newton's third law of motion asserts that

$$p_1 + p_2 = p_1' + p_2'.$$

If there are several particles interacting, then Newton's third law of motion asserts that

$$p_1 + p_2 + \cdots + p_n = p_1' + p_2' + \cdots + p_n'. \tag{6.9}$$

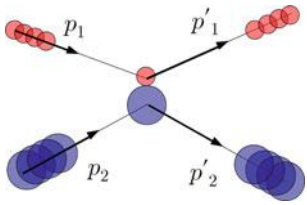


Fig. 6.7 Collision of two objects

If the speeds of the objects are much less than one, there is conservation of momentum and conservation of kinetic energy.

A photon moves at a constant velocity, and its speed is 1, which means that its worldline is lightlike and that it does not admit a natural parametrization. According to physics, a photon interacts with ordinary objects. One can ask if it still has a four-momentum that obeys Newton's third law of motion ((6.9)) during the interaction. The four-momentum in (6.8) is not well-defined for a photon because

$$1 - \|v\|^2 = 0.$$

Instead, one may consider a photon as a limiting case of a massive object whose speed approaches one in (6.8). For the photon to have a well-defined four-momentum, its mass should approach zero in this limiting process. In this way, we expect that a photon is massless. Let $p = (p_x, p_y, p_z, E)$ be the four-momentum of a photon; then,

$$p = \lim_{m \to 0^+} p_m,$$

where $p_m = \frac{m}{\sqrt{1-\|v\|^2}}(v, 1)$. Since $\frac{m}{\sqrt{1-\|v\|^2}} > 0$, we expect that $E > 0$. Noting that $\|p_m\|^2 = -m^2$, we conclude that

$$\|p\|^2 = \lim_{m \to 0^+} (-m^2) = 0.$$

Since $\|p\|^2 = 0$, the relativistic kinetic energy of a photon is

$$E_k = E - \sqrt{-\|p\|^2} = E,$$

which is the total energy.

According to physics, an ordinary object may not travel faster than light. However, if an object continues to be "pushed" forward, its speed will increase. Hence, one can still suspect that a pushed object may travel faster than light.

Imagine a situation in which a photon pushes an object through a collision. Let

$$(p, 0, 0, p), (-p', 0, 0, p')$$

be the four-momentums of the photon before and after the collision ($p,\ p' > 0$). Let

$$\frac{m}{\sqrt{1-v^2}}(v,0,0,1) \text{ and } \frac{m}{\sqrt{1-v'^2}}(v',0,0,1)$$

be the four-momentums of the object before and after the collision, respectively, where v and v' are the speeds of the object before and after the collision, respectively. We assume that the object initially does not travel faster than light, and thus, $v < 1$. By Newton's third law of motion,

$$(p,0,0,p) + \frac{m}{\sqrt{1-v^2}}(v,0,0,1) = (-p',0,0,p') + \frac{m}{\sqrt{1-v'^2}}(v',0,0,1),$$

which results in

$$p + \frac{mv}{\sqrt{1-v^2}} = -p' + \frac{mv'}{\sqrt{1-v'^2}}$$

and

$$p + \frac{m}{\sqrt{1-v^2}} = p' + \frac{m}{\sqrt{1-v'^2}}.$$

Therefore,

$$2p + \frac{(1+v)m}{\sqrt{1-v^2}} = m\frac{1+v'}{\sqrt{1-v'^2}},$$

which yields

$$v' = \frac{m^2v + 2p^2(1-v) + 2mp\sqrt{1-v^2}}{m^2 + 2p^2(1-v) + 2mp\sqrt{1-v^2}} = \frac{b}{a},$$

where

$$a = m^2 + 2p^2(1-v) + 2mp\sqrt{1-v^2},$$

and

$$b = m^2v + 2p^2(1-v) + 2mp\sqrt{1-v^2}.$$

Since $0 \leq v < 1$, $a > 0$, $b > 0$ and

$$a - b = m^2(1-v) > 0.$$

Hence,

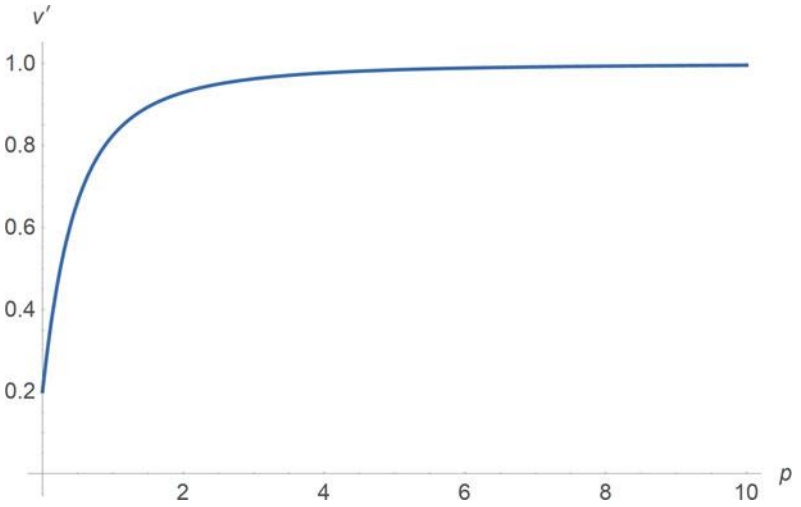


Fig. 6.8 The speed of a massive object cannot reach the speed of light

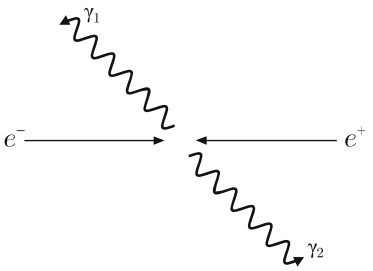


Fig. 6.9 Electron–positron annihilation

$$v' = \frac{b}{a} < 1,$$

which means that the speed of the object after the collision is still less than that of light. Even though the object is pushed forward by a photon, the speed is less than that of light. Hence, regardless of how hard and how many times an object is pushed by photons, its speed cannot reach the speed of light (Figure 6.8).

According to elementary particle physics theory, every type of elementary particle has an associated antiparticle with the same mass but with opposite physical charges (such as electric charge). For example, the antiparticle of the electron is the anti-electron (which is called a positron). While an electron has a negative electric charge, a positron has a positive electric charge. When a positron (e^+) collides with an electron (e^-), annihilation occurs, resulting in the creation of gamma ray photons (Υ_1, Υ_2) (Figure 6.9):

$$e^- + e^+ \to \Upsilon_1 + \Upsilon_2.$$

Let E_-, E_+, E_1, and E_2 be the total energies of the electron, the positron, and the two photons, respectively. According to Newton's third law,

$$E_- + E_+ = E_1 + E_2.$$

Suppose that the electron and the positron do not initially move. Then, the left-hand side of the above equation is the sum of their masses. However, the right-hand side of the above equation includes all the kinetic energies of the photons. Hence, we conclude that

$$2m_e = E_1 + E_2,$$

where m_e is the mass of an electron, which is equal to that of a positron.

Therefore, we can say that all the mass of the electron and the positron turns into kinetic energy of the photons. Since every type of particle has an associated antiparticle and particle–antiparticle pairs can annihilate each other, producing photons, every object of the universe can turn into photons, whose total energy is the mass of the object:

$$E = m.$$

This leads to the famous conclusion by Einstein that mass is just another form of energy (mass–energy equivalence). The above equation involves the speed of light, which is 1 in our scale of time. In the ordinary unit system, the equation of the mass–energy equivalence becomes

$$E = mc^2,$$

where c is the speed of light.

Exercises

6.10. For $u, v \in \mathbb{R}$ with $|u|, |v| < 1$, show that

$$\left| \frac{u + v}{1 + uv} \right| < 1.$$

6.11. Show that the velocity and, accordingly, the speed of an object are independent of how its worldline is parameterized.

6.12. Consider a worldline γ_1 whose speed is less than one at every event of it. Map it to another worldline γ_2 by an isometry $\mathbb{R}^{3,1}$. Show that the speed of γ_2 is still less than one at every event of it.

Answers to Selected Exercises

Chapter 1

1.2 Note that $a \neq 0$ or $b \neq 0$. First assume that $a \neq 0$.
Let

$$p_1 = \left(a - \frac{c}{a}, b\right)$$

and

$$p_2 = \left(-a - \frac{c}{a}, -b\right).$$

Then, for $p = (x, y) \in \mathbb{R}^2$,

$$\begin{aligned}
p \in L_{p_1,p_2} &\Leftrightarrow d(p_1, p) = d(p_2, p) \\
&\Leftrightarrow d(p_1, p)^2 = d(p_2, p)^2 \\
&\Leftrightarrow \left(x - a + \frac{c}{a}\right)^2 + (y - b)^2 = \left(x + a + \frac{c}{a}\right)^2 + (y + b)^2 \\
&\Leftrightarrow 2\left(-a+\frac{c}{a}\right)x+\left(-a+\frac{c}{a}\right)^2 -2by=2\left(a+\frac{c}{a}\right)x+\left(a+\frac{c}{a}\right)^2+2by \\
&\Leftrightarrow 0 = 4ax + 4by + 4c \\
&\Leftrightarrow 0 = ax + by + c.
\end{aligned}$$

Therefore, $L = L_{p_1,p_2}$.
Now assume $a = 0$. Then $b \neq 0$ and the line L is defined by the equation

N.-H. Lee, *Geometry: from Isometries to Special Relativity*, Undergraduate Texts in Mathematics, https://doi.org/10.1007/978-3-030-42101-4

$$y = -\frac{c}{b}.$$

Let

$$p_1 = \left(0, -\frac{c}{b} - 1\right)$$

and

$$p_2 = \left(0, -\frac{c}{b} + 1\right).$$

Then we have $L = L_{p_1,p_2}$ also in this case.

1.3 For an isometry ϕ and a circle

$$C = \{p \in \mathbb{R}^2 \mid d(p, q) = r\},$$

where q is the center and r is the radius, let

$$C' = \{p \in \mathbb{R}^2 \mid d(p, \phi(q)) = r\},$$

a circle of radius r, centered at $\phi(q)$.

We will show that

$$C' = \phi(C).$$

$$\begin{aligned} p \in C' &\Leftrightarrow d(p, \phi(q)) = r \\ &\Leftrightarrow d(\phi^{-1}(p), \phi^{-1}(\phi(q))) = r \\ &\Leftrightarrow d(\phi^{-1}(p), q)) = r \\ &\Leftrightarrow \phi^{-1}(p) \in C \\ &\Leftrightarrow p \in \phi(C). \end{aligned}$$

Hence, $C' = \phi(C)$.

1.6 Note that

$$r_{(a,b),\theta} = t_{(a,b)} \circ r_\theta \circ t_{-(a,b)}.$$

Hence,

$$\begin{aligned} r_{(a,b),\theta}(x, y) &= (t_{(a,b)} \circ r_\theta \circ t_{-(a,b)})(x, y) \\ &= (t_{(a,b)} \circ r_\theta)(x - a, y - b) \\ &= t_{(a,b)}((x - a)\cos\theta - (y - b)\sin\theta, (x - a)\sin\theta + (y - b)\cos\theta) \\ &= (a+(x - a)\cos\theta-(y - b)\sin\theta, b+(x - a)\sin\theta+(y - b)\cos\theta). \end{aligned}$$

1.7

1. For all $p, q \in \mathbb{R}^2$,

$$d(\mathrm{id}_{\mathbb{R}^2}(p), \mathrm{id}_{\mathbb{R}^2}(q)) = d(p, q).$$

   Hence, the identity map is an isometry.
2. Let ϕ be an isometry of $\mathbb{R}^2$. For all $p, q \in \mathbb{R}^2$,

$$\begin{aligned} d(\phi^{-1}(p), \phi^{-1}(q)) &= d(\phi(\phi^{-1}(p)), \phi(\phi^{-1}(q)))) \\ &= d(p, q). \end{aligned}$$

   Hence, ϕ^{-1} is also an isometry.
3. Let ϕ and ψ be isometries of $\mathbb{R}^2$. For all $p, q \in \mathbb{R}^2$,

$$\begin{aligned} d((\phi \circ \psi)(p), (\phi \circ \psi)(q)) &= d(\phi(\psi(p)), \phi(\psi(q))) \\ &= d(\psi(p), \psi(q)) \\ &= d(p, q). \end{aligned}$$

   Hence, $\phi \circ \psi$ is also an isometry.

1.11 Choose non-collinear points p_1, p_2, and p_3. The isometry ϕ maps the triangle $\triangle p_1 p_2 p_3$ to a congruent triangle

$$\triangle\phi(p_1)\phi(p_2)\phi(p_3),$$

which implies that the points $\phi(p_1)$, $\phi(p_2)$, and $\phi(p_3)$ are also non-collinear. As in the proof of Theorem 1.6, by composing reflections, we can build an isometry ψ such that

$$\phi(p_1) = \psi(p_1), \phi(p_2) = \psi(p_2) \text{ and } \phi(p_3) = \psi(p_3).$$

Using arguments similar to those in the proof of Theorem 1.4, we show that $\phi = \psi$. Since ψ is bijective, ϕ is also bijective.

1.13 For given two reflections $\bar{r}_L, \bar{r}_M$ in lines L, M, there is an isometry ψ such that $\psi(M) = L$.

$$^{\psi}\bar{r}_M = \bar{r}_{\psi(M)} = \bar{r}_L.$$

Hence, $\bar{r}_L$ and $\bar{r}_M$ are conjugate.

1.14 Define a map $\phi : \mathbb{R}^2 \to \mathbb{R}^2$ using the formula in the exercise;

$$\phi(p) = p - \frac{2(p-u)\cdot v}{\|v\|^2} v.$$

We must show that $\phi = \bar{r}_L$.

First, for two points α_1 and α_2,

$$\begin{aligned} d(\phi(\alpha_1), \phi(\alpha_2)) &= \|\phi(\alpha_1) - \phi(\alpha_2)\|^2 \\ &= \left\| \alpha_1 - \alpha_2 - \frac{2(\alpha_1 - \alpha_2)\cdot v}{\|v\|^2} v \right\|^2 \\ &= \|\alpha_1 - \alpha_2\|^2 + \frac{4((\alpha_1 - \alpha_2)\cdot v)^2}{\|v\|^4} \|v\|^2 - \frac{4((\alpha_1 - \alpha_2)\cdot v)^2}{\|v\|^2} \\ &= \|\alpha_1 - \alpha_2\|^2 \\ &= d(\alpha_1, \alpha_2). \end{aligned}$$

Therefore, ϕ is an isometry by Exercise 1.11. Then,

$$\phi(p_1) = p_1 - \frac{2(p_1 - u)\cdot v}{\|v\|^2} v = p_1 - \frac{2v\cdot v}{\|v\|^2} v = p_1 - 2v = p_2,$$

and similarly, $\phi(p_2) = p_1$.

If p lies on the line L_{p_1,p_2}, then $d(p, p_1) = d(p, p_2)$, i.e., $\|p - p_1\|^2 = \|p - p_2\|^2$, which implies that

$$\frac{1}{4}(\|p_1\|^2 - \|p_2\|^2) = \frac{1}{2}(p_1 - p_2)\cdot p,$$

i.e., $u \cdot v = v \cdot e$; thus, $(p - u)\cdot v = 0$. Hence,

$$\phi(p) = p - \frac{2(p-u)\cdot v}{\|v\|^2} v = p.$$

Note that p, p_1, and p_2 are non-collinear for $p \in L$, and we showed that

$$\phi(p) = p = \bar{r}_L(p), \phi(p_1) = p_2 = \bar{r}_L(p_1), \phi(p_2) = p_1 = \bar{r}_L(p_2).$$

Therefore, according to Theorem 1.4, $\phi = \bar{r}_L$.

1.23 Consider a glide reflection $\phi = g_{L,\alpha}$. Choose a coordinate system such that L coincides with the x-axis. Then, $\alpha = (a, 0)$ for some a. Then,

$$\phi = g_{L,\alpha} = t_{(a,0)} \circ \bar{r}.$$

For $(x, y) \in \mathbb{R}^2$,

$$\begin{aligned}
\phi^2(x, y) &= (t_{(a,0)} \circ \bar{r})^2(x, y) \\
&= (t_{(a,0)} \circ \bar{r})(a + x, -y) \\
&= (2a + x, y) \\
&= t_{(2a,0)}(x, y) \\
&= t_{2\alpha}(x, y).
\end{aligned}$$

Therefore, $\phi^2 = t_{2\alpha}$, a translation.

1.27 Let $\phi : \mathbb{R} \to \mathbb{R}$ be an isometry. Let $\psi = \bar{r}_a \circ \phi$, where $a = \frac{\phi(0)}{2}$. Then, ψ is an isometry with $\psi(0) = 0$.

Case 1. Assume that $\psi(1) = 1$. We will show that $\psi = \mathrm{id}_{\mathbb{R}}$. Suppose that $\psi(x) \neq x$ for some $x \in \mathbb{R}$. Since

$$|x| = d(x, 0) = d(\psi(x), \psi(0)) = d(\psi(x), 0) = |\psi(x)|,$$

$\psi(x) = -x$. Note

$$|x - 1| = d(x, 1) = d(\psi(x), \psi(1)) = d(\psi(x), 1) = |\psi(x) - 1| = |-x - 1|,$$

which implies $x = 0$. But then $0 = \psi(0) = \psi(x) \neq x = 0$, which is impossible. Hence, $\bar{r}_a \circ \phi = \psi = \mathrm{id}_{\mathbb{R}}$ and so $\phi = \bar{r}_a$.

Case 2. Assume that $\psi(1) \neq 1$. Note

$$|\psi(1)| = |\psi(1) - 0| = d(\psi(1), 0) = d(\psi(1), \psi(0)) = d(1, 0) = 1.$$

Hence $\psi(1) = -1$.

Let $\psi' = \bar{r}_0 \circ \psi$, then $\psi'(1) = 1$. Since $\psi'(0) = 0$, we can use Case 1 to show that $\psi' = \mathrm{id}_{\mathbb{R}}$, i.e.,

$$\bar{r}_0 \circ \bar{r}_a \circ \phi = \mathrm{id}_{\mathbb{R}},$$

which implies $\phi = \bar{r}_a \circ \bar{r}_0$, a composition of two reflections.

Chapter 2

2.5

(a) Let G_x, G_y, and G_z be the intersection of the sphere with the yz-plane, zx-plane, and xy-plane, respectively. It can be readily verified that

$$\hat{a} = \bar{r}_{G_z} \circ \bar{r}_{G_y} \circ \bar{r}_{G_x},$$

which is an orientation-reversing isometry.

(b) If $\phi = \hat{a} \circ r_{l,\theta}$ for some rotation $r_{l,\theta}$, then ϕ can be expressed as a composition of $(3 + 2)$ reflections. Hence, it is orientation-reversing.

Conversely, if ϕ is orientation-reversing, then $\hat{a} \circ \phi$ is orientation-preserving. According to Euler's rotation theorem, it is a rotation $r_{l,\theta}$, i.e.,

$$\hat{a} \circ \phi = r_{l,\theta},$$

which implies that

$$\phi = \hat{a} \circ r_{l,\theta}.$$

2.6 If ϕ is orientation-preserving, then according to Euler's rotation theorem, it is a rotation whose fixed points are composed of two points or infinite points (when it is a rotation by zero angle).

If ϕ is orientation-reversing, then according to Exercise 2.5,

$$\phi = \hat{a} \circ r_{l,\theta}$$

for some line l through the origin and an angle θ $(0 \leq \theta < 2\pi)$. One can choose the coordinate system such that l coincides with the z-axis. Let $p = (x, y, z)$ be a fixed point of ϕ; then,

$$(x, y, z) = p = \phi(p) = -(r_\theta(x, y), z).$$

Hence, $z = 0$, and $\theta = \pi$. Now it is readily seen that all the points on the equator of the sphere are fixed points of ϕ. Note that ϕ is actually a reflection.

In summary, if the isometry ϕ has a fixed point, then it is a rotation or a reflection. Now the claims in the Exercise are readily verified.

2.14 Let p_1, p_2, and p_3 be the vertices of the spherical triangle T, right-angled at p_1. We can choose a coordinate system such that

$$p_1 = (1, 0, 0),\ p_2 = (\cos a, \sin a, 0),\ p_3 = (\cos a, 0, \sin a).$$

Project the triangle T from the origin to the plane $x = 1$, forming a right-angled isosceles triangle with like sides of length a. Hence, the area of its projected image is

$$\frac{1}{2}\left(\frac{\sin a}{\cos a}\right)^2 = \frac{1}{2}\tan^2 a,$$

and it is greater than the area of T. Conversely, project the triangle T from the origin to the plane $x = \cos a$, forming a right-angled isosceles triangle with like sides of length $\sin a$. Hence, the area of its projected image is

$$\frac{1}{2}\sin^2 a,$$

which is less than the area of T. In summary,

$$\frac{1}{2}\tan^2 a > \text{Area}(T) > \frac{1}{2}\sin^2 a,$$

and therefore,

$$\frac{\frac{1}{2}\tan^2 a}{\frac{1}{2}a^2} > \frac{\text{Area}(T)}{\frac{1}{2}a^2} > \frac{\frac{1}{2}\sin^2 a}{\frac{1}{2}a^2}.$$

Therefore,

$$\frac{\text{Area}(T)}{\frac{1}{2}a^2}$$

converges to one as a approaches zero.

Chapter 3

3.3 Let P be the plane in $\mathbb{R}^3$ whose intersection with $\mathbb{S}^2$ is C. Choosing a suitable coordinate system, we can assume that P is orthogonal to the zx-plane. Let $C \cap zx\text{-plane} = \{p_1, p_2\}$. Now we only need to show (why?) that $\Phi(q)$ is the middle point of $\Phi(p_1)$ and $\Phi(p_2)$, i.e.,

$$\Phi(q) = \frac{1}{2}(\Phi(p_1) + \Phi(p_2)).$$

Note that $p_i = (\cos\theta_i, 0, \sin\theta_i)$ for some θ_1 and θ_2, and thus,

$$\Phi(p_i) = \left(\frac{\cos\theta_i}{1-\sin\theta_i}, 0\right).$$

By a direct calculation, one can show that

$$q = \operatorname{cosec}(\theta_1 - \theta_2)((\sin\theta_1 - \sin\theta_2), 0, -(\cos\theta_1 - \cos(\theta_2))).$$

By another direct calculation, one can check that

$$\Phi(q) = \frac{1}{2}(\Phi(p_1) + \Phi(p_2)).$$

3.6

(a) For $\alpha \in \mathbb{R}^2$,

$$I_{C_{\mathbf{0},r}}(\alpha) = \frac{r^2}{\|\alpha\|^2}\alpha.$$

However,

$$\begin{aligned}(d_r \circ I \circ d_{\frac{1}{r}})(\alpha) &= (d_r \circ I)(\alpha/r)\\ &= d_r\left(\frac{1}{\|\alpha/r\|^2}\alpha/r\right)\\ &= d_r\left(\frac{r}{\|\alpha\|^2}\alpha\right)\\ &= \frac{r^2}{\|\alpha\|^2}\alpha.\end{aligned}$$

Hence, $I_{C_{\mathbf{0},r}} = d_r \circ I \circ d_{\frac{1}{r}}$.

(b) For $\alpha \in \mathbb{R}^2$,

$$I_{C_{p,r}}(\alpha) - p = \frac{r^2}{\|\alpha - p\|^2}(\alpha - p).$$

Hence,

$$I_{C_{p,r}}(\alpha) = p + \frac{r^2}{\|\alpha - p\|^2}(\alpha - p) = \left(t_p \circ I_{C_{\mathbf{0},r}} \circ t_{-p}\right)(\alpha).$$

3.13

1. Let

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \quad B = \begin{pmatrix} \alpha & \beta \\ \delta & \gamma \end{pmatrix}.$$

Then

$$f_A\left(f_B(x)\right) = f_A\left(\frac{\alpha x+\beta}{\delta x+\gamma}\right) = \frac{a\frac{\alpha x+\beta}{\delta x+\gamma}+b}{c\frac{\alpha x+\beta}{\delta x+\gamma}+d} = \frac{(a\alpha+b\delta)x+a\beta+d\gamma}{(c\alpha+d\delta)x+c\beta+d\gamma} = f_{AB}(x).$$

Thus, $f_A \circ f_B = f_{AB}$.
Note that

$$f_{I_2} = \mathrm{id}_{\mathbb{R}_\infty}$$

for

$$I_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

and

$$AA^{-1} = A^{-1}A = I_2.$$

Thus

$$f_A \circ f_{A^{-1}} = f_{A^{-1}} \circ f_A = f_{I_2} = \mathrm{id}_{\mathbb{R}_\infty},$$

which means $f_A^{-1} = f_{A^{-1}}$.

2. If none of the elements x_2, x_3, x_4 is ∞, let

$$f(x) = \frac{x-x_3}{x-x_4}\frac{x_2-x_3}{x_2-x_4}.$$

If x_2, x_3 or $x_4 = \infty$, let $f(x)$ be

$$\frac{x-x_3}{x-x_4}, \quad \frac{x_2-x_4}{x-x_4}, \quad \frac{x-x_3}{x_2-x_3}$$

respectively. Then f has the property. Let g be another linear fractional transformation with the same property. Then $f \circ g^{-1}$ is also a linear fractional transformation that leaves $1, 0, \infty$ invariant. Direct calculation shows that

$$f \circ g^{-1} = \mathrm{id}_{\mathbb{R}_\infty}.$$

Hence $f = g$ and we conclude that f is uniquely determined.
When $x_1, x_2, x_3, x_4 \in \mathbb{R}$,

$$(x_1, x_2; x_3, x_4) = f(x_1) = \frac{x_1 - x_3}{x_1 - x_4}\frac{x_2 - x_3}{x_2 - x_4}.$$

3. For given $x_1, x_2, x_3, x_4 \in \mathbb{R}_\infty$, define a map $g : \mathbb{R}_\infty \to \mathbb{R}_\infty$ by $g(x) = (x, x_2; x_3, x_4)$. Explicitly,

$$g(x) = \frac{x - x_3}{x - x_4}\frac{x_2 - x_4}{x_2 - x_3}$$

for $x \neq x_4$ and $g(x_4) = \infty$.
It is easy to verify that g is a linear fractional transformation. For a given linear fractional transformation f, $g \circ f^{-1}$ is also a linear fractional transformation and

$$(g \circ f^{-1})(f(x_2)) = g(x_2) = 1,$$

$$(g \circ f^{-1})(f(x_3)) = g(x_3) = 0$$

and

$$(g \circ f^{-1})(f(x_4)) = g(x_4) = \infty.$$

Hence, by the very definition of the cross ratio,

$$(g \circ f^{-1})(f(x_1)) = (f(x_1), f(x_2); f(x_3), f(x_4)).$$

On the other hand,

$$(g \circ f^{-1})(f(x_1)) = g(x_1) = (x_1, x_2; x_3, x_4).$$

Therefore,

$$(f(x_1), f(x_2); f(x_3), f(x_4)) = (x_1, x_2; x_3, x_4),$$

i.e., f preserves the cross ratio.

4. An inversion I_C of $\mathbb{R}^2_\infty$ in a circline C leaves the x-axis invariant if and only if the circline C meets with the x-axis orthogonally.
When $\psi = \bar{r}_a$ for some $a \in \mathbb{R}$, we have

$$I_C(x, 0) = (\psi(x), 0)$$

for $x \in \mathbb{R}_\infty$ if and only if C is a vertical line that meets the x-axis at $x = a$.
When $\psi = \bar{r}_{\alpha,r}$ for some $\alpha \in \mathbb{R}$ and $r > 0$, we also have

$$I_C(x, 0) = (\psi(x), 0)$$

for $x \in \mathbb{R}_\infty$ if and only if C is a circle that meets the x-axis orthogonally at $x = \alpha - r, \alpha + r$.

5.

$a \Leftrightarrow b$: It is easy to see that an inversion of $\mathbb{R}_\infty$ is a linear fractional transformation. By 1 of Exercise 3.13, a composition of linear fractional transformations is a linear fractional transformation. Hence, a composition of inversions of $\mathbb{R}_\infty$ is a linear fractional transformation.
Conversely, consider a linear fractional transformation

$$f(x) = \frac{ax + b}{cx + d}.$$

For $r \neq 0$, define a map $d_r : \mathbb{R}_\infty \to \mathbb{R}_\infty$ by $d_r(x) = rx$. When $r > 0$,

$$d_r(x) = \frac{\left(\sqrt{r}\right)^2}{\frac{1}{x}} = I_{\mathbf{0},\sqrt{r}}\left(I_{\mathbf{0},1}(x)\right).$$

So $d_r = I_{\mathbf{0},\sqrt{r}} \circ I_{\mathbf{0},1}$. When $r < 0$,

$$d_r = I_{\mathbf{0},\sqrt{-r}} \circ I_{\mathbf{0},1} \circ \bar{r}_{\mathbf{0}}.$$

Therefore, d_r is a composition of inversions of $\mathbb{R}_\infty$ in either case.
Assume that $a, c \neq 0$.

$$f(x) = \frac{ax+b}{cx+d} = -\frac{a}{c}\left(-1 - \frac{\frac{d}{c} - \frac{b}{a}}{-\frac{d}{c} - x}\right) = d_{-\frac{a}{c}}\left(\bar{r}_{-\frac{1}{2}}\left(d_{\left(\frac{d}{c}-\frac{b}{a}\right)}\left(\bar{r}_{-\frac{d}{2c}}(x)\right)\right)\right).$$

Hence,

$$f = d_{-\frac{a}{c}} \circ \bar{r}_{-\frac{1}{2}} \circ d_{\left(\frac{d}{c}-\frac{b}{a}\right)} \circ \bar{r}_{-\frac{d}{2c}}$$

is a composition of inversions. Similarly one can easily show that f is a composition of inversions for the case that $a = 0$ or $c = 0$.

$b \Leftrightarrow c$: "$b \Rightarrow c$" is shown already in 3 of Exercise 3.13.
Assume that f preserves the cross ratio. We will show that f is a linear fractional transformation. Let

$$x_2 = f(1), x_3 = f(0), x_4 = f(\infty).$$

Define a map $g : \mathbb{R}_\infty \to \mathbb{R}_\infty$ by

$$g(x) = (f(x), f(x_2); f(x_3), f(x_4)) = (x, x_2; x_3, x_4).$$

Note that g is a linear fractional transformation with

$$g(x_2) = 1, g(x_3) = 0, g(x_4) = \infty.$$

Let $h = g \circ f$, then it also preserves the cross ratio. Note that $h(1) = 1$, $h(0) = 0$, and $h(\infty) = \infty$. Hence,

$$x = (x, 1; 0, \infty) = (h(x), h(1); h(0), h(\infty)) = (h(x), 1; 0, \infty) = h(x),$$

so

$$g \circ f = h = \mathrm{id}_{\mathbb{R}_\infty}$$

and we conclude that $f = g^{-1}$ is a linear fractional transformation.

Chapter 4

4.6

(a) Choose two points $p_1, p_2 \in H' = I_M(H)$. Then, $I_M(p_i) \in H$. Note that $I_{H'}(p_i) = p_i$. Note also that

$$\begin{aligned} {}^{I_M}I_H(p_i) &= (I_M \circ I_H \circ I_M)(p_i) \\ &= I_M(I_H(I_M(p_i))) \\ &= I_M(I_M(p_i)) \\ &= p_i \end{aligned}$$

for $i = 1, 2$. Choose another point p_3 that does not belong to H'; then, $H' = H_{p_3, p_3'}$ for some point p_3'. Note that

$$H = (I_M)^{-1}(H') = I_M(H') = I_M(H_{p_3, p_3'}) = H_{I_M(p_3), I_M(p_3')}.$$

Clearly, $I_{H'}(p_3) = p_3'$. Note that

$$\begin{aligned} {}^{I_M}I_H(p_3) &= (I_M \circ I_H \circ I_M)(p_3) \\ &= I_M(I_H(I_M(p_3))) \\ &= I_M(I_M(p_3')) \\ &= p_3'. \end{aligned}$$

In summary,

$$I_{H'}(p_i) = {}^{I_M}I_H(p_i)$$

for $i = 1, 2, 3$, and p_1, p_2, and p_3 are non-collinear. According to Theorem 4.19,

$$I_{H'} = {}^{I_M}I_H.$$

(b) According to Theorem 4.20,

$$\phi = I_{H_n} \circ I_{H_{n-1}} \circ \cdots \circ I_{H_2} \circ I_{H_1}$$

for some lines $H_1, H_2, \cdots, H_n$ with $0 \leq n \leq 3$. Let

$$H'_k = (I_{H_k} \circ I_{H_{k-1}} \circ \cdots \circ I_{H_2} \circ I_{H_1})(H).$$

Note that $H'_n = \phi(H) = H'$. Then,

$$\begin{aligned}
{}^{\phi}I_H &= \phi \circ I_H \circ \phi^{-1} \\
&= I_{H_n} \circ I_{H_{n-1}} \circ \cdots \circ I_{H_2} \circ I_{H_1} \circ I_H \circ I_{H_1} \circ I_{H_2} \circ \cdots \circ I_{H_n} \\
&= I_{H_n} \circ I_{H_{n-1}} \circ \cdots \circ I_{H_2} \circ I_{H'_1} \circ I_{H_2} \circ \cdots \circ I_{H_n} \\
&= I_{H_n} \circ I_{H_{n-1}} \circ \cdots \circ I_{H_3} \circ I_{H'_2} \circ I_{H_3} \circ \cdots \circ I_{H_n} \\
&\vdots \\
&= I_{H'_n} \\
&= I_{H'}.
\end{aligned}$$

4.12

1. Let T be an ideal $\mathbb{H}^2$-triangle with vertices p_1, p_2, and p_3. We need to find an isometry ϕ of $\mathbb{H}^2$ such that

$$\phi(p_1) = \mathbf{0}, \phi(p_2) = (1, 0), \phi(p_3) = \infty.$$

Case 1. If all the p_i's are on the x-axis. We can assume that $\pi_x(p_2) < \pi_x(p_1)$ by re-labeling if necessary. Let

$$\phi = d_r \circ t_{(-k,0)} \circ I_C,$$

where C is a circle, centered at p_3, $k = \pi_x(I_C(p_1))$ and

$$r = \frac{1}{\pi_x(t_{(-k,0)}(I_C(p_2)))}.$$

Now it is easy to verify

$$\phi(p_1) = \mathbf{0}, \phi(p_2) = (1, 0), \phi(p_3) = \infty.$$

Case 2. If any of p_i's is ∞. Let $p_3 = \infty$. Then p_1 and p_2 lie on the x-axis and we can assume that $\pi_x(p_1) < \pi_x(p_2)$.

Let

$$\phi = d_s \circ t_{(-l,0)},$$

where $l = \pi_x(p_1)$ and $s = \frac{1}{t_{(-l,0)}(p_2)}$. Again it is easy to verify

$$\phi(p_1) = \mathbf{0}, \phi(p_2) = (1, 0), \phi(p_3) = \infty.$$

2. Let T be the ideal $\mathbb{H}^2$-triangle with vertices $\mathbf{0}$, $(1, 0)$, ∞. Using the integral in the proof of Lemma 4.23, we can show that

$$\text{Area}^2_{\mathbb{H}}(T) = \pi.$$

Since all the ideal $\mathbb{H}^2$-triangles are congruent with T, as shown above, and an isometry of $\mathbb{H}^2$ preserves the $\mathbb{H}^2$-area, the $\mathbb{H}^2$-area of an ideal $\mathbb{H}^2$-triangle is π.

4.18

Let C be a circle on S of radius ρ, centered at p, and let $\Gamma(\rho)$ be the circumference (measured in S) of C. Note that

$$C = \{q \in S \mid d(p, q) = \rho\}.$$

Hence,

$$\begin{aligned}
\phi(C) &= \{\phi(q) \mid q \in S, d(p, q) = \rho\} \\
&= \{q' \in S' \mid \phi^{-1}(q') \in S, d(p, \phi^{-1}(q')) = \rho\} \\
&= \{q' \in S' \mid d(p, \phi^{-1}(q')) = \rho\} \\
&= \{q' \in S' \mid d'(\phi(p), q') = \rho\},
\end{aligned}$$

which is a circle on S' of radius ρ, centered at $\phi(p)$. Since ϕ is an isometry, the circumference $\Gamma'(\rho)$ (measured in S') of $\phi(C)$ is the circumference of C (measured in S), which is $\Gamma(\rho)$.

Therefore, the Gaussian curvature of S' at the points $\phi(p)$ is

$$K' = \lim_{\rho \to 0^+} 3 \cdot \frac{2\pi\rho - \Gamma'(\rho)}{\pi\rho^3} = \lim_{\rho \to 0^+} 3 \cdot \frac{2\pi\rho - \Gamma(\rho)}{\pi\rho^3} = K,$$

which is the Gaussian curvature of S at point p.

4.20 Note that the image of a $\mathbb{D}^2$-circle on $\mathbb{J}^2$ is a Euclidean circle. Hence, a $\mathbb{K}^2$-circle, which is a projection of the circle on $\mathbb{J}^2$ to the xy-plane, is, in general, a Euclidean ellipse. A $\mathbb{K}^2$-circle is a Euclidean circle if and only if the circle on $\mathbb{J}^2$ has the north pole as it spherical center.

4.21 Let $\theta = \frac{2\pi}{t}$, then

$$\theta < \frac{(s-2)\pi}{s}.$$

Hence, as in Proposition 4.37, we can build a regular s-gon on the hyperbolic plane whose interior angle is θ. Perform hyperbolic reflections in all the edges of the s-gon. Then we have new s regular s-gons. Note that they overlap only in their edges. Perform hyperbolic reflections in all the edges of all the s-gons again. Then we have more regular s-gons and they overlap only in their edges. Repeat this process to get a regular hyperbolic tessellation. This tessellation is of type $[s, t]$.

Chapter 5

5.2

1.

$$\begin{aligned} e \in L_{e_1,e_2} &\Leftrightarrow d^{\mathrm{II}}(e, e_1) = d^{\mathrm{II}}(e, e_2) \\ &\Leftrightarrow \|e - e_1\|^2 = \|e - e_2\|^2 \\ &\Leftrightarrow \|e\|^2 + \|e_1\|^2 - 2e \cdot e_1 = \|e\|^2 + \|e_2\|^2 - 2e \cdot e_2 \\ &\Leftrightarrow \|e_1\|^2 - 2e \cdot e_1 - \|e_2\|^2 + 2e \cdot e_2 = 0 \\ &\Leftrightarrow (e_1 + e_2) \cdot (e_1 - e_2) - 2e \cdot (e_1 - e_2) = 0 \\ &\Leftrightarrow 2u \cdot 2v - 2e \cdot 2v = 0 \\ &\Leftrightarrow (e - u) \cdot v = 0. \end{aligned}$$

2. Since $u' \in L_{e_1,e_2}$,

$$(u' - u) \cdot v = 0.$$

Let $L' = \{e \in \mathbb{R}^{1,1} \mid (e - u') \cdot v' = 0\}$.

$$e \in L_{e_1,e_2} \Leftrightarrow (e - u) \cdot v = 0$$

$$\begin{aligned}
&\Leftrightarrow (e-u)\cdot v-(u'-u)\cdot v=0\\
&\Leftrightarrow (e-u')\cdot v=0\\
&\Leftrightarrow a(e-u')\cdot v=0\\
&\Leftrightarrow (e-u')\cdot v'=0\\
&\Leftrightarrow e\in L'.
\end{aligned}$$

Hence, $L_{e_1,e_2}=L'$.

5.5 A direct calculation yields

$$\begin{aligned}
(b_{\lambda_2}\circ b_{\lambda_1})(x,\tau) =&(\cosh(\lambda_2)(x\cosh(\lambda_1)+\tau\sinh(\lambda_1))+\\
&(\tau\cosh(\lambda_1)+x\sinh(\lambda_1))\sinh(\lambda_2),\\
&\cosh(\lambda_2)(\tau\cosh(\lambda_1)+x\sinh(\lambda_1))+\\
&(x\cosh(\lambda_1)+\tau\sinh(\lambda_1))\sinh(\lambda_2))\\
=&(x\cosh(\lambda_1)\cosh(\lambda_2)+\tau\cosh(\lambda_2)\sinh(\lambda_1)+\\
&\tau\cosh(\lambda_1)\sinh(\lambda_2)+x\sinh(\lambda_1)\sinh(\lambda_2),\ \tau\cosh(\lambda_1)\cosh(\lambda_2)\\
&+x\cosh(\lambda_2)\sinh(\lambda_1)+x\cosh(\lambda_1)\sinh(\lambda_2)+\\
&\tau\sinh(\lambda_1)\sinh(\lambda_2))\\
=&(x\cosh(\lambda_1+\lambda_2)+\tau\sinh(\lambda_1+\lambda_2),\ \tau\cosh(\lambda_1+\lambda_2)+\\
&x\sinh(\lambda_1+\lambda_2))\\
=&b_{\lambda_1+\lambda_2}(x,\tau).
\end{aligned}$$

Therefore, $b_{\lambda_2}\circ b_{\lambda_1}=b_{\lambda_1+\lambda_2}$.

5.8 Let $e_1,e_2\in L+\alpha$. Then, $e_i-\alpha\in L$. Note that $\bar{r}_{L+\alpha}(e_i)=e_i$. Note also that

$$\begin{aligned}
(t_\alpha\circ\bar{r}_L\circ t_{-\alpha})(e_i) &= (t_\alpha\circ\bar{r}_L)(e_i-\alpha)\\
&= t_\alpha(e_i-\alpha)\\
&= e_i
\end{aligned}$$

for $i=1,2$. Choose an event e_3 that does not lie on the line $L+\alpha$; then, one can choose another event e_3' such that

$$L+\alpha=L_{e_3,e_3'}.$$

Note that

$$L = L_{e_3-\alpha, e_3'-\alpha}.$$

Clearly, $\bar{r}_{L+\alpha}(e_3) = e_3'$. Note also that

$$\begin{aligned}(t_\alpha \circ \bar{r}_L \circ t_{-\alpha})(e_3) &= (t_\alpha \circ \bar{r}_L)(e_3 - \alpha) \\ &= t_\alpha(e_3' - \alpha) \\ &= e_3'.\end{aligned}$$

In summary,

$$\bar{r}_{L+\alpha}(e_i) = (t_\alpha \circ \bar{r}_L \circ t_{-\alpha})(e_i)$$

for $i = 1, 2, 3$, and e_1, e_2, and e_3 are non-collinear. According to Lemma 5.2,

$$\bar{r}_{L+\alpha} = t_\alpha \circ \bar{r}_L \circ t_{-\alpha}.$$

5.10 Let $L = L_{e_1,e_2}$ be a timelike line, then $\|e_1 - e_2\|^2 > 0$.

$$\phi(L) = \phi(L_{e_1,e_2}) = L_{\phi(e_1),\phi(e_2)}.$$

Note that

$$\begin{aligned}\|\phi(e_1) - \phi(e_2)\|^2 &= d^{\mathrm{II}}(\phi(e_1), \phi(e_2)) \\ &= d^{\mathrm{II}}(e_1, e_2) \\ &= \|e_1 - e_2\|^2 \\ &> 0.\end{aligned}$$

Hence, $\phi(L)$ is a timelike line.

5.12 From

$$\|a(\cosh\lambda, \sinh\lambda)\|^2 = \|a'(\cosh\lambda', \sinh\lambda')\|^2,$$

we have $a^2 = a'^2$, so $a = \pm a'$. Since $\cosh\lambda, \cosh\lambda' > 0$ and

$$a\cosh\lambda = a'\cosh\lambda',$$

we conclude that $a = a'$. From $a\sinh\lambda = a'\sinh\lambda'$, we have

$$\sinh\lambda = \sinh\lambda',$$

which implies that $\lambda = \lambda'$.

5.13

$$\begin{aligned} b_\lambda(e_1) &= \check{e}_1(\cosh(\lambda_1)\cosh(\lambda) + \sinh(\lambda_1)\sinh(\lambda), \cosh(\lambda)\sinh(\lambda_1) \\ &\quad + \cosh(\lambda_1)\sinh(\lambda)) \\ &= \check{e}_1(\cosh(\lambda_1+\lambda), \sinh(\lambda_1+\lambda)), \end{aligned}$$

$$\begin{aligned} b_\lambda(e_2) &= \check{e}_2(\cosh(\lambda)\sinh(\lambda_2) + \cosh(\lambda_2)\sinh(\lambda), \cosh(\lambda_2)\cosh(\lambda) \\ &\quad + \sinh(\lambda_2)\sinh(\lambda)) \\ &= \check{e}_2(\sinh(\lambda_2+\lambda), \cosh(\lambda_2+\lambda)). \end{aligned}$$

5.15 Let $e_0 = (1, 0)$ and

$$\lambda(e) = \angle e_0 \mathbf{0} e$$

for a non-zero event e. Then,

$$\lambda(e) = \lambda(-e) = \lambda(\tilde{e}) = \lambda(-\tilde{e}).$$

Hence, we can assume that e_1, e_2, and e_3 are spacelike with positive x-coordinates. Now the claim in the exercise comes directly from (5.1).

5.19 $\phi = b_{\alpha,\lambda}$ or $\bar{b}_{\alpha,\lambda}$. If $\phi = b_{\alpha,\lambda}$, then

$$\mathrm{id}_{\mathbb{R}^{1,1}} = b^n_{\alpha,\lambda} = b_{\alpha,n\lambda}.$$

Hence, we have

$$t_\alpha \circ b_{n\lambda} \circ t_{-\alpha} = \mathrm{id}_{\mathbb{R}^{1,1}},$$

which implies

$$b_{n\lambda} = \circ t_{-\alpha} \circ \mathrm{id}_{\mathbb{R}^{1,1}} \circ t_\alpha = \mathrm{id}_{\mathbb{R}^{1,1}}.$$

Hence $n\lambda = 0$ and we have $\lambda = 0$, which means that $\phi = b_0 = \mathrm{id}_{\mathbb{R}^{1,1}}$.

If $\phi = \bar{b}_{\alpha,\lambda}$, then

$$\mathrm{id}_{\mathbb{R}^{1,1}} = \bar{b}^n_{\alpha,\lambda} = \begin{cases} b_{\alpha,n\lambda}, & \text{if } n \text{ is even;} \\ \bar{b}_{\alpha,n\lambda}, & \text{if } n \text{ is odd.} \end{cases}$$

When n is even, we have

$$b_{\alpha,n\lambda} = t_\alpha \circ b_{n\lambda} \circ t_{-\alpha} = \mathrm{id}_{\mathbb{R}^{1,1}},$$

which implies

$$b_{n\lambda} = \circ t_{-\alpha} \circ \mathrm{id}_{\mathbb{R}^{1,1}} \circ t_\alpha = \mathrm{id}_{\mathbb{R}^{1,1}}.$$

Hence $n\lambda = 0$ and we have $\lambda = 0$. So $\phi = \bar{b}_{\alpha,0}$.

When n is odd,

$$\phi^n = \bar{b}_{\alpha,n\lambda} \neq \mathrm{id}_{\mathbb{R}^{1,1}}.$$

Hence, this case is excluded.

In summary, we have the claims in the exercise.

5.20 First, suppose that

$$t_\gamma \circ T_C = \mathrm{id}_{\mathbb{R}^{1,1}}$$

for some event γ and 2×2 matrix C. Note that

$$\gamma = t_\gamma(\mathbf{0}) = (t_\gamma \circ T_C)(\mathbf{0}) = \mathrm{id}_{\mathbb{R}^{1,1}}(\mathbf{0}) = \mathbf{0}.$$

Let

$$C = \begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

Since

$$(a, c) = T_C(1, 0) = \mathrm{id}_{\mathbb{R}^{1,1}}(1, 0) = (1, 0)$$

and

$$(b, d) = T_C(0, 1) = \mathrm{id}_{\mathbb{R}^{1,1}}(0, 1) = (0, 1),$$

$$C = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = I_2.$$

From the condition given in the exercise,

$$t_\beta^{-1} \circ t_\alpha \circ T_A \circ T_B^{-1} = \mathrm{id}_{\mathbb{R}^{1,1}},$$

i.e.,

$$t_{\alpha-\beta} \circ T_{AB^{-1}} = \mathrm{id}_{\mathbb{R}^{1,1}}.$$

As shown above, $\alpha - \beta = \mathbf{0}$, and $AB^{-1} = I_2$; thus, $\alpha = \beta$, $A = B$.

5.27 Assume that γ is a spacelike curve. Then, according to Exercise 5.26, δ is also a spacelike curve.

$$
\begin{aligned}
l_R(\delta) &= \int_c^d \sqrt{\left\| \frac{d\delta(t)}{dt} \right\|^2} dt \\
&= \int_c^d \sqrt{\left\| \frac{\gamma(s)}{ds} \frac{f(t)}{dt} \right\|^2}_{s=f(t)} dt \\
&= \int_c^d \sqrt{\left\| \frac{\gamma(s)}{ds} \right\|^2}_{s=f(t)} \left| \frac{f(t)}{dt} \right| dt \\
&= \int_c^d \sqrt{\left\| \frac{\gamma(s)}{ds} \right\|^2}_{s=f(t)} \frac{f(t)}{dt} dt \\
&= \int_a^b \sqrt{\left\| \frac{\gamma(s)}{ds} \right\|^2} ds \\
&= l_R(\gamma).
\end{aligned}
$$

One can similarly show that if γ is a timelike curve, δ is also timelike.

5.28 $\alpha = \gamma(a) = (\sinh a, \cosh a)$. Thus, $L = v^{\perp}$, where

$$v = (\cosh a, -\sinh a) = (\cosh(-a), \sinh(-a)).$$

Hence, $\bar{r}_L = T_{\Lambda(-2a)}$. Let $e = (\sinh \lambda, \cosh \lambda)$.

$$
\begin{aligned}
(\gamma \circ \bar{r} \circ \gamma^{-1})(e) &= (\gamma \circ \bar{r})(\lambda) \\
&= \gamma(2a - \lambda) \\
&= (\sinh(2a - \lambda), \cosh(2a - \lambda)) \\
&= T_{\Lambda(-2a)}(\sinh \lambda, \cosh \lambda) \\
&= T_{\Lambda(-2a)}(e) \\
&= \bar{r}_L(e).
\end{aligned}
$$

5.30 Choose an event e from $\phi(\mathbb{U}^2) \cap \mathbb{U}^2$. Let $e' \in \phi(\mathbb{U}^2)$. Then, $e = \phi(e_1)$ and $e' = \phi(e_1')$ for some $e_1, e_1' \in \mathbb{U}^2$. Note that

$$\|e'\|^2 = \|\phi(e_1')\|^2 = \|e_1'\|^2 = -1.$$

Hence, e' or $-e'$ belongs to $\mathbb{U}^2$. Suppose that $-e' \in \mathbb{U}^2$. Then, according to Corollary 5.36,

$$\begin{aligned} 1 &\leq \cosh d_{\mathbb{U}^2}(e, -e') \\ &= -e \cdot (-e') \\ &= e \cdot e' \\ &= \phi(e_1) \cdot \phi(e_1') \\ &= e_1 \cdot e_1' \\ &= -(-e_1 \cdot e_1') \\ &= -\cosh d_{\mathbb{U}^2}(e_1, e_1') \\ &\leq -1, \end{aligned}$$

which is a contradiction. Therefore, $e' \in \mathbb{U}^2$, and we conclude that

$$\phi(\mathbb{U}^2) \subset \mathbb{U}^2.$$

On the other hand, note that

$$e \in \phi(\mathbb{U}^2) \cap \mathbb{U}^2 \Rightarrow e_1 = \phi^{-1}(e) \in \mathbb{U}^2 \cap \phi^{-1}(\mathbb{U}^2),$$

which implies that

$$\phi^{-1}(\mathbb{U}^2) \cap \mathbb{U}^2 \neq \emptyset.$$

Applying the same argument to the isometry ϕ^{-1},

$$\phi^{-1}(\mathbb{U}^2) \subset \mathbb{U}^2,$$

which implies that $\mathbb{U}^2 \subset \phi(\mathbb{U}^2)$.

In summary, $\phi(\mathbb{U}^2) = \mathbb{U}^2$.

Chapter 6

6.5 If $A = -\Lambda(\lambda)$, then

$$\begin{aligned} x' &= x \cosh \lambda - \tau \sinh \lambda, \\ \tau' &= x \sinh \lambda - \tau \cosh \lambda. \end{aligned}$$

Hence,

$$\begin{aligned}\tau_1' - \tau_2' &= x_1 \sinh\lambda - \tau_1 \cosh\lambda - x_2 \sinh\lambda + \tau_2 \cosh\lambda \\ &= (x_1 - x_2)\sinh\lambda - (\tau_1 - \tau_2)\cosh\lambda \\ &\geq -|x_1 - x_2||\sinh\lambda| - (\tau_1 - \tau_2)\cosh\lambda) \\ &\geq -(\tau_2 - \tau_1)|\sinh\lambda| - (\tau_1 - \tau_2)\cosh\lambda \\ &= (\tau_2 - \tau_1)\,(-|\sinh\lambda| + \cosh\lambda) \\ &\geq 0\end{aligned}$$

and so $\tau_1' \geq \tau_2'$.

If $A = R(\lambda)$, then

$$\begin{aligned}x' &= x\cosh\lambda + \tau\sinh\lambda, \\ \tau' &= x\sinh\lambda + \tau\cosh\lambda.\end{aligned}$$

Hence,

$$\begin{aligned}\tau_1' - \tau_2' &= x_1 \sinh\lambda + \tau_1 \cosh\lambda - x_2 \sinh\lambda - \tau_2 \cosh\lambda \\ &= (x_1 - x_2)\sinh\lambda + (\tau_1 - \tau_2)\cosh\lambda \\ &\leq |x_1 - x_2||\sinh\lambda| + (\tau_1 - \tau_2)\cosh\lambda \\ &\leq (\tau_2 - \tau_1)|\sinh\lambda| + (\tau_1 - \tau_2)\cosh\lambda \\ &= (\tau_2 - \tau_1)\,(|\sinh\lambda| - \cosh\lambda) \\ &\leq 0.\end{aligned}$$

and so $\tau_1' \leq \tau_2'$.

If $A = -R(\lambda)$, then

$$\begin{aligned}x' &= -x\cosh\lambda - \tau\sinh\lambda, \\ \tau' &= -x\sinh\lambda - \tau\cosh\lambda.\end{aligned}$$

Hence,

$$\begin{aligned}\tau_1' - \tau_2' &= -x_1 \sinh\lambda - \tau_1 \cosh\lambda + x_2 \sinh\lambda + \tau_2 \cosh\lambda \\ &= (x_2 - x_1)\sinh\lambda + (\tau_2 - \tau_1)\cosh\lambda \\ &\geq -|x_1 - x_2||\sinh\lambda| + (\tau_2 - \tau_1)\cosh\lambda \\ &\geq -(\tau_2 - \tau_1)|\sinh\lambda| + (\tau_2 - \tau_1)\cosh\lambda \\ &= (\tau_2 - \tau_1)\,(-|\sinh\lambda| + \cosh\lambda) \\ &\geq 0.\end{aligned}$$

and so $\tau_1' \geq \tau_2'$.

6.6 For events $e_1, e_2 \in \mathbb{R}^{3,1}$, $e_i = (x_i, y_i, z_i, \tau_i)$, let $\alpha_i = (x_i, y_i, z_i) \in \mathbb{R}^3$. Then $e_i = (\alpha_i, \tau_i)$ and

$$\|e_1 - e_2\|^2 = \|\alpha_1 - \alpha_2\|^2 - (\tau_1 - \tau_2)^2.$$

Hence,

$$\begin{aligned} d^{\mathrm{II}}\left(\widetilde{\phi}(e_1), \widetilde{\phi}(e_2)\right) &= \left\|\widetilde{\phi}(e_1) - \widetilde{\phi}(e_2)\right\|^2 \\ &= \|(\phi(\alpha_1), \tau_1) - (\phi(\alpha_2), \tau_2)\|^2 \\ &= \|(\phi(\alpha_1) - \phi(\alpha_2)\|^2 - (\tau_1 - \tau_2)^2 \\ &= d(\phi(\alpha_1), \phi(\alpha_2)) - (\tau_1 - \tau_2)^2 \\ &= d(\alpha_1, \alpha_2) - (\tau_1 - \tau_2)^2 \\ &= \|\alpha_1 - \alpha_2\|^2 - (\tau_1 - \tau_2)^2 \\ &= \|e_1 - e_2\|^2 \\ &= d^{\mathrm{II}}(e_1 - e_2). \end{aligned}$$

Hence, $\widetilde{\phi}$ is an isometry of $\mathbb{R}^{3,1}$.

6.7 For an event $e = (x, y, z, \tau) \in \mathbb{R}^{3,1}$, let

$$\widehat{e} = (x, \tau) \in \mathbb{R}^{1,1}.$$

Note that the event $\widehat{e}_2 = (a, b)$ is spacelike. Hence,

$$\widehat{e}_2 = \widehat{e}_2^{\,\times}(\cosh t, \sinh t)$$

for some $t \in \mathbb{R}$ with $t > 0, \widehat{e}_2^{\,\times} > 0$. Let $\lambda = -2t$. Then,

$$B_\lambda(e_2) = (\widehat{e}_2^{\,\times} \cosh(-t), 0, 0, \widehat{e}_2^{\,\times} \sinh(-t)).$$

Now, clearly, $B_\lambda(e_2) - B_\lambda(e_1)$ is past-directed.

6.8 According to Theorem 6.7, a causal isometry of $\mathbb{R}^{1,1}$ has the form of

$$t_\alpha \circ T_A,$$

where $A = R(\lambda)$ or $\Lambda(\lambda)$ for some $\lambda \in \mathbb{R}$. Note that this isometry is orientation-preserving only if $A = R(\lambda)$ and then $T_A = b_\lambda$. So we are done.

6.9

1. First, note that

$$\gamma(t_0) = \frac{1}{a}(\cosh t_0 - 1, 0, 0, \sinh t_0) = \left(\frac{d}{2}, 0, 0, T'\right).$$

Hence,

$$t_0 = \cosh^{-1}\left(\frac{ad}{2} + 1\right)$$

and so

$$T' = \frac{1}{a}\sinh t_0 = \frac{1}{a}\sinh\left(\cosh^{-1}\left(\frac{ad}{2} + 1\right)\right).$$

The proper time elapse is

$$\begin{aligned} T &= \int_{t_0}^{0} \sqrt{-\left\|\frac{d\gamma}{dt}\right\|^2}\, dt \\ &= \int_{t_0}^{0} \sqrt{\frac{1}{a^2}(-\sinh^2 t + \cosh^2 t)}\, dt \\ &= \int_{t_0}^{0} \frac{1}{a} dt = \frac{t_0}{a} = \frac{1}{a}\cosh^{-1}\left(\frac{ad}{2} + 1\right). \end{aligned}$$

2. From the previous calculation,

$$t_0 = \cosh^{-1}\left(\frac{gd}{2} + 1\right), \quad T = \frac{1}{g}\cosh^{-1}\left(\frac{gd}{2} + 1\right).$$

$$T = \frac{46 - 10}{4} = 9.$$

The distance d between the Earth and the star is

$$d = \frac{2}{g}(\cosh(gT) - 1) \approx 10304 \text{ light-years}.$$

The number of years that pass on Earth is

$$4T' = \frac{4}{g}\sinh\left(\cosh^{-1}\left(\frac{gd}{2} + 1\right)\right) \approx 20611 \text{ years}.$$

6.11 Let $\gamma(t)$ and $\gamma'(t)$ be parametrizations of the worldline. Then,

$$\gamma'(t) = \gamma(s(t))$$

for some one variable function $s(t)$ with $\frac{ds}{dt} \neq 0$. Let

$$\gamma(t) = (r(t), \tau(t)) \text{ and } \gamma'(t) = (r'(t), \tau'(t))$$

with $r(t), r'(t) \in \mathbb{R}^3$. The velocity with respect to the parametrization $\gamma(t)$ is

$$\frac{1}{\frac{d\tau(t)}{dt}} \frac{dr(t)}{dt}.$$

The velocity with respect to the parametrization $\gamma'(t)$ is

$$\begin{aligned} \frac{1}{\frac{d\tau'(t)}{dt}} \frac{dr'(t)}{dt} &= \frac{1}{\frac{d\tau(s(t))}{dt}} \frac{dr(s(t))}{dt} \\ &= \frac{1}{\frac{d\tau(s)}{ds}\frac{ds}{dt}} \frac{d\gamma(s)}{ds} \frac{ds}{dt} \\ &= \frac{1}{\frac{d\tau(s)}{ds}} \frac{d\gamma(s)}{ds}, \end{aligned}$$

which is the velocity with respect to the parametrization $\gamma(t)$. Hence, the velocity and the speed are independent of the parametrization.

Bibliography

1. Aarts, J.M.: Plane and Solid Geometry. Translated from the 2000 Dutch original by Reinie Erné. Universitext. Springer, New York (2008) [Chapter 1]
2. Anderson, J.: Hyperbolic Geometry. Springer Undergraduate Mathematics Series, 2nd edn. Springer-Verlag London, Ltd., London (2005) [Chapters 4 and 5]
3. Audin, M.: Geometry. Translated from the 1998 French original. Universitext. Springer, Berlin (2003) [Chapters 1–3]
4. Barker, W., Howe, R.: Continuous Symmetry: From Euclid to Klein. American Mathematical Society, Providence, RI (2007) [Chapter 1]
5. Beardon, A.F.: Algebra and Geometry. Cambridge University Press, Cambridge (2005) [Chapters 1, 2 and 4]
6. Benedetti, R., Petronio C.: Lectures on Hyperbolic Geometry. Universitext. Springer, Berlin (1992) [Chapters 4 and 5]
7. Benz, W.: Classical Geometries in Modern Contexts: Geometry of Real Inner Product Spaces, 3rd edn. Birkhäuser/Springer Basel AG, Basel (2012) [Chapters 5 and 6]
8. Birman, G.S., Nomizu, K.: Trigonometry in Lorentzian geometry. Am. Math. Mon. **91**(9), 543-549 (1984) [Chapter 5]
9. Brannan, D.A., Esplen, M.F., Gray, J.J.: Geometry. Cambridge University Press, Cambridge (1999) [Chapters 2 and 4]
10. Callahan, J.: The Geometry of Spacetime: An Introduction to Special and General Relativity. Undergraduate Texts in Mathematics. Springer, New York (2000) [Chapters 5 and 6]
11. Fenchel, W.: Elementary Geometry in Hyperbolic Space. With an editorial by Heinz Bauer. De Gruyter Studies in Mathematics, vol. 11. Walter de Gruyter & Co., Berlin (1989) [Chapter 4]
12. Gourgoulhon, É.: Special Relativity in General Frames: From Particles to Astrophysics. With a foreword by Thibault Damour. Translated from the 2010 French edition. Graduate Texts in Physics. Springer, Heidelberg (2013) [Chapter 6]
13. Iversen, B.: Hyperbolic Geometry. London Mathematical Society Student Texts, vol. 25. Cambridge University Press, Cambridge (1992) [Chapters 4 and 5]
14. Jennings, G.A.: Modern Geometry with Applications. Universitext. Springer, New York (1994) [Chapters 1, 2 and 6]
15. Lewis, F.P.: Questions and discussions: history of the parallel postulate. Am. Math. Mon. **27**(1), 16–23 (1920). https://www.jstor.org/stable/2973238 [Chapter 4]

The related chapters are mentioned at the end of each reference.

N.-H. Lee, *Geometry: from Isometries to Special Relativity*, Undergraduate Texts in Mathematics, https://doi.org/10.1007/978-3-030-42101-4

16. Martin, G.: Transformation Geometry: An Introduction to Symmetry. Undergraduate Texts in Mathematics. Springer, New York/Berlin (1982) [Chapter 1]
17. Naber, G.: The Geometry of Minkowski Spacetime: An Introduction to the Mathematics of the Special Theory of Relativity. Applied Mathematical Sciences, vol. 92, 2nd edn. [of MR1174969]. Springer, New York (2012) [Chapters 5 and 6]
18. Ratcliffe, J.: Foundations of Hyperbolic Manifolds. Graduate Texts in Mathematics, vol. 149, 2nd edn. Springer, New York (2006) [Chapters 3–5]
19. Rees, E.G.: Notes on Geometry. Universitext. Springer, Berlin/New York (1983) [Chapters 1 and 4]
20. Ryan, P.J.: Euclidean and Non-Euclidean Geometry. An Analytical Approach. Cambridge University Press, Cambridge (1986) [Chapters 1, 2 and 5]
21. Sobczyk, G.: New Foundations in Mathematics: The Geometric Concept of Number. Birkhäuser Springer, New York (2013) [Chapters 5 and 6]
22. Stillwell J.: Geometry of Surfaces. Corrected reprint of the 1992 original. Universitext. Springer, New York (1992) [Chapters 1, 2 and 4]
23. Ungar, A.A.: Analytic Hyperbolic Geometry: Mathematical Foundations and Applications. World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ (2005) [Chapters 4–6]
24. Yaglom, I.M.: A Simple Non-Euclidean Geometry and Its Physical Basis: An Elementary Account of Galilean Geometry and the Galilean Principle of Relativity. Heidelberg Science Library. Translated from the Russian by Abe Shenitzer. With the editorial assistance of Basil Gordon. Springer, New York/Heidelberg (1979) [Chapter 5]

Index

N.-H. Lee, *Geometry: from Isometries to Special Relativity*, Undergraduate Texts in Mathematics, https://doi.org/10.1007/978-3-030-42101-4

Symbol Index

N.-H. Lee, *Geometry: from Isometries to Special Relativity*, Undergraduate Texts in Mathematics, https://doi.org/10.1007/978-3-030-42101-4