Undergraduate Texts in Mathematics

Arseniy Sheydvasser

Linear Fractional Transformations

An Illustrated Introduction

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.

Arseniy Sheydvasser

Linear Fractional Transformations

An Illustrated Introduction

Springer

Arseniy Sheydvasser
Southborough, MA, USA

ISSN 0172-6056 ISSN 2197-5604 (electronic)
Undergraduate Texts in Mathematics
ISBN 978-3-031-25001-9 ISBN 978-3-031-25002-6 (eBook)
https://doi.org/10.1007/978-3-031-25002-6

Mathematics Subject Classification: 51N30, 53A35, 20A05

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 parents and to my sister.
This book would not exist without you.

Preface

This book started as a collaboration between myself and students of the Massachusetts Academy of Arts and Sciences: I wrote notes and they responded with questions and what they thought could be done better. One of the requests was for a preface to the book describing how best to read it. The reader might well be confused about why this is necessary: surely, one reads a book from left to right, top to bottom, starting at the beginning and finishing at the end? And, indeed, this is one possible way to read it, but it might not be the best one, particularly if this is the reader's first foray into a mathematics text that is primarily proof-driven. Such a reader (but not only such a reader) might naturally consider some of the following.

- Should you read the proofs of statements in this book?
- How should you read proofs? Should you try to memorize them?
- Should you read every single chapter of this book?
- In each chapter, which exercises should you do? Should you try to do all of them, or should you prioritize in some way?

There are no right or wrong answers to these questions. A reader who is primarily interested in dipping their toes into new mathematical waters and feeling the direction of the current—as opposed to diving in headlong in search of deeper mathematical understanding—can get an entirely valid (if somewhat superficial) experience of this book just by reading through the chapters but ignoring the proofs and exercises.

However, for a reader looking for greater conceptual understanding, I would very strongly recommend a different approach. In the first place, I think that such a reader should read through every single proof in the text. There are a number of reasons for this. In the first place, mathematics is nothing without proofs: logical deduction guided by intuitive reasoning is at the heart of what mathematics is and it has been this way since the days of the ancient Greeks. Mathematical literature aimed at grade school students and even lower level undergraduate students very commonly ignores this, but I think it is a mistake—it deprives such students of witnessing what mathematics really is. I will not belabor this point too much, as it has already been made by far more eloquent writers [10]. The second reason is more specific to this particular text: the proofs of basic results about linear fractional transformations can be both elegant and deeply enlightening—ideally, they should leave the reader not only with the feeling that, yes, these statements are true, but also give them a deep conviction as to why they are true. The careful reader may find that some basic ideas come up again and again in these proofs and hopefully this will provide some insight into how such results are discovered and how they can be reproduced. Readers without much experience in reading proofs may be well served by remembering the words of celebrated mathematician Paul Halmos [4]:

> Don't just read it; fight it! Ask your own question, look for your own examples, discover your own proofs. Is the hypothesis necessary? Is the converse true? What happens in the classical special case? What about the degenerate cases? Where does the proof use the hypothesis?

In my opinion, memorizing proofs is usually a waste of time. Human memory is a fickle thing and without substantial training, it is difficult to memorize something verbatim without fear that it will not morph into something different with the passage of time. This is perhaps not so great a concern when memorizing a poem or novel—a misremembered word is unlikely to drastically change the meaning. In mathematics, however, changing any part of a proposition is highly likely to produce something blatantly wrong or simply word salad. A reader who wants to actually learn a theorem should proceed in a different way: strive to understand the theorem in its totality. This means:

- Understand the statement of the theorem.
- Boil down the proof of the theorem down to its essential ideas.
- Connect the theorem and its proof to other theorems and concepts that you have learned.
- Convince yourself that this theorem was the right thing to have written down.

I guarantee that any theorem that has sunk into your bones in this manner is a theorem that you will never forget, and it will instead become a foundation upon which you can continue building. For an even greater understanding, I recommend performing a similar analysis for the definitions in the book: try to understand not just what they say but why these definitions were chosen.

I wrote this book with the intention of fostering mathematical growth. The exercises in this book are written accordingly and are organized into sections at the end of each chapter. The difficulty of the exercises varies greatly: some are very simple continuations of proofs written out in the main text; others ask for proofs of entirely new results, but broken down into many steps to guide the reader through the process; still others ask for entirely new proofs without any guidance. Depending on the mathematical maturity of the reader, these exercises will range from essentially trivial to deeply challenging. Being unable to do all of the exercises should not be taken as a sign of defeat but as a chance for continued growth. I would recommend going through the exercises that cover some missing pieces in the exposition of the main text—these can be easily identified by the fact that they are cited as “(See Exercise xxx)” in the text. In particular, wherever a proof in the main text is left as an exercise to the reader, that is something that should be prioritized. On the other hand, for readers looking to avoid busywork, I recommend the following litmus test to decide whether you should skip a problem: when you look at it, is it obvious to you how to solve it? Do you understand it well enough that you could explain to another person how to do it? If the answer to both questions is an honest “yes”, then I think that skipping the exercise is permissible. If you are unsure, try to find a friend and explain to them your reasoning. If all of your friends are busy and/or don’t want to hear about math, a rubber duck will usually do in a pinch.

If you are unfamiliar with writing mathematical proofs and you find yourself struggling with the exercises as a result, there are various excellent resources out there that might be of help. There is, for example, Polya’s classic book *How to Solve It* [12] which describes various mathematical strategies that exist and how one can implement them. At the time of writing, the Art of Problem Solving maintains a helpful wiki and forums for discussing problems and posting solutions; the same company has some helpful books geared toward particular subject areas such as algebra, precalculus, and others. There are many, many other books and websites out there. However, even with all of these aides, learning proof-writing is challenging, and it is important to remember that is okay.

Above all, have fun! This is a truly wonderful subject and it deserves to be enjoyed. Play with it, explore, and I wish you good hunting.

Southborough, USA Arseniy Sheydvasser
May 2022

Acknowledgements

I am very grateful to the students of the Massachusetts Academy of Arts and Sciences who participated in the collaboration that grew into this book. Specifically, I want to thank Shane Ferrante, Ishita Goluguri, Toyesh Jayaswal, Neil Kale, Darshan Krishnaswamy, and Matt Suyer for their help, as their feedback gave me a better idea of what needed to be improved. In that same vein, I want to thank their teacher, Thomas Regele, who helped organize the whole thing.

There are many other people who deserve mention. Jeffrey Lagarias and Melvyn Nathanson were both very kind to offer advice on how to publish. William Barker went above and beyond in reading portions of the book and offering his honest critiques. Andre Hernandez-Espiet did much the same. As they say, it takes a village.

Contents

1 Euclidean Isometries and Similarities

> In which we think deeply of simple things.
>
> Arnold Ross (paraphrased).

This is a book all about functions of the form

$$\varphi(z) = \frac{az+b}{cz+d},$$

where a, b, c, d are complex numbers. Such functions are usually known as either linear fractional transformations, or sometimes Möbius transformations. Our goal for this chapter is to understand intimately the simplest kind of linear fractional transformations, where $c = 0$ and $d = 1$—that is, functions of the form

$$\varphi(z) = az + b,$$

which are usually called (complex) affine maps. We will see that these transformations will describe isometries and similarities of the Euclidean plane, and we will make good use of this to prove some basic geometric theorems. Before we do that, however, we should remind ourselves of the basics of Cartesian geometry as expressed in terms of complex numbers.

1.1 The Complex Plane and Affine Maps

Usually, one describes the Cartesian plane in terms of pairs of real numbers (x, y). However, for our purposes, it is more convenient to write everything in terms of complex numbers $z = x + iy$—here $x = \Re(z)$ is the *real part* and $y = \Im(z)$ is the *imaginary part*. This has the immediate benefit of making various definitions more compact. For example, we know that the distance between two points (x_1, y_1), (x_2, y_2) in Cartesian geometry is

A. Sheydvasser, *Linear Fractional Transformations*, Undergraduate Texts in Mathematics, https://doi.org/10.1007/978-3-031-25002-6_1

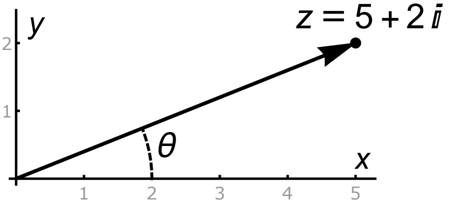


Fig. 1.1 A point z in the complex plane and the angle θ between the $+x$-axis and z.

$$d_{\text{Euclid}}((x_1, y_1), (x_2, y_2)) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}.$$

Using complex numbers, this can be phrased instead as $d_{\text{Euclid}}(z_1, z_2) = |z_1 - z_2|$, where

$$|x + iy| = \sqrt{x^2 + y^2}$$

is called the norm. The norm is particularly easy to think about because $|z|^2 = z\overline{z}$, where $\overline{x + iy} = x - iy$ is the complex conjugate. This immediately implies that the norm is *multiplicative*—that is, for all $z_1, z_2 \in \mathbb{C}$[1], $|z_1 z_2| = |z_1||z_2|$.

Complex numbers make it very convenient to describe translations. Specifically, a translation is just a transformation of the form $z \mapsto z + z_0$ for some $z_0 \in \mathbb{C}$. Why is this? Well, suppose our translation is supposed to shift everything in the x-direction by x_0, and in the y-direction by y_0. For any $z = x + iy$,

$$z + z_0 = (x + x_0) + (y + y_0)i,$$

which is exactly the desired effect.

The description we have given describes complex numbers in terms of Cartesian coordinates. Alternatively, any point z in the complex plane can be specified by its distance r away from the origin and the angle θ between the rays through 1 and z. By basic trigonometry, $z = r\cos(\theta) + i\sin(\theta)$—see Figure 1.1 for an illustration. This can be written in an equivalent way using Euler's formula that

$$\cos(\theta) + i\sin(\theta) = e^{i\theta},$$

which is not particularly hard to prove if you are familiar with Taylor series. (See Exercise 1.3.1.) Therefore, any complex number z can either be written as $x + iy$ or as $re^{i\theta}$ where $r = |z| = \sqrt{x^2 + y^2}$ is the distance from the origin, and θ is the angle between the rays through 1 and z. This makes it very easy to describe rotations: concretely, $z \mapsto e^{i\theta}z$ is a rotation by θ radians counter-clockwise around the origin. Why is this? Well, if $z = re^{i\alpha}$, then

[1] I will be making frequent use of set-theoretical notation in this book. If the reader is unfamiliar with it, I strongly recommend looking at Appendix A.

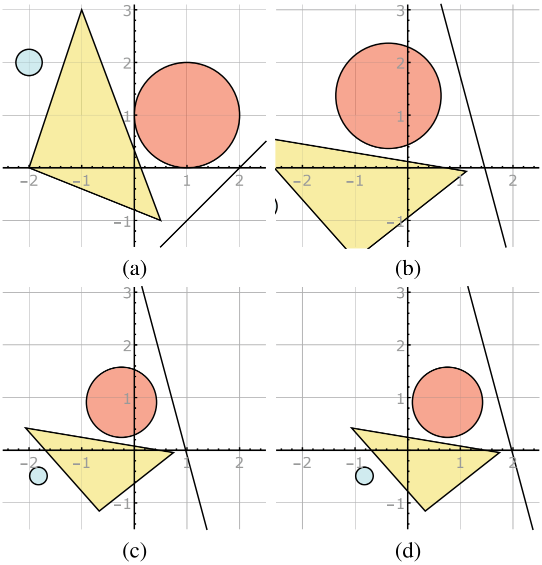


Fig. 1.2 An illustration of the proof of Theorem 1.1 for the specific case of the map $z \mapsto \frac{2}{3}e^{i\pi/3}z+1$. (a) shows an initial configuration; (b) shows the effect of $z \mapsto e^{i\pi/3}z$ on (a); (c) shows the effect of $z \mapsto 2/3z$ on (b); finally, (d) shows the effect of $z \mapsto z+1$ on (c).

$$e^{i\theta}z = re^{i\alpha}e^{i\theta} = re^{i(\alpha+\theta)},$$

which is indeed a rotation.

One final kind of transformation that is easy to describe are dilations.

Definition 1.1 A *dilation* of $\mathbb{C}$ is a transformation of the form $z \mapsto rz$ for some $r > 0$.

Intuitively, a dilation is rescaling or a "zoom" of $\mathbb{C}$. Dilations are the final ingredient we need to describe all complex affine maps.

Theorem 1.1 (Composition Theorem for Complex Affine Maps)
Let a, b be complex numbers with $a \neq 0$. Let

$$\begin{aligned} \varphi : \mathbb{C} &\to \mathbb{C} \\ z &\mapsto az + b. \end{aligned}$$

Then φ is a composition of a rotation, a dilation, and a translation.

Remark 1.1 The restriction that $a \neq 0$ is important, since otherwise φ simply maps all of $\mathbb{C}$ to a single point b.

Proof First, we write $a = re^{i\theta}$. Then, we define three maps

$$\begin{aligned}\varphi_1(z) &= e^{i\theta}z\\ \varphi_2(z) &= rz\\ \varphi_3(z) &= z+b.\end{aligned}$$

It is easy to see that $\varphi = \varphi_3 \circ \varphi_2 \circ \varphi_1$—indeed,

$$\varphi_3(\varphi_2(\varphi_1(z))) = \varphi_3(\varphi_2(e^{i\theta}z)) = \varphi_3(re^{i\theta z}) = re^{i\theta}z + b = az + b = \varphi(z).$$

This decomposition is illustrated in Figure 1.2. □

Philosophical Principle

This basic result showcases a technique that we will see over and over again: if you want to understand some kind of mathematical structure, try to break it into simple pieces that are easy to analyze, then see how you can put this information together.

▶ **Example** *Find* $a, b \in \mathbb{C}$ *such that* $\varphi(z) = az + b$ *moves* 1 *to* i *and* i *to* $1+i$*. Decompose* φ *as a rotation, dilation, and translation.*

Since $\varphi(1) = a + b = i$ and $\varphi(i) = ai + b = 1 + i$, we see that $a(i-1) = 1$, so $a = 1/(i-1) = 1/(i-1)\cdot(i+1)/(i+1) = -(i+1)/2$. From this, we get that $b = i - a = i + (i+1)/2 = (3i+1)/2$.

Next, we must write a in the form $re^{i\theta}$. We have $r = |-(i+1)/2| = |i+1|/4 = \sqrt{2}/4$. To calculate θ, we note that

$$\tan(\theta) = \frac{r\sin(\theta)}{r\cos(\theta)} = \frac{-1/2}{-1/2} = 1,$$

and since $-(i+1)/2$ is in the third quadrant of the complex plane, it follows that $\theta = \pi + \pi/4 = 5\pi/4$. Therefore, φ can be decomposed as first a rotation counterclockwise by $5\pi/4$ radians, then a dilation by $\sqrt{2}/4$, and finally a translation by $(3i+1)/2$.

1.2 Isometries

With Theorem 1.1 as our launch point, we shall now endeavor to classify all of the affine maps $z \mapsto az + b$ by sorting them into various types of transformations depending on what types of properties they preserve. We begin this quest by talking about one of the most fundamental types of transformations in geometry: isometries.

Definition 1.2 A function $\Psi : \mathbb{C} \to \mathbb{C}$ is an *isometry* if it does not change the distance between points—that is, if

$$d_{\text{Euclid}}(\Psi(z_1), \Psi(z_2)) = d_{\text{Euclid}}(z_1, z_2)$$

for all $z_1, z_2 \in \mathbb{C}$.

While we have only defined isometries of $\mathbb{C}$, the concept is much more broadly applicable. It can be defined for any metric space—roughly speaking, any set together with a distance function satisfying a few reasonable assumptions. We will delay discussing this general theory for now; we will return to it in Chapter 4.

Intuitively, we say that isometries are those functions that *preserve distance*, meaning precisely that although they may move points to other points, they do not change the distances between different points. This is an idea that will come up over and over again in this book: when you have some families of transformations, try to study what it is that they preserve.

Since $d_{\text{Euclid}}(z_1, z_2) = |z_1 - z_2|$, if Ψ is an isometry of $\mathbb{C}$, then $|\Psi(z_1) - \Psi(z_2)| = |z_1 - z_2|$ for all $z_1, z_2 \in \mathbb{C}$. Conversely, if $|\Psi(z_1) - \Psi(z_2)| = |z_1 - z_2|$ for all $z_1, z_2 \in \mathbb{C}$, then Ψ is an isometry. We will make use of this observation to make the proofs of various statements more convenient, such as the following.

Lemma 1.1 *Rotations around the origin and translations are isometries.*

Proof Let $\varphi(z) = e^{i\theta}z$ for some angle θ. We check directly that

$$\begin{aligned}|\varphi(z_1) - \varphi(z_2)| &= \left|e^{i\theta}z_1 - e^{i\theta}z_2\right| = \left|e^{i\theta}(z_1 - z_2)\right| \\ &= |e^{i\theta}||z_1 - z_2| = |z_1 - z_2|.\end{aligned}$$

Thus, rotations are isometries. The case for translations is even easier: let $\varphi(z) = z+b$ for some complex number b; then

$$|\varphi(z_1) - \varphi(z_2)| = |z_1 + b - z_1 - b| = |z_1 - z_2|,$$

directly showing that it is an isometry. □

One of the key facts about isometries is that composing them together gives you another isometry. An example of this is provided in Figure 1.3.

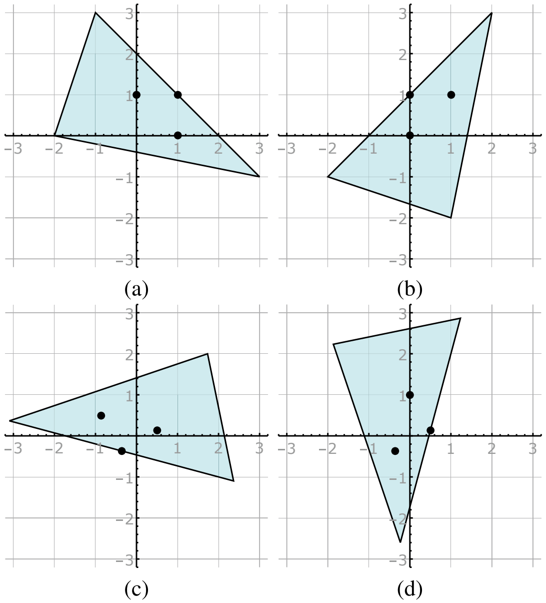
Theorem 1.2 *If* Ψ_1, Ψ_2 *are isometries, then* $\Psi_1 \circ \Psi_2$ *is an isometry.*

Remark 1.2 Although we only prove this for functions $\mathbb{C} \to \mathbb{C}$, this is true for isometries in general, and with effectively the same justification.

Proof We simply check the definition—for any $z_1, z_2 \in \mathbb{C}$,

$$\begin{aligned}d_{\text{Euclid}}\left(\Psi_1(\Psi_2(z_1)), \Psi_1(\Psi_2(z_2))\right) &= d_{\text{Euclid}}\left(\Psi_2(z_1), \Psi_2(z_2)\right) \\ &= d_{\text{Euclid}}\left(z_1, z_2\right),\end{aligned}$$

where we first used that Ψ_1 is an isometry and then that Ψ_2 is an isometry. □



(a) (b)

(c) (d)

Fig. 1.3 An illustration of the composition of two isometries. (a) shows an initial configuration; (b) shows the effect of an isometry Ψ_1 on (a); (c) shows the effect of an isometry Ψ_2 on (a); (d) shows the effect of $\Psi_2 \circ \Psi_1$ on (a).

An immediate corollary of this is that any combination of translations and rotations about the origin is an isometry. In particular, all affine maps of the form $z \mapsto e^{i\theta}z + b$ are necessarily isometries by Lemma 1.1. On the other hand, we shall shortly see two things: first, not all affine maps are isometries; second, not all isometries are affine maps. Let's begin with the latter assertion—we will cover the former in the next section.

Lemma 1.2 *Complex conjugation $z \mapsto \overline{z}$ is an isometry, but it is not an affine map.*

Proof Note that for any $z_1, z_2 \in \mathbb{C}$,

$$|\overline{z_1 - z_2}| = \sqrt{(z_1 - z_2)\overline{(z_1 - z_2)}} = |z_1 - z_2|,$$

so $z \mapsto \overline{z}$ is an isometry. Now, suppose that there exist $a, b \in \mathbb{C}$ such that $\overline{z} = az + b$ for all $z \in \mathbb{C}$. If we evaluate at $z = 0$, we get that $b = 0$. If we evaluate at $z = 1$, we get that $a = 1$. But it is certainly false that $z = \overline{z}$ for all complex numbers z. □

What sort of an isometry is $z \mapsto \overline{z}$? Since $\overline{x + iy} = x - iy$, we see it is just a reflection across the real line! We will see later that, in general, reflections are never affine maps. On the other hand, this map $z \mapsto \overline{z}$ is in some sense the only obstruction preventing affine maps from describing all Euclidean isometries.

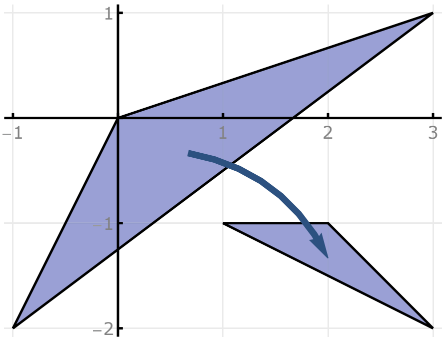


Fig. 1.4 The effect of a similarity on a triangle.

▶ **Example** *Find an isometry* Ψ *such that* $\Psi(0) = 1$, $\Psi(1) = 1 + i$, $\Psi(i) = 2$. Unfortunately, we don't know of very many isometries yet, so we will simply guess that we can find one of the form $z \mapsto az+b$ or $z \mapsto a\overline{z}+b$. In either case, the fact that $\Psi(0) = 1$ forces $b = 1$, and the fact that $\Psi(1) = a+1 = 1+i$ forces $a = i$. However, if it were the case that $\Psi(z) = iz + 1$, then $\Psi(i) = i^2 + 1 = 0 \neq 2$. So, if this approach works at all, then it must be that $\Psi(z) = i\overline{z} + 1$. Since $\Psi(i) = i\overline{i} + 1 = 2$, this is a valid solution.

1.3 Similarities

It is easy to see that dilations $\varphi(z) = rz$ are not isometries unless $r = 1$. Indeed,

$$d_{\text{Euclid}}\left(\varphi(1), \varphi(0)\right) = d_{\text{Euclid}}\left(r, 0\right) = r \neq 1 = d_{\text{Euclid}}\left(1, 0\right).$$

However, while such transformations don't preserve distances, they do preserve ratios between distances—that is, they are *similarities*.

Definition 1.3 A function $\Psi : \mathbb{C} \to \mathbb{C}$ is a *similarity* if it does not change the ratio between distances—that is, for all distinct triples of points $z_1, z_2, z_3 \in \mathbb{C}$,

$$\frac{d_{\text{Euclid}}(\Psi(z_1), \Psi(z_2))}{d_{\text{Euclid}}(\Psi(z_1), \Psi(z_3))} = \frac{d_{\text{Euclid}}(z_1, z_2)}{d_{\text{Euclid}}(z_1, z_3)}.$$

One can define similarities in the same generality as one can define isometries. Surprisingly, similarities that are not isometries are comparatively rare—for example, an exercise in one of the later chapters shows that a function on hyperbolic space is a similarity if and only if it is an isometry!

It is easy to see that any isometry is a similarity. Are there similarities that are not isometries? Yes: Figure 1.4 illustrates an example. More explicitly, take any non-trivial dilation.

Lemma 1.3 *Dilations are similarities. Non-trivial dilations are not isometries.*

Proof Let $\varphi(z) = rz$ for some $r > 0$, and check

$$\frac{|\varphi(z_1) - \varphi(z_2)|}{|\varphi(z_1) - \varphi(z_3)|} = \frac{|rz_1 - rz_2|}{|rz_1 - rz_3|} = \frac{|r||z_1 - z_2|}{|r||z_1 - z_3|} = \frac{|z_1 - z_2|}{|z_1 - z_3|},$$

which works for all $z_1, z_2, z_3 \in \mathbb{C}$. Therefore, φ is a similarity. On the other hand, as we already remarked at the beginning of this section, non-trivial dilations (i.e. dilations that are not the identity map $z \mapsto z$) are not isometries. □

Can we produce more examples of similarities? Absolutely: we can build them from other similarities and isometries, as we shall now show.

Lemma 1.4 *Let Ψ be a similarity. Then there exists a constant c (called the* constant of proportionality*) such that for any two points $z_1, z_2 \in \mathbb{C}$,*

$$c = \frac{d_{Euclid}(\Psi(z_1), \Psi(z_2))}{d_{Euclid}(z_1, z_2)}.$$

Proof For any two points $z_1, z_2 \in \mathbb{C}$ define

$$\lambda_{z_1,z_2} = \frac{d_{\text{Euclid}}(\Psi(z_1), \Psi(z_2))}{d_{\text{Euclid}}(z_1, z_2)} = \frac{|\Psi(z_1) - \Psi(z_2)|}{|z_1 - z_2|}.$$

We need to prove that λ_{z_1,z_2} is the same for any choice of z_1, z_2. We do this in the following way—we first prove that $\lambda_{z_1,z_2} = \lambda_{z_1,z_3}$ for any three points z_1, z_2, z_3. To see that this is true, notice that

$$\frac{\lambda_{z_1,z_2}}{\lambda_{z_1,z_3}} = \frac{\frac{|\Psi(z_1)-\Psi(z_2)|}{|z_1-z_2|}}{\frac{|\Psi(z_1)-\Psi(z_3)|}{|z_1-z_3|}} = \frac{\frac{|\Psi(z_1)-\Psi(z_2)|}{|\Psi(z_1)-\Psi(z_3)|}}{\frac{|z_1-z_2|}{|z_1-z_3|}} = 1,$$

where we used that Ψ is a similarity. Additionally, it is easy to check that $\lambda_{z_1,z_2} = \lambda_{z_2,z_1}$. But now, this means that for any two pairs of points z_1, z_2 and z_3, z_4, we can conclude that $\lambda_{z_1,z_2} = \lambda_{z_1,z_3} = \lambda_{z_3,z_1} = \lambda_{z_3,z_4}$. Consequently, we can simply write $\lambda = \lambda_{z_1,z_2}$, secure in the knowledge that λ doesn't depend on the choice of z_1 or z_2. □

Lemma 1.5 *Let Ψ_1, Ψ_2 be similarities with constants of proportionality c_1, c_2, respectively. Then $\Psi_1 \circ \Psi_2$ is also a similarity with a constant of proportionality c_1c_2.*

Proof I leave the proof to the reader. (See Exercise 1.2.3.) □

Theorem 1.3 *Let Ψ be a similarity. There exists some constant $r > 0$ and an isometry ψ such that $\Psi = \varphi \circ \psi$ where $\varphi(z) = rz$.*

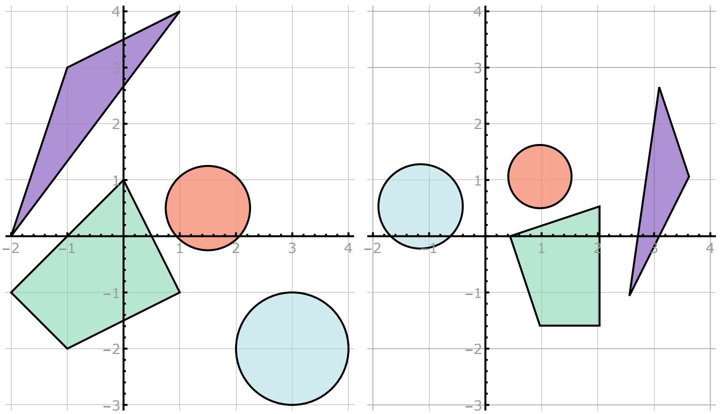


Fig. 1.5 An illustration of the effect of the map $z \mapsto \frac{3}{4}e^{3\pi i/4}\overline{z} + \frac{3}{2}$ on the complex plane.

Proof Let r be the constant of proportionality of Ψ, and define $\varphi(z) = rz$. Consider the function $\psi = \varphi^{-1} \circ \Psi$—if we prove that this is an isometry, we will be done. To show this, consider the ratio

$$\begin{aligned}\frac{|\psi(z_1) - \psi(z_2)|}{|z_1 - z_2|} &= \frac{|\varphi^{-1}(\Psi(z_1)) - \varphi^{-1}(\Psi(z_2))|}{|z_1 - z_2|} \\ &= \frac{|r^{-1}\Psi(z_1) - r^{-1}\Psi(z_2)|}{|z_1 - z_2|} \\ &= \frac{|r^{-1}||\Psi(z_1) - \Psi(z_2)|}{|z_1 - z_2|} \\ &= \frac{1}{r}\frac{|\Psi(z_1) - \Psi(z_2)|}{|z_1 - z_2|} = \frac{r}{r} = 1,\end{aligned}$$

where we used at the end the definition of the constant of proportionality. Thus, we have shown that for all z_1, z_2, $|\psi(z_1) - \psi(z_2)| = |z_1 - z_2|$, which is to say that ψ is an isometry. Ergo, $\Psi = \varphi \circ \psi$, as desired. □

This gives us an intuitive picture of what similarities are: they are just like isometries, except that they allow us to rescale everything by some constant factor. Much like isometries, similarities are crucially important in Euclidean geometry. For example, we will prove later that two triangles are similar if and only if there is a similarity that takes one to the other. We also now have enough information to be able to characterize affine maps.

Theorem 1.4 *Let a, b be complex numbers with $a \neq 0$. Then both the functions $z \mapsto az + b$ and $z \mapsto a\overline{z} + b$ are similarities.*

Remark 1.3 An example of a similarity of this form is provided in Figure 1.5.

Proof We know from Lemma 1.1 that any map $\varphi(z) = az + b$ is a composition of a rotation, a dilation, and a translation—we now know that those are all similarities, hence φ is a similarity. On the other hand, $a\overline{z} + b = \varphi(\overline{z})$, so this is just the composition of a similarity with the similarity $z \mapsto \overline{z}$; it must also be a similarity. □

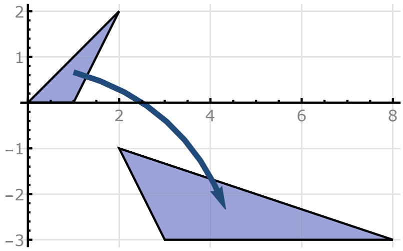
This fact serves a dual purpose: on the one hand, it gives an intuitive idea of what affine maps are. On the other hand, it gives an algebraic description of (some) similarities. Both of these are useful, and allow us to jump between algebra and geometry as necessary.

▶ **Example** *Compute the constant of proportionality of the transformation* $\varphi(z) = (1 + 2i)z + 3 - i$.
Since we can use any two points to compute the constant of proportionality, it behooves us to choose the two simplest points: 0 and 1. Then we note that if c is the constant of proportionality, then

$$c = \frac{|\varphi(1) - \varphi(0)|}{|1 - 0|} = |(1 + 2i) \cdot 1 + 3 - i - ((1 + 2i) \cdot 0 + 3 - i)|$$
$$= |1 + 2i| = \sqrt{5}.$$

(See also Exercise 1.2.4.)

▶ **Example** *Find a similarity that takes a triangle with vertices at* $0, 1, 2 + 2i$ *to a triangle with vertices at* $2 - i, 3 - 3i, 8 - 3i$.
It is a good idea to graph these two triangles to confirm that they are actually similar and to figure out which vertices should correspond to which.



We see that our similarity should send $0 \mapsto 2-i$, $1 \mapsto 3-3i$, and $2+2i \mapsto 8-3i$. We shall try to find a, b such that $\varphi(z) = az + b$ has the desired effect. (We do not know at this time that this is guaranteed to work, but we don't know of any other way to create similarities.) Since $0 \mapsto 2 - i$, we must have $\varphi(0) = b = 2 - i$. Since $1 \mapsto 3 - 3i$, we must have

$$\varphi(1) = a \cdot 1 + 2 - i = a + 2 - i = 3 - 3i,$$

whence $a = 3 - 3i - (2 - i) = 1 - 2i$. It remains to confirm that the last point is sent to the right place. Indeed,

$$\begin{aligned}\varphi(2 + 2i) &= (1 - 2i)(2 + 2i) + (2 - i)\\ &= 6 - 2i + 2 - i = 8 - 3i,\end{aligned}$$

as desired. Therefore, $\varphi = (1 - 2i)z + (2 - i)$ is a similarity that has the correct effect on our triangle.

1.4 Classifying Similarities

Theorem 1.4 tells us that maps $z \mapsto az + b$ and $z \mapsto a\overline{z} + b$ are similarities. Marvelously, the converse is also true: all similarities are of one of these two forms.

Theorem 1.5 (Classification of Similiarities)

Let Ψ be a similarity. Then there exist complex numbers a, b such that either $\Psi(z) = az + b$ for all $z \in \mathbb{C}$, or $\Psi(z) = a\overline{z} + b$ for all $z \in \mathbb{C}$.

Proof We will begin by considering a simple case: assume that $\Psi(0) = 0$ and $\Psi(1) = 1$. We will think about what $\Psi(z)$ can be. First, note that the constant of proportionality is

$$\frac{|\Psi(1) - \Psi(0)|}{|1 - 0|} = \frac{|1 - 0|}{|1 - 0|} = 1,$$

hence Ψ is an isometry by Theorem 1.3. So, choose any point $z \in \mathbb{C}$. We know that

$$\begin{aligned}|\Psi(z) - \Psi(0)| &= |z - 0|\\ |\Psi(z) - \Psi(1)| &= |z - 1|\,.\end{aligned}$$

How many points $w = \Psi(z)$ are there that satisfy those conditions? Well, let's simplify a little and note that the above two conditions can be written as $|w| = |z|$ and $|w - 1| = |z - 1|$. Furthermore,

$$\begin{aligned}|w - 1|^2 = (w - 1)\overline{(w - 1)} &= w\overline{w} - w - \overline{w} + 1\\ &= |w|^2 - 2\Re(w) + 1,\end{aligned}$$

and by the same logic $|z - 1|^2 = |z|^2 - 2\Re(z) + 1$. Knowing $|w| = |z|$, we get that $\Re(w) = \Re(z)$. Well, $|z|^2 = \Re(z)^2 + \Im(z)^2$ and $|w|^2 = \Re(w)^2 + \Im(w)^2$; ergo, $\Im(w) = \pm\Im(z)$. Therefore, either $w = \Re(z) + \Im(z)i = z$ or $w = \Re(z) - \Im(z)i = \overline{z}$. We would still like to know that it can't be that $\Psi(z) = z$ for some z, but $\Psi(z) = \overline{z}$

for other z. Suppose that this happens—i.e. there exist (non-real) complex numbers z_1, z_2 such that $\Psi(z_1) = z_1$ and $\Psi(z_2) = \overline{z_2}$. Then

$$|\Psi(z_1) - \Psi(z_2)| = |z_1 - \overline{z_2}| = |z_1 - z_2|.$$

Write $z_1 = x_1 + y_1 i$ and $z_2 = x_2 + y_2 i$, and expand out the above.

$$\begin{aligned} |z_1 - \overline{z_2}|^2 &= |x_1 + y_1 i - x_2 + y_2 i|^2 \\ &= (x_1 - x_2)^2 + (y_1 + y_2)^2 \\ |z_1 - z_2|^2 &= |x_1 + y_1 i - x_2 - y_2 i|^2 \\ &= (x_1 - x_2)^2 + (y_1 - y_2)^2, \end{aligned}$$

so the two can be equal only if $(y_1 + y_2)^2 = (y_1 - y_2)^2$. But

$$\begin{aligned} (y_1 + y_2)^2 &= y_1^2 + 2y_1y_2 + y_2^2 \\ (y_1 - y_2)^2 &= y_1^2 - 2y_1y_2 + y_2^2, \end{aligned}$$

so equality only holds if $y_1y_2 = -y_1y_2$, which is impossible since $y_1, y_2 \neq 0$ by assumption. Therefore, we have to conclude that either $\Psi(z) = z$ for all $z \in \mathbb{C}$, or $\Psi(z) = \overline{z}$ for all $z \in \mathbb{C}$.

However, we have only proved the case where $\Psi(0) = 0$ and $\Psi(1) = 1$. To prove the general case, we will show we can actually always reduce to this simple scenario. To wit, suppose that $\Psi(0) = w_0$ and $\Psi(1) = w_1$. Consider the transformation

$$\begin{aligned} \varphi &: \mathbb{C} \to \mathbb{C} \\ z &\mapsto \frac{1}{w_1 - w_0} z - \frac{w_0}{w_1 - w_0}. \end{aligned}$$

We know that it is a similarity by Theorem 1.4, and therefore $\psi = \varphi \circ \Psi$ is a similarity. On the other hand,

$$\begin{aligned} \psi(0) &= \varphi(\Psi(0)) = \varphi(w_0) \\ &= \frac{w_0}{w_1 - w_0} - \frac{w_0}{w_1 - w_0} = 0 \\ \psi(1) &= \varphi(\Psi(1)) = \varphi(w_1) \\ &= \frac{w_1}{w_1 - w_0} - \frac{w_0}{w_1 - w_0} = 1. \end{aligned}$$

By our preceding discussion, either $\psi(z) = z$ for all $z \in \mathbb{C}$, or $\psi(z) = \overline{z}$ for all $z \in \mathbb{C}$. Since $\Psi = \varphi^{-1} \circ \psi$, and it is easy to check that

$$\varphi^{-1}(z) = (w_1 - w_0)z + w_0$$

(see Exercise 1.2.5), we see that either $\Psi(z) = (w_1 - w_0)z + w_0$ for all z, or $\Psi(z) = (w_1 - w_0)\overline{z} + w_0$ for all z. In either case, we have proved what we wanted. □

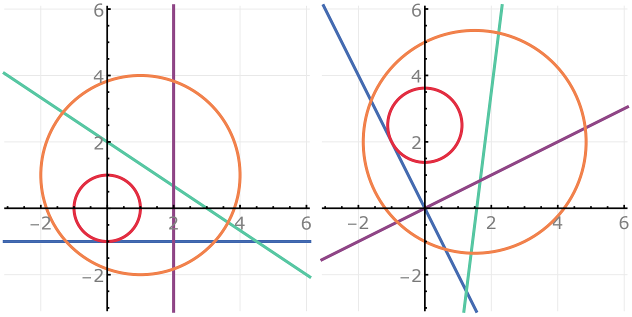


Fig. 1.6 On the left, the original image. On the right, the image under a similarity. Note that the lines are still lines, the circles are still circles, and the angle measures between lines do not change.

Theorem 1.5 is very useful; now that we understand what similarities are, it is easy to prove various properties that they have.

Theorem 1.6 *Let Ψ be a similarity of $\mathbb{C}$. Then all of the following statements are true.*

1. *Ψ is continuous.*
2. *Ψ has an inverse Ψ^{-1}, which is itself a similarity.*
3. *The image of any line under Ψ is a line.*
4. *The image of any circle under Ψ is a circle.*
5. *Given two lines l_1, l_2 that intersect at an angle θ, their image under Ψ are two lines $\Psi(l_1)$, $\Psi(l_2)$ that intersect at an angle θ.*

Remark 1.4 A visual illustration of the type of properties that are preserved under similarities is provided in Figure 1.6.

Proof By the composition theorem for complex affine maps and the classification of similarities, we know that any similarity is a composition of rotations, dilations, translations, and reflections. It is easy to see that all of those transformations are continuous, they have inverses that are similarities, they map lines to lines, they map circles to circles, and they don't change the angles between intersecting lines. Therefore, compositions of them have all those same properties. □

Remark 1.5 One can prove something stronger than mere continuity here. If you think of Ψ as being a function in two real variables, then it is (real) differentiable, and indeed smooth, meaning that its derivative is also real differentiable, and so on and so forth.

Philosophical Principle

In the proof of this theorem, we have hit on an important idea: to prove that a family of mathematical objects has a property, try to decompose those objects into

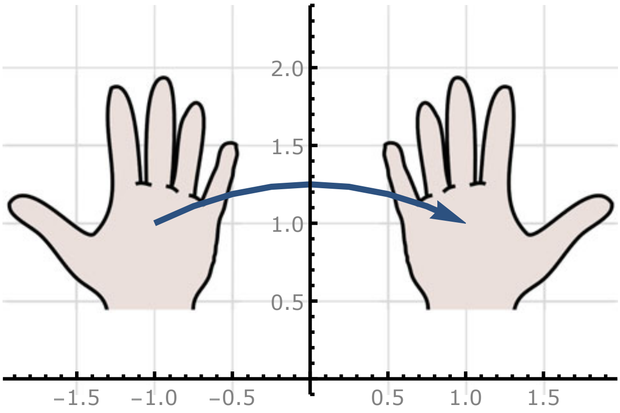


Fig. 1.7 The map $z \mapsto -\overline{z}$ reverses orientation.

simple ones. Then, show that those simple objects have that property, and try to leverage this to prove that every member of the family has the desired property.

In light of Theorem 1.6, we say that similarities are *angle preserving*. We will see later that all linear fractional transformations are angle preserving, even though they will no longer necessarily map lines to lines. This is a very useful property that we will exploit extensively, particularly in later chapters.

There is a different property that is preserved by some similarities but not others—specifically, orientation. Intuitively, we know that mirrors reverse "handedness"; the mathematical term for this property is called *orientation*. Furthermore, it is not hard to see from illustrations like Figure 1.5 and 1.7 that transformations of the form $z \mapsto a\overline{z} + b$ reverse orientation. However, formally defining orientation is difficult to do in general: it requires knowledge of either differential topology or homology. This is far more machinery than we want to introduce. Thankfully, we can do it much more simply for our specific case.

Definition 1.4 Let $\varphi : \mathbb{C} \to \mathbb{C}$ be a transformation that maps circles to circles. We say that φ is *orientation-preserving* if for any circular path C traversed counter-clockwise, the image is a circular path $\varphi(C)$ that is also traversed counter-clockwise. We say that φ is *orientation-reversing* if for any circular path C traversed counter-clockwise, the image is a circular path $\varphi(C)$ that is traversed clockwise.

Remark 1.6 Definition 1.4 can only make sense in the context of maps that preserve circles—thankfully, by Theorem 1.6, we know that similarities qualify. We will have to revisit this definition in Chapter 2 when we consider maps that do not always send circles to circles.

We can now easily show that similarities split into orientation-preserving and orientation-reversing along the expected lines.

Theorem 1.7 (Classification of Orientation-Preserving/Reversing Similarities) *Let Ψ be a similarity of $\mathbb{C}$—exactly one of the following is true.*

1. *$\Psi(z) = az + b$ for some $a, b \in \mathbb{C}$, and Ψ is orientation-preserving.*
2. *$\Psi(z) = a\bar{z} + b$ for some $a, b \in \mathbb{C}$, and Ψ is orientation-reversing.*

Proof By the classification of similarities, we know that Ψ is either of the form $z \mapsto az + b$, or $z \mapsto a\bar{z} + b$. In the first case, by the decomposition theorem for complex affine maps, Ψ is a composition of a translation, a reflection around the origin, and a dilation—it is easy to see that all three of these basics types of transformations are orientation-preserving and therefore their composition is orientation-preserving. In the second case, Ψ is also composed with the reflection $z \mapsto \bar{z}$; it is easy to see that this reflection is orientation-reversing. However, the composition of an orientation-preserving map and an orientation-reversing map is an orientation-reversing map. □

To reiterate, we can now describe the affine maps in the following beautiful way: they are precisely the orientation-preserving similarities on $\mathbb{C}$!

▶ **Example** *Let $\varphi(z) = i\bar{z}+2$. Compute its inverse and confirm that it is a similarity.*
If $\varphi(z) = i\bar{z} + 2$, then $z = i\overline{\varphi^{-1}(z)} + 2$, hence

$$\begin{aligned} i\overline{\varphi^{-1}(z)} &= z - 2 \\ \overline{\varphi^{-1}(z)} &= -iz + 2i \\ \varphi^{-1}(z) &= \overline{-iz + 2i} = i\bar{z} - 2i, \end{aligned}$$

which is indeed a similarity. (See also Exercise 1.2.5.)

1.5 Applications

We have spent a significant amount of time classifying similarities and showing how they relate to linear fractional transformations. It would be good to know that all of this effort is actually worth it. We have already seen part of the payoff via Theorem 1.6. To add to this, we assemble here a collection of various ways that our machinery can be used to give short proofs of classical results in Euclidean geometry.

Theorem 1.8 *Let Δ_1, Δ_2 be two triangles. They are similar if and only if there exists a similarity Ψ such that $\Psi(\Delta_1) = \Delta_2$.*

Proof Saying that Δ_1 and Δ_2 are similar is the same as saying that Δ_1 has vertices v_1, v_2, v_3 and Δ_2 has vertices w_1, w_2, w_3 such that

$$\frac{d_{\text{Euclid}}(v_1, v_2)}{d_{\text{Euclid}}(v_1, v_3)} = \frac{d_{\text{Euclid}}(w_1, w_2)}{d_{\text{Euclid}}(w_1, w_3)},$$
$$\frac{d_{\text{Euclid}}(v_2, v_3)}{d_{\text{Euclid}}(v_2, v_1)} = \frac{d_{\text{Euclid}}(w_2, w_3)}{d_{\text{Euclid}}(w_2, w_1)},$$
$$\frac{d_{\text{Euclid}}(v_3, v_1)}{d_{\text{Euclid}}(v_3, v_2)} = \frac{d_{\text{Euclid}}(w_3, w_1)}{d_{\text{Euclid}}(w_3, w_2)}.$$

If there exists a similarity Ψ such that $\Psi(\Delta_1) = \Delta_2$, then the above will be satisfied—all of these relations come from the defining property of a similarity! So, the only difficulty is proving that if Δ_1 and Δ_2 are similar then that had to have come from some similarity taking one to the other. To prove this, we will first consider a very basic case: suppose that $v_1 = w_1 = 0$ and $v_2 = w_2 = 1$. If this is so, then we must have

$$\frac{|v_1 - v_2|}{|v_1 - v_3|} = \frac{|w_1 - w_2|}{|w_1 - w_3|},$$
$$\frac{|v_2 - v_3|}{|v_2 - v_1|} = \frac{|w_2 - w_3|}{|w_2 - w_1|},$$

whence

$$\frac{|0 - 1|}{|0 - v_3|} = \frac{|0 - 1|}{|0 - w_3|},$$
$$\frac{|1 - v_3|}{|1 - 0|} = \frac{|1 - w_3|}{|1 - 0|},$$

and so

$$|v_3| = |w_3|$$
$$|v_3 - 1| = |w_3 - 1|.$$

We saw previously in the course of the proof of the classification of similarities that such equations have only two solutions: either $v_3 = w_3$ or $v_3 = \overline{w_3}$. Therefore, we can take either $\Psi(z) = z$ or $\Psi(z) = \overline{z}$ and have $\Psi(\Delta_1) = \Delta_2$, as desired.

How do we reduce to this basic case, though? The observation is the following: if we apply a similarity to Δ_1 or Δ_2, then the new triangles will still be similar. Furthermore, if we can find a similarity between these two new triangles, then we can compose it with the similarities used to move Δ_1 and Δ_2 into their special position to get a new similarity Ψ such that $\Psi(\Delta_1) = \Delta_2$. So, all we need to do is to find a similarity that will move $v_1 \mapsto 0$, $v_2 \mapsto 1$. This is easy: use

$$\varphi(z) = \frac{1}{v_2 - v_1}(z - v_1).$$

Obtaining an analogous similarity that sends $w_1 \mapsto 0$, $w_2 \mapsto 1$, we are done. □

Corollary 1.1 *Let Δ_1, Δ_2 be two triangles. They are congruent if and only if there exists an isometry Ψ such that $\Psi(\Delta_1) = \Delta_2$.*

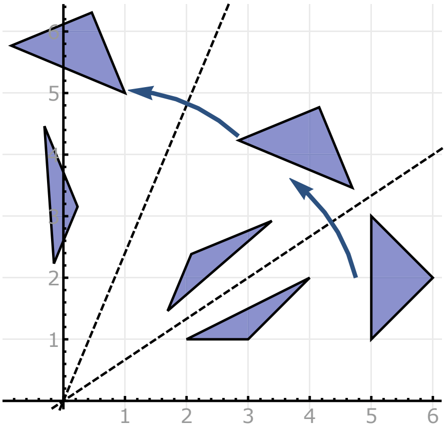


Fig. 1.8 A pair of reflections can be composed to give a rotation around the origin.

Remark 1.7 Generically, the isometry such that $\Psi(\Delta_1) = \Delta_2$ is unique—however, if Δ_1 is somehow symmetric, then there can be multiple different isometries with the same properties. Indeed, in the next section, we shall consider examples of isometries that move a polygon back onto itself—the identity map will always do this, but there may well be other examples.

Proof By Theorem 1.8, we know if $\Delta_1 \cong \Delta_2$, we can find a similarity Ψ such that $\Psi(\Delta_1) = \Psi(\Delta_2)$. However, since $\Delta_1 \cong \Delta_2$, the constant of proportionality of Ψ must be 1, so it is an isometry. □

Theorem 1.9 *Every isometry of $\mathbb{C}$ can be written as a composition of reflections.*

Proof I leave the proof to the reader. Figure 1.8 gives a hint of how to do it for rotations. (See Exercise 1.2.9.) □

Theorem 1.10 *If Ψ is an orientation-preserving isometry of $\mathbb{C}$, then either Ψ is a translation, or it is a rotation around some point.*

Proof Since Ψ is orientation-preserving, by the classification of orientation-preserving and orientation-reversing isometries, we know that $\Psi(z) = az + b$ for some complex numbers a, b. Since Ψ is an isometry, we know that $|a| = 1$. (See Exercise 1.2.4.) If $a = 1$, Ψ is a translation. Otherwise, write $a = e^{i\theta}$—I claim that Ψ is a rotation by θ around some point w. This is so if Ψ is of the form $\Psi(z) = e^{i\theta}z + \left(1 - e^{i\theta}\right) w$ for some $w \in \mathbb{C}$. (See Exercise 1.2.2.) But since $e^{i\theta} \neq 1$, we see that if we simply take

$$w = \frac{b}{1 - e^{i\theta}},$$

then we have shown that Ψ can indeed be put in the desired form. □

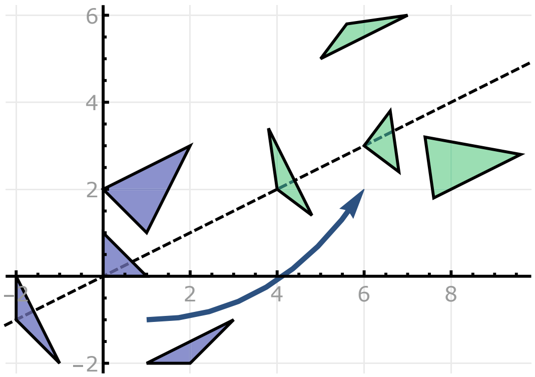


Fig. 1.9 A glide reflection.

Theorem 1.11 *If Ψ is an orientation-reversing isometry of $\mathbb{C}$, then either Ψ is a reflection across some line, or Ψ is a glide reflection (that is, a reflection across some line together with a translation along the direction of this line, such as what is illustrated in Figure 1.9).*

Proof I leave the proof to the reader. (See Exercise 1.2.10.) □

Theorem 1.12 *Any isometry of $\mathbb{C}$ fixes either no points, one point, a line, or the entire plane.*

Proof The identity map $z \mapsto z$ fixes the entire plane. Assume that our isometry is not the identity map. There aren't many other options:

1. Translations and glide reflections fix no points.
2. Rotations fix one point.
3. Reflections fix a line.

This enumerates all possibilities. □

▶ **Example** *Determine the set of points fixed by the similarity $\varphi(z) = e^{i\pi/5}\overline{z}$.*
The set of fixed points is the collection of $z \in \mathbb{C}$ satisfying $z = e^{i\pi/5}\overline{z}$. Multiplying by z on both sides, we get $z^2 = |z|^2 e^{i\pi/5}$. Write $z = re^{i\theta}$. Then this becomes $r^2e^{2i\theta} = r^2e^{i\pi/5}$. Ergo, the set of fixed points is the line of points of the form $z = re^{i\pi/10}$ for some $r \in \mathbb{R}$.

1.6 A Little Bit of Group Theory

The focus of this chapter can be summarized as trying to understand Euclidean geometry by studying transformations on it. For example, we might study the collection of isometries: this gives the usual notion of congruence that we are used to. We might study the collection of similarities: this gives the usual notion of similarity that comes up extensively in trigonometry. In short, we have had the following guiding thought.

Philosophical Principle

> Rather than studying a type of geometry directly, study the collection of transformations that preserve its basic properties.

This philosophy is very prominent in modern mathematics. In fact, we can go further and try to define a geometry by starting with a collection of transformations and seeing what sort of properties they preserve—this is more or less precisely how we will introduce inversive geometry, and later hyperbolic geometry. Before I end this chapter, I want to briefly develop this philosophical idea further and specify what types of collections of transformations we are interested in—that is, I want to finish with a short introduction to *groups*.

Definition 1.5 A *group* $(G, *, \iota)$ is a set G together with a binary operation $* : G \times G \to G$ (which we usually call the *group operation* or *group multiplication*) satisfying the following properties.

1. For all $a, b, c \in G$, $(a * b) * c = a * (b * c)$—that is, the multiplication $*$ is *associative*.
2. There exists an element $\iota \in G$ such that for all $a \in G$, $a * \iota = \iota * a = a$—that is, there is an *identity*.
3. For every element $a \in G$, there exists an element $b \in G$ such that $a * b = b * a = \iota$—that is, every element a has an *inverse*.

A group is called *abelian*[2] if additionally for all $a, b \in G$, $ab = ba$.

Remark 1.8 Some authors include a "closure" axiom that states that for all $a, b \in G$, $a * b \in G$. For us, this is packaged into the definition of a binary operation—after all, we define the co-domain of $*$ to be G.

Remark 1.9 It is customary to denote the inverse of an element b by b^{-1}. This is justified by the fact that any element has only one inverse. (See Exercise 1.4.3.)

If it is clear from context what the group operation $*$ is, we will simply write ab to mean $a * b$. We will also often refer to G itself as a group—so we might refer, for

[2] Why on Earth are groups whose multiplication is commutative called abelian? They are named in honor of Niels Henrik Abel, one of the very first group theorists.

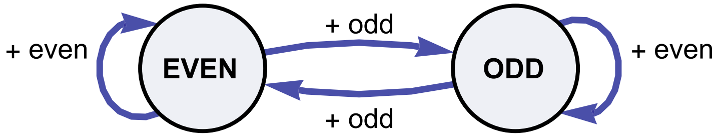


Fig. 1.10 A diagram illustrating the arithmetic of even and odd numbers.

instance, to "the group of similarities Sim($\mathbb{C}$)." This is technically abuse of notation, but it is far more convenient and I have never seen it be confusing in practice.

You might think that you haven't seen groups before, but I assure you that you have: you just haven't seen them under that name. Let me provide a few examples.

1. The set of complex numbers $\mathbb{C}$ is an abelian group if we take $*$ to be addition and $\iota = 0$. Indeed, addition of complex numbers is associative, 0 is an identity, and every complex number z has an additive inverse $-z$.
2. The set of real numbers $\mathbb{R}$, the set of rational numbers $\mathbb{Q}$, and the set of integers $\mathbb{Z}$ are all abelian groups if we take $*$ to be addition and $\iota = 0$, for the same reasons as above.
3. The set of non-zero complex numbers $\mathbb{C}^\times$ is an abelian group if we take $*$ to be multiplication and $\iota = 1$. Indeed, multiplication of complex numbers is associative, 1 is an identity, and every non-zero complex number z has a multiplicative inverse z^{-1}.
4. The set of non-zero real numbers $\mathbb{R}^\times$, the set of all positive real numbers $\mathbb{R}^+$, the set of all non-zero rational numbers $\mathbb{Q}^\times$, and the set of all positive rational numbers $\mathbb{Q}^+$ are all abelian groups if we take $*$ to be multiplication and $\iota = 1$, for the same reasons as above.
5. The set {even, odd} is an abelian group if we take $*$ to be addition with the usual rules that

$$\begin{aligned}\text{even} + \text{even} &= \text{even},\\ \text{even} + \text{odd} &= \text{odd},\\ \text{odd} + \text{even} &= \text{odd},\\ \text{odd} + \text{odd} &= \text{even},\end{aligned}$$

   and we take $\iota = $ even. Indeed, one can check that this addition is associative, "even" is an identity, and each element has an inverse (see Exercise 1.4.1). Figure 1.10 gives a visual guide to understanding this group.

All of the above are very important and worthy groups—however, there are a few examples that are more relevant to us.

Theorem 1.13 *Define Sim($\mathbb{C}$) to be the collection of similarities on $\mathbb{C}$. Then Sim($\mathbb{C}$) is a non-abelian group if we take the operation to be composition and the identity to be $\iota(z) = z$.*

Proof Let's attack this piece by piece. First, we confirm that function composition $\circ$ is a binary operation on $\mathrm{Sim}(\mathbb{C})$—that is, if we compose two similarities, we get another similarity. We know that this is true by Lemma 1.5. Second, we show that the group operation is associative. But of course it is: if $\varphi_1, \varphi_2, \varphi_3 \in \mathrm{Sim}(\mathbb{C})$, then

$$\begin{aligned}(\varphi_1 \circ (\varphi_2 \circ \varphi_3))\,(z) &= \varphi_1(\varphi_2(\varphi_3(z))) \\ &= ((\varphi_1 \circ \varphi_2) \circ \varphi_3)\,(z),\end{aligned}$$

whence $\varphi_1 \circ (\varphi_2 \circ \varphi_3) = (\varphi_1 \circ \varphi_2) \circ \varphi_3$, and so $\circ$ is associative. For all $\varphi \in \mathrm{Sim}(\mathbb{C})$,

$$\begin{aligned}(\varphi \circ \iota)(z) &= \varphi(\iota(z)) = \varphi(z) \\ &= \iota(\varphi(z)) = (\iota \circ \varphi)(z),\end{aligned}$$

whence $\varphi \circ \iota = \iota \circ \varphi = \varphi$, and so ι is the identity. We need to show that for every $\varphi \in \mathrm{Sim}(\mathbb{C})$ there exists $\psi \in \mathrm{Sim}(\mathbb{C})$ such that $\varphi \circ \psi = \psi \circ \varphi = \iota$—we know that this is true by Theorem 1.6. Finally, why is this a non-abelian group? Consider the transformations

$$\begin{aligned}\phi_1(z) &= z + 1 \\ \phi_2(z) &= iz.\end{aligned}$$

Both of these are elements in $\mathrm{Sim}(\mathbb{C})$, but

$$\begin{aligned}(\phi_1 \circ \phi_2)(z) &= iz + 1 \\ (\phi_2 \circ \phi_1)(z) &= iz + i,\end{aligned}$$

which are different. □

We saw that $\mathbb{C}$ and $\mathbb{C}^\times$ both contain smaller groups that are interesting in their own right—so does $\mathrm{Sim}(\mathbb{C})$.

Theorem 1.14 *All of the following are non-abelian groups if we take the operation to be composition and the identity to be $\iota(z) = z$.*

1. *Sim$^0(\mathbb{C})$: the collection of all orientation-preserving similarities of $\mathbb{C}$.*
2. *Isom$(\mathbb{C})$: the collection of all isometries of $\mathbb{C}$.*
3. *Isom$^0(\mathbb{C})$: the collection of all orientation-preserving isometries of $\mathbb{C}$.*

Proof I leave the proof that they are groups to the reader. (See Exercise 1.4.5.) To see that they are non-abelian, it is sufficient to note that the two transformations ϕ_1, ϕ_2 that we used to prove that $\mathrm{Sim}(\mathbb{C})$ is non-abelian are also elements of all of these groups. □

What gives? Why is every interesting collection of transformations a group? If you think about it, this makes perfect sense. First of all, our collection had better be closed under composition—at worst, if it is not, then we enlarge it until it is. Composition of functions is always associative, so we get that property for free. Whatever collection of transformations we have, the identity transformation $z \mapsto z$ that just doesn't do anything to our underlying space is always an option. The only requirement that is at all restrictive is that every transformation must have an inverse. However, in the context of geometrical transformations, this requirement will typically be satisfied because whatever transform we do, we usually can simply undo it, and that will be our desired inverse. This simple observation encapsulates why group theory is central to much of modern geometry, and allows us to refine the philosophical statement that we voiced previously.

Philosophical Principle

> To study a geometry, determine some invariants (such as length or angles) that characterize that geometry. Then, study groups of transformations of this geometry that preserve these invariants (such as isometries or similarities).[3]

Moreover, sometimes you will want to consider geometries that are refinements of each other—we know, for instance, that studying Euclidean geometry via congruence is a refinement of studying it up to similarity. In these cases, we look for a subset of the original group of transformations; of course, that subset should itself be a group via the same operation as the original. In other words, it should be a subgroup.

Definition 1.6 Let $(G, *, \iota)$ be a group. A *subgroup* H of G is a non-empty subset $H \subset G$ such that for all $g, h \in H$, $gh \in H$ and $h^{-1} \in H$.

Remark 1.10 It isn't hard to prove that H is a subgroup if and only if $(H, *, \iota)$ is a group, thus justifying the name. (See Exercise 1.4.4.)

So, for example, we have shown that $\mathrm{Sim}^0(\mathbb{C})$, $\mathrm{Isom}(\mathbb{C})$, and $\mathrm{Isom}^0(\mathbb{C})$ are all subgroups of $\mathrm{Sim}(\mathbb{C})$. These are all subgroups that have infinitely many elements—they fit into the portion of modern geometry known as Lie theory. Before we finish, I want to give an example of a subgroup of $\mathrm{Sim}(\mathbb{C})$ that is finite—such examples are also interesting, but typically appear in slightly different areas of mathematics, such as geometric group theory. To be concrete, we are going to define the isometry group of the square.

Definition 1.7 Let $\diamond$ denote the square with vertices $\pm 1, \pm i$. Define $\mathrm{Isom}(\diamond)$ to be the set of $\varphi \in \mathrm{Isom}(\mathbb{C})$ such that $\varphi(\diamond) = \diamond$.

Lemma 1.6 *$\mathit{Isom}(\diamond)$ is a subgroup of $\mathit{Isom}(\mathbb{C})$.*

Proof First, note that if $\varphi_1, \varphi_2 \in \mathrm{Isom}(\diamond)$, then $\varphi_1(\varphi_2(\diamond)) = \diamond$, hence $\varphi_1 \circ \varphi_2 \in \mathrm{Isom}(\diamond)$. Secondly, $\varphi \in \mathrm{Isom}(\diamond)$, then certainly there exists $\varphi^{-1} \in \mathrm{Isom}(\diamond)$, and since $\varphi(\diamond) = \diamond$, we see that $\varphi^{-1}(\diamond) = \diamond$, hence $\varphi^{-1} \in \mathrm{Isom}(\diamond)$. □

[3] This philosophy is known as the Erlangen program; it was originally proposed by Felix Klein in 1872 [8].

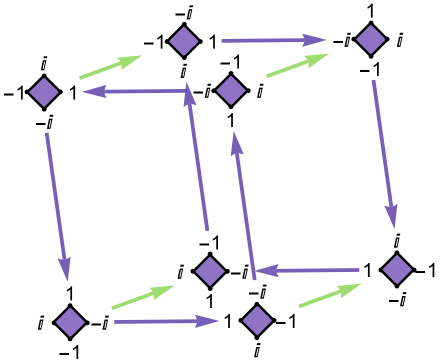


Fig. 1.11 A diagram illustrating all of the various isometries of the square—the purple arrows correspond to rotations, while the green arrows correspond to reflections.

To figure out what transformations this group consists of, we will first think about a simpler group; namely, $\mathrm{Isom}^0(\diamond)$, the collection of orientation-preserving transformations that send $\diamond \mapsto \diamond$. This set can also be understood as $\mathrm{Isom}(\diamond) \cap \mathrm{Isom}^0(\mathbb{C})$, and it is easily seen that it is a group.

Lemma 1.7 *There are only four elements in Isom*$^0(\diamond)$*:* $z \mapsto z$, $z \mapsto iz$, $z \mapsto -z$, $z \mapsto -iz$.

Proof Notice that 0 is the intersection of the diagonals of $\diamond$—therefore, its image under any isometry φ must be the intersection of the diagonals of the square $\varphi(\diamond)$. However, $\varphi(\diamond) = \diamond$ if $\diamond \in \mathrm{Isom}^0(\diamond)$, so $\varphi(0) = 0$. Since φ is orientation-preserving, it must be a rotation around the point 0. Any such rotation will be totally determined by where it sends 1. But since 1 is a vertex of the square, its image must be one of the vertices of the square. This gives precisely the four options listed. □

The rest is easy.

Theorem 1.15 *Isom*$(\diamond)$ *is a group consisting of the following eight elements.*

$$\begin{array}{cccc} z \mapsto z & z \mapsto iz & z \mapsto -z & z \mapsto -iz \\ z \mapsto \bar{z} & z \mapsto i\bar{z} & z \mapsto -\bar{z} & z \mapsto -i\bar{z} \end{array}$$

Remark 1.11 Figure 1.11 depicts the structure of this group.

Proof We already know that $\mathrm{Isom}(\diamond)$ is a group and it is easy to see that $\psi(z) = \bar{z}$ is an orientation-reversing transformation in $\mathrm{Isom}(\diamond)$. For any $\varphi \in \mathrm{Isom}(\diamond)$ that is orientation-reversing, we see that $\varphi \circ \psi$ is orientation-preserving—that is, it is inside $\mathrm{Isom}^0(\diamond)$. However, we already know everything inside $\mathrm{Isom}^0(\diamond)$! Thus, there are only the eight given choices of isometries. □

While a nice result, there is something a little artificial about it in that we have only worked out the isometry group of this particular square. However, one can check that the isometry group of any other square “looks the same” in some sense. (See also Exercise 1.4.11.) In this broader context, the group $\mathrm{Isom}(\diamond)$ is better known as the *dihedral group of order 8*, or D_8. It is often one of the first examples of a group depicted in any course on the subject due to being easy to visualize yet already demonstrating some of the complexities of the subject. (See also Exercise 1.4.12.)

Problems

1.1 COMPUTATIONAL EXERCISES

1. For each of the following, compute the image under $\varphi : \mathbb{C} \to \mathbb{C}$.

   a) Line $y = 3x - 1$, $\varphi(z) = (2 + i)z - 1$.
   b) Line $x = 4$, $\varphi(z) = i\overline{z} + i$.
   c) Circle $|z| = 2$, $\varphi(z) = (1 - 3i)z$.
   d) Circle $|z - 2| = 1$, $\varphi(z) = (1 - i)\overline{z} + 2$.

2. For each of the following, given $\varphi_1, \varphi_2 : \mathbb{C} \to \mathbb{C}$, compute $\varphi_1 \circ \varphi_2$ and $\varphi_2 \circ \varphi_1$. Are they generally the same or different?

   a) $\varphi_1(z) = \frac{1+i}{\sqrt{2}} z$, $\varphi_2(z) = iz$.
   b) $\varphi_1(z) = (3 + 4i)z$, $\varphi_2(z) = (1 - i)z$.
   c) $\varphi_1(z) = z + 3 - i$, $\varphi_2(z) = z - 1$.
   d) $\varphi_1(z) = (3 + 4i)z$, $\varphi_2(z) = z + 1$.
   e) $\varphi_1(z) = iz$, $\varphi_2(z) = z + i$.
   f) $\varphi_1(z) = \overline{z}$, $\varphi_2(z) = z + i$.
   g) $\varphi_1(z) = \overline{z}$, $\varphi_2(z) = z + 2$.
   h) $\varphi_1(z) = \overline{z}$, $\varphi_2(z) = (1 + i)z$.

3. Find $a, b \in \mathbb{C}$ such that $z \mapsto az + b$ or $z \mapsto a\overline{z} + b$ is the desired similarity.

   a) A translation that moves $2 - 3i \mapsto 1 + i$.
   b) A rotation around the origin that moves $3 + 5i \mapsto (4 + i)\sqrt{2}$.
   c) A rotation around $2 + \sqrt{3} + i$ by $\pi/6$ radians.
   d) A reflection through the line $y = 1$.
   e) A reflection through the line $y = x$.
   f) A reflection through the line $y = 3x - 1$.
   g) A glide reflection through the line $y = x$ moving $0 \mapsto 1 + i$.
   h) An orientation-preserving similarity taking $1 \mapsto 7 - 3i$ and $-2 + i \mapsto 6i$.

1.2 PROOFS

1. a) Prove de Moivre's theorem that $\cos(nx) + i\sin(nx) = (\cos(x) + i\sin(x))^n$. *(Hint: Use Euler's formula.)*
   b) Set $n = 2$ in the above, and expand the right-hand side. Use this to compute $\cos(2x)$ and $\sin(2x)$ in terms of $\sin(x)$ and $\cos(x)$.
2. Our goal for this exercise is to find an explicit formula (of the form $z \mapsto az + b$) describing a rotation of θ radians around a fixed point w.

   a) Let Ψ be the desired rotation. If we let φ be a translation taking w to 0 (so that φ^{-1} is a translation taking 0 to w), then what is the isometry $\varphi^{-1} \circ \Psi \circ \varphi$? *(Hint: it is a rotation.)*
   b) Use your answer to part a) to prove that $\Psi(z) = e^{i\theta} z + \left(1 - e^{i\theta}\right) w$.

3. a) Let Ψ_1, Ψ_2 be similarities. Show that $\Psi_1 \circ \Psi_2$ is a similarity.
   b) Why is
   $$\frac{d_{\text{Euclid}}(\Psi_1(\Psi_2(z_1)), \Psi_1(\Psi_2(z_2))}{d_{\text{Euclid}}(\Psi_2(z_1), \Psi_2(z_2))}$$
   equal to the constant of proportionality of Ψ_1?
   c) Why is
   $$\frac{d_{\text{Euclid}}(\Psi_1(\Psi_2(z_1)), \Psi_1(\Psi_2(z_2)))}{d_{\text{Euclid}}(\Psi_2(z_1), \Psi_2(z_2)} \cdot \frac{d_{\text{Euclid}}(\Psi_2(z_1), \Psi_2(z_2))}{d_{\text{Euclid}}(z_1, z_2)}$$
   equal to the constant of proportionality of $\Psi_1 \circ \Psi_2$?
   d) Why does Lemma 1.5 follow from parts a)- c)?
4. Prove that the constant of proportionality of the similarity $\varphi(z) = az + b$ is $|a|$.
5. a) Given $\varphi(z) = az + b$, compute the inverse $\varphi^{-1}(z)$.
   b) Given $\varphi(z) = a\overline{z} + b$, compute the inverse $\varphi^{-1}(z)$.
6. Let l be a line passing through two points w_1, w_2. Prove that the reflection through that line has the form
   $$\varphi(z) = \frac{w_1 - w_2}{\overline{w_1} - \overline{w_2}}\overline{z} + \frac{\overline{w_1}w_2 - w_1\overline{w_2}}{\overline{w_1} - \overline{w_2}}.$$
   (Hint: You know that $\varphi(z) = a\overline{z} + b$ for some $a, b \in \mathbb{C}$ and that $\varphi(w_1) = w_1$, $\varphi(w_2) = w_2$. Use this to solve for a and b.)
7. Let l be a line and let w be the closest point on l to the origin. Our goal is to prove that the reflection through l has the form
   $$\varphi(z) = -\frac{w}{\overline{w}}\overline{z} + 2w.$$
   a) Prove this assuming that $w \geq 0$.
   b) Write $w = re^{i\theta}$ and $\psi = e^{i\theta}z$. If φ is the reflection through l, describe what sort of similarity $\psi^{-1} \circ \varphi \circ \psi$ is.
   c) Use your answers to the previous parts to prove the result for the general case.
8. Let l be a line passing through two points w_1, w_2. Prove that the glide reflection through the line moving $w_1 \mapsto w_2$ has the form
   $$\varphi(z) = \frac{w_1 - w_2}{\overline{w_1} - \overline{w_2}}\overline{z} + \frac{|w_1|^2 - 2\overline{w_1}w_2 + |w|^2}{\overline{w_2} - \overline{w_1}}.$$
   (Hint: Use the result of Exercise 1.2.6.)
9. Prove Theorem 1.9. *(Hint: you only need to prove that any translation and any rotation around the origin are compositions of reflections.)*
10. Prove Theorem 1.11.

1.3 PROOFS (Calculus)

1. Our goal for this exercise is to give a semi-rigorous proof of Euler's theorem. (For a fully rigorous proof we would need to properly *define* what we mean by e^z for complex inputs. There are various ways to do this—one approach is to say that $f(z) = e^z$ is the unique solution to the differential equation $y' = y$ with initial condition $y(0) = 1$, but this requires defining the complex derivative.)

   a) Compute the Taylor series of e^x centered at $x = 0$.
   b) Substituting ix for x, compute the Taylor series of e^{ix} centered at $x = 0$.
   c) Compute the Taylor series of $\cos(x)$ and $\sin(x)$ centered at $x = 0$.
   d) Using the results of the previous parts, show that the Taylor series of e^{ix} matches the Taylor series of $\cos(x) + i \sin(x)$.
   e) Use Taylor's Remainder Theorem to prove that $e^x, \sin(x), \cos(x)$ are all equal to their Taylor series for all x. Conclude that $e^{ix} = \cos(x) + i \sin(x)$ for all x.

1.4 PROOFS (Group Theory)

1. Let $G = \{\text{even}, \text{odd}\}$ and define

$$\begin{aligned} \text{even} + \text{even} &= \text{even}, \\ \text{even} + \text{odd} &= \text{odd}, \\ \text{odd} + \text{even} &= \text{odd}, \\ \text{odd} + \text{odd} &= \text{even}. \end{aligned}$$

   We will prove that G is a group.

   a) Prove that $+$ is associative. *(Hint: There are only finitely many choices for a, b, c in $a + (b + c) = (a + b) + c$ to check.)*
   b) Prove that "even" is an identity. *(Hint: There are only finitely many choices for a in $a +$ even $=$ even $+ a = a$ to check.)*
   c) Prove that every element a has an inverse.

2. Let G be a group with an identity ι. Suppose that there is another element $e \in G$ with the property that $g * e = e * g = g$ for all $g \in G$.

   a) Why is $\iota * e = \iota$?
   b) Why is $\iota * e = e$?
   c) Why does this prove that the inverse of a group is unique?

3. Let G be a group with an inverse ι. Let g be an element of G, and suppose that there are two elements $h, k \in G$ such that $g * h = h * g = \iota$ and $g * k = k * g = \iota$.

a) Why is $k * g * h = k$?
b) Why is $k * g * h = k$?
c) Why does this prove that the inverse of any element is unique?

4. Let $(G, *, \iota)$ be a group. Prove that H is a subgroup if and only if $(H, *, \iota)$ is a group.
5. Prove Theorem 1.14. *(Hint: you may want to go through Exercise 1.4.4 first.)*
6. Let G, H be groups with identities ι_G and ι_H, and operations $*$ and $\circ$. An *isomorphism* between G and H is a bijective map $\varphi : G \rightarrow H$ such that for all $a, b \in G$, $\varphi(a * b) = \varphi(a) \circ \varphi(b)$. Intuitively, we say that an isomorphism "preserves multiplication." It can also be thought about as a map that renames elements, but keeps the underlying arithmetic the same.

a) Prove that $\varphi(\iota_G) = \iota_H$. *(Hint: use the result of Exercise 1.4.2.)*
b) Prove that $\varphi(a^{-1}) = \varphi(a)^{-1}$. *(Hint: use the result of Exercise 1.4.3.)*
c) Prove that φ^{-1} is also a group isomorphism.

7. Prove that the natural logarithm ln is an isomorphism between $\mathbb{R}^+$ (considered as a group under multiplication) and $\mathbb{R}$ (considered as a group under addition). What is its inverse?
8. For some fixed symbol g, let G be the set of all symbols g^n where $n \in \mathbb{Z}$. Give G a multiplication by $g^m * g^n = g^{m+n}$.

a) Prove that G is a group.
b) Prove that there is an isomorphism between $\mathbb{Z}$ and G.

9. For some two fixed collections of symbols $g_1, g_2, \ldots g_k$, define the *free group* $\langle g_1, g_2, \ldots g_k \rangle$ to be the set of all sequences written in terms of symbols $g_1^{n_1}, \ldots g_k^{n_k}$, where $n_1, \ldots n_k \in \mathbb{Z}$ (e.g. $g_1^3 g_2^4 g_1^{-2}$ is an element of the free group), with the rule that $g_i^m g_i^n = g_i^{m+n}$ for any i in any such sequence (e.g. $g_1 g_4^3 g_4^2 g_6 = g_1 g_4^5 g_6$). For convenience, rather than writing the empty sequence consisting of no symbols as a blank space, we instead write it as ι. Define a multiplication on the free group via concatenation—that is, given any two sequences, we can just write one after the other (e.g. $g_2^3 g_1^2 * g_1^{-3} g_7 = g_2^3 g_1^{-1} g_7$).

a) Prove that $\langle g_1, g_2, \ldots g_k \rangle$ is a group.
b) Prove that there exist elements $x, y \in \langle g_1, g_2, \ldots g_k \rangle$ such that $x * y \neq y * x$ as long as $k > 1$.

10. For some two fixed collection of symbols $g_1, g_2, \ldots g_k$, and some fixed collection of elements $X_1, X_2 \ldots X_r \in \langle g_1, \ldots g_k \rangle$ define the *quotient group* $\langle g_1, g_2, \ldots g_k | X_1, X_2 \ldots X_r \rangle$ to be the elements of $\langle g_1, \ldots g_k \rangle$ with the additional rules that $X_1 = X_2 = \ldots = X_r = \iota$. For example, in $\langle g, h | ghg^{-1}h^{-1} \rangle$, $ghg^{-1}h^{-1} = \iota$, so $gh = hg$—that is, we can freely exchange the order of g and h in this quotient group.

a) Prove that $\langle g_1, g_2, \ldots g_k | X_1, X_2, \ldots X_r\rangle$ is a group.
b) Prove that $\langle g | g^2\rangle$ has just two elements in it. Does it remind you of any other group you have seen? Can you find an isomorphism?

11. a) Let D be any square in the Euclidean plane. Define $\text{Isom}(D)$ to be the collection of isometries φ with the property that $\varphi(D) = D$. Prove that $\text{Isom}(D)$ is a group.
b) Let D_1, D_2 be any two squares in the Euclidean plane. Since they are similar, there exists a similarity $\psi : \mathbb{C} \to \mathbb{C}$ such that $\psi(D_1) = D_2$. Prove that if $\varphi \in \text{Isom}(D_1)$, then $\psi \circ \varphi \circ \psi^{-1} \in \text{Isom}(D_2)$.
c) Prove that

$$\begin{aligned}\Psi : \text{Isom}(D_1) &\to \text{Isom}(D_2)\\ \varphi &\mapsto \psi \circ \varphi \circ \psi^{-1}\end{aligned}$$

is an isomorphism.
d) Why can we now conclude that $\text{Isom}(D)$ has precisely 8 elements, regardless of the choice of square D?

12. The result of Exercise 1.4.11 tells us that the underlying multiplication of $\text{Isom}(D)$ does not really change regardless of which square D is—all that changes is how we write down the elements of the group. Our goal here is to give a standard way to write down this group that makes this property explicit.

a) Define $R(z) = iz$ and $L(z) = \overline{z}$. Prove that $R^4 = \iota$, $L^2 = \iota$, and $(L \circ R)^2 = \iota$. (Here, as for all groups, a^n should be understood as a multiplied by itself n times—recall that in this case, multiplication means composition.)
b) Define $D_8 = \langle L, R | R^4, L^2, LRLR\rangle$ (see Exercise 1.4.10). Prove that

$$\begin{aligned}D_8 &\to \text{Isom}(\diamond)\\ L &\mapsto L\\ R &\mapsto R\end{aligned}$$

defines an isomorphism.

13. Define a set $\mathcal{M}$ consisting of all 3×3 real matrices of the form

$$\begin{pmatrix} a & -b & x\\ b & a & y\\ 0 & 0 & 1\end{pmatrix} \text{ or } \begin{pmatrix} a & b & x\\ b & -a & y\\ 0 & 0 & 1\end{pmatrix}$$

for some a, b, x, y such that $a^2 + b^2 = 1$.

a) Prove that $\mathcal{M}$ is a group if we take the operation to be matrix multiplication.

b) Prove that

$$\begin{aligned} M : \mathrm{Isom}(\mathbb{C}) &\to \mathcal{M} \\ az + b &\mapsto \begin{pmatrix} \Re(a) & -\Im(a) & \Re(b) \\ \Im(a) & \Re(a) & \Im(b) \\ 0 & 0 & 1 \end{pmatrix} \\ a\bar{z} + b &\mapsto \begin{pmatrix} \Re(a) & \Im(a) & \Re(b) \\ \Im(a) & -\Re(a) & \Im(b) \\ 0 & 0 & 1 \end{pmatrix} \end{aligned}$$

is a group isomorphism.

Inversive Geometry

2

In which the true faces of our main characters are revealed, and we consider their actions.

Having understood the simplest linear fractional transformations, our next goal should be to understand maps

$$\varphi(z) = \frac{az+b}{cz+d}$$

as functions on the complex plane. Of course, this statement isn't quite right: if $c \neq 0$, such maps cannot possibly be functions on the complex plane. Indeed, since $c(-d/c) + d = 0$, $\varphi(-d/c)$ can't be defined in the usual way. This is a problem that can be fixed, but it will require introducing a point at infinity. This sentence also contains a more subtle inaccuracy: we won't actually consider all functions of this form, because some of them are very boring. For instance, suppose that $d \neq 0$ and $a = bc/d$. Then

$$\frac{az+b}{cz+d} = \frac{\frac{bc}{d}z+b}{cz+d} = \frac{b(cz+d)}{d(cz+d)} = \frac{b}{d}$$

as long as $cz + d \neq 0$. Constant functions are not interesting and so we will exclude them. What we shall discover is that as long as we require that $ad - bc \neq 0$, φ can be defined and it will be a non-constant function on the complex plane augmented with a point at infinity.

A. Sheydvasser, *Linear Fractional Transformations*, Undergraduate Texts in Mathematics, https://doi.org/10.1007/978-3-031-25002-6_2

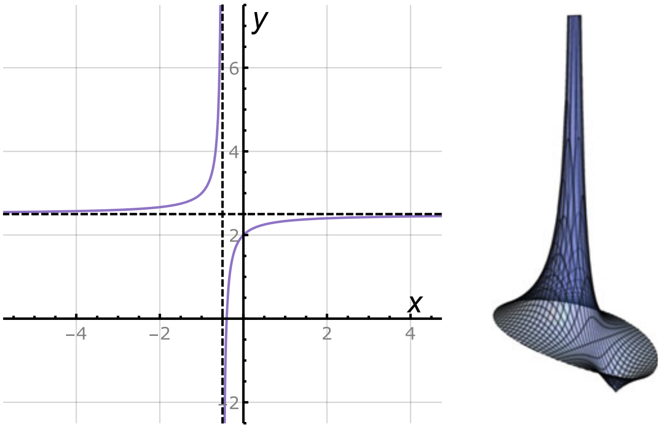


Fig. 2.1 On the left, the graph of $x \mapsto \frac{5x+2}{2x+1}$ plotted over the real numbers. On the right, a small piece of the graph of $z \mapsto \left|\frac{z-5i}{iz+6}\right|^2$ over the complex numbers.

2.1 The Extended Euclidean Plane

Before we begin discussing what a point at infinity is, let's first think about what we would like to define $\varphi(-d/c)$ to be. It is useful to consider what $\varphi(z)$ tends toward as z approaches $-d/c$. For illustrative purposes, let us first consider the case where a, b, c, d are all real numbers and we look at what happens as x approaches $-d/c$ from the left and from the right. From basic calculus, we know that if $c \neq 0$ and $a(-d/c) + b \neq 0$ then

$$\lim_{x \to -d/c^+} \frac{ax+b}{cx+d} = \pm\infty \qquad\qquad \lim_{x \to -d/c^-} \frac{ax+b}{cx+d} = \mp\infty$$

where the sign of the limit depends on the signs of c and $a(-d/c)+b$. Furthermore, the signs of the two limits are always opposite to one another. In either case, as x gets very close to $-d/c$, $|(ax+b)/(cx+d)|$ will either get larger and larger without bound. These asymptotes are shown in Figure 2.1.

This observation suggests that we may want to define $\varphi(-d/c) = \infty$ or $-\infty$—except, which one should it be? After all, which one it depends on the direction we approach from. This gets even more complicated when we generalize to the case where a, b, c, d might be complex numbers and we are looking at approaching from any conceivable direction in the complex plane. Thankfully, there are no such ambiguities if we consider the norm: if a, b, c, d are complex numbers, $c \neq 0$, and $a(-d/c) + b \neq 0$, then

$$\lim_{z \to -d/c} \left| \frac{az+b}{cz+d} \right| = \infty.$$

(See Exercise 2.3.1.) This suggests a simple solution to our problem: augment the complex plane with a single point, which we call the point at infinity.

Definition 2.1 The *extended complex plane*, also known as the *Riemann sphere*, also known as the *complex projective line*, is defined as

$$\mathbb{C}P^1 = \mathbb{C} \cup \{\infty\},$$

where ∞ is an extra formal symbol.

Remark 2.1 This definition is not the only possible way to add points at infinity to $\mathbb{C}$ or, equivalently, $\mathbb{R}^2$. In fact, it isn't even the only widely used construction: another very common one is $\mathbb{R}P^2$, the real projective plane, which adds an entire line at infinity. While this construction is important, we will not make use of it.

The term "extended complex plane" is unlikely to be surprising, but the terms "Riemann sphere" and "complex projective line" probably are—after all, it certainly doesn't look like there is a sphere or a line here. How spheres show up will become apparent momentarily; the reason why this space is a "line" in any sense comes from projective geometry and requires some familiarity with abstract algebra. I refer the interested reader to Hartshorne's *Foundations of Projective Geometry* [6].

In some sense, this definition is incomplete: it only describes what $\mathbb{C}P^1$ is like as a set. It does not give any indication about what other structure $\mathbb{C}P^1$ might have. Is it a group? Does it have some sort of distance function defined on it? The short answer is that neither of those two structures applies (at least, not in natural ways). It *does* have natural structure as a topological space, a manifold, and an algebraic variety, but, unfortunately, all of those are beyond the scope of this book. Again, I refer the reader to Hartshorne [6].

While we define $\mathbb{C}P^1$ in an altogether formal way—it is the set of complex numbers with a single new point added—the intuition about what that point represents is fairly clear: it is supposed to be a point that is infinitely far away from the origin. There are various ways to make this precise, but we will only note that this construction can be understood in terms of *stereographic projection*.

Definition 2.2 Let S^2 be the unit sphere in $\mathbb{R}^3$ centered at the origin. Define the *north pole* $p_N = (0, 0, 1)$. For every point $p \in S^2$ that is not the north pole, there is a unique line $l(p)$ through p and p_N, and this line has a unique intersection $Q(p)$ with the xy-plane. This point has a natural interpretation as a complex number, which we call $S(p)$, the *stereographic projection of* p *onto* $\mathbb{C}$. We call the map

$$\begin{aligned} \text{stereo} : S^2 \backslash \{p_N\} &\to \mathbb{C} \\ p &\mapsto S(p) \end{aligned}$$

stereographic projection.

Figure 2.2 gives an example of how this map works; Figure 2.3 shows a slice of the same. It is geometrically clear that the closer that the point p gets to p_N, the

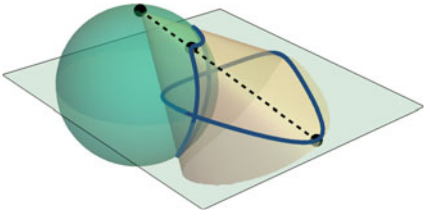


Fig. 2.2 The stereographic projection of a curve on the sphere to a curve on the complex plane.

further away stereo(p) will be from the origin. This suggests an obvious extension of the map stereo, as follows:

$$\overline{\text{stereo}} : S^2 \to \mathbb{C}P^1$$
$$p \mapsto \begin{cases} S(p) & \text{if } p \neq p_N \\ \infty & \text{otherwise.} \end{cases}$$

By slight abuse of notation, we shall also call this map stereographic projection. It has many nice properties, but what we shall care about primarily is that it is bijective and can be given a wonderfully simple algebraic description.

Theorem 2.1 *The map $\overline{stereo}$ is bijective. Indeed,*

$$\overline{stereo} : S^2 \to \mathbb{C}P^1$$
$$(x, y, z) \mapsto \begin{cases} \frac{x+iy}{1-z} & \textit{if } z \neq 1 \\ \infty & \textit{if } z = 1, \end{cases}$$

and its inverse is

$$z \mapsto \begin{cases} \left(\frac{2\Re(z)}{1+|z|^2}, \frac{2\Im(z)}{1+|z|^2}, \frac{|z|^2-1}{|z|^2+1}\right) & \textit{if } z \neq \infty \\ (0, 0, 1) & \textit{if } z = \infty, \end{cases}$$

Proof Choose a point (x, y, z) on the sphere. If $z = 1$, then this point must $(0, 0, 1)$, which we know stereographic projection sends to ∞. Otherwise, we can determine what point in $\mathbb{C}$ it will be sent to as follows: the line through $p_N = (0, 0, 1)$ and (x, y, z) is the set of points of the form

$$(0, 0, 1)(1 - t) + (x, y, z)t = (xt, yt, (z - 1)t + 1)$$

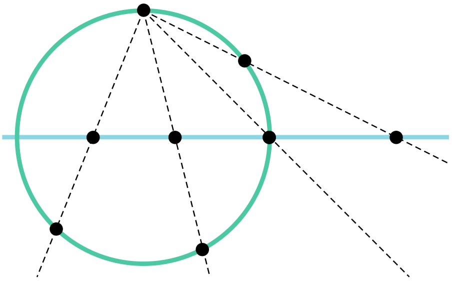


Fig. 2.3 Stereographic projection restricted to the $y = 0$ plane.

for $t \in \mathbb{R}$, and the intersection of this line with the xy-plane happens precisely when the z-component is zero—that is, when $(z-1)t+1 = 0$, or $t = 1/(1-z)$. Therefore, $S((x, y, z)) = \frac{x+iy}{1-z}$, as claimed. Next, we check that the inverse is as claimed. For any $(x, y, z) \in S^2\backslash\{p_N\}$, let

$$w = \frac{x+iy}{1-z}$$

and note that

$$\begin{aligned}\overline{\text{stereo}}^{-1}(w) &= \left(\frac{2\Re(w)}{1+|w|^2}, \frac{2\Im(w)}{1+|w|^2}, \frac{|w|^2-1}{|w|^2+1}\right)\\
&= \left(\frac{2\frac{x}{1-z}}{1+\frac{x^2+y^2}{(1-z)^2}}, \frac{2\frac{y}{1-z}}{1+\frac{x^2+y^2}{(1-z)^2}}, \frac{\frac{x^2+y^2}{(1-z)^2}-1}{\frac{x^2+y^2}{(1-z)^2}+1}\right)\\
&= \left(\frac{2x(1-z)}{(1-z)^2+x^2+y^2}, \frac{2y(1-z)}{(1-z)^2+x^2+y^2}, \frac{x^2+y^2-(1-z)^2}{x^2+y^2+(1-z)^2}\right)\\
&= \left(\frac{2x(1-z)}{1-2z+x^2+y^2+z^2}, \frac{2y(1-z)}{1-2z+x^2+y^2+z^2}, \frac{x^2+y^2-1+2z-z^2}{x^2+y^2+z^2+1-2z}\right)\\
&= \left(\frac{2x(1-z)}{2(1-z)}, \frac{2y(1-z)}{2(1-z)}, \frac{2z-2z^2}{2-2z}\right) = (x, y, z),\end{aligned}$$

whence $(\overline{\text{stereo}}^{-1} \circ \overline{\text{stereo}})(p) = p$ for all $p \in S^2$. I leave proving that composing in the other order also gives the identity as an exercise for the reader. (See Exercise 2.2.3.) □

It now makes perfect sense to define $\varphi(z) = (az+b)/(cz+d)$ as a function that returns points in the extended complex plane $\mathbb{C}P^1$. However, we would like the domain to be the same as the codomain so that φ is really a transformation of a particular space. We can accomplish this without too much difficulty. All we need to do is to define $\varphi(\infty)$ and once again calculus comes to the rescue—if $ad - bc \neq 0$ and $c \neq 0$, then

$$\lim_{z\to\infty} \frac{az+b}{cz+d} = \frac{a}{c}.$$

(See Exercise 2.3.2.) In words, what we are saying is that as z gets farther and farther away from the origin (irrespective of which direction), $(az+b)/(cz+d)$ gets closer and closer to a/c. This makes perfect intuitive sense: if $|z|$ is large, then $az+b \approx az$ and $cz+d \approx cz$, and so $(az+b)/(cz+d) \approx (az)/(cz) = a/c$. This, finally, allows us to make an unambiguous definition of linear fractional transformations.

Definition 2.3 A *linear fractional transformation* on $\mathbb{C}P^1$ is a map of the form

$$\begin{aligned} \varphi : \mathbb{C}P^1 &\to \mathbb{C}P^1 \\ z &\mapsto \begin{cases} \frac{a}{c} & \text{if } z = \infty,\ c \neq 0 \\ \infty & \text{if } z = \infty,\ c = 0 \text{ or if } z = -\frac{d}{c},\ c \neq 0 \\ \frac{az+b}{cz+d} & \text{otherwise,} \end{cases} \end{aligned}$$

for some $a, b, c, d \in \mathbb{C}$ such that $ad - bc \neq 0$.

Remark 2.2 If $ad - bc = 0$, then either φ will be undefined, or it will be a constant function. (See Exercise 2.2.1.)

One possible objection to this definition is that it is not very elegant: it requires splitting into four different cases. There are various ways to rectify this. One possible solution is to write

$$z \mapsto \lim_{w\to z} \frac{aw+b}{cw+d},$$

with the understanding that a limit that fails to exist should be understood as returning ∞. Another option is to phrase everything in terms of projective geometry, but we will not pursue this notion in this text.

► **Example** *Let* $\varphi(z) = (3+z)/(z-1)$. *Find* $\varphi(\infty)$ *and find the* $z \in \mathbb{C}P^1$ *such that* $\varphi(z) = \infty$. *How many such* z *are there? Would the answer be different for a different linear fractional transformation* φ'?
Since

$$\lim_{z\to\infty} \frac{z+3}{z-1} = 1,$$

we see that $\varphi(\infty) = 1$. On the other hand, per the definition of the linear fractional transformation, $\varphi(z) = \infty$ only if the denominator is 0—that is, if $z - 1 = 0$. Therefore, $z = 1$ is the unique element such that $\varphi(z) = \infty$. It is enough to see that this same result will hold true for any linear fractional transformation: either $c = 0$, and $\varphi(z) = \infty$ if and only if $z = \infty$, or $c \neq 0$, and $\varphi(z) = \infty$ if and only if $cz + d = 0$. In either case, there is always exactly one point for which this happens.

2.2 A Little Bit More Group Theory

Linear fractional transformations have many beautiful properties. The first among these that we will prove is that they form a group. To be a bit more precise, it is convenient to give a definition.

Definition 2.4 The set of all linear fractional transformations $\varphi : \mathbb{C}P^1 \to \mathbb{C}P^1$ shall be denoted $\text{Möb}^0(2)$.

Remark 2.3 The notation $\text{Möb}^0(2)$ to denote linear fractional transformations is liable to raise some eyebrows. I promise that there is a good reason for it, which will be explained later in this chapter.

What we are going to prove is that $\text{Möb}^0(2)$, together with function composition $\circ$ as the operation, is a group. First, we will need a few lemmas.

Lemma 2.1 *Let $\varphi_1, \varphi_2 \in Möb^0(2)$. Then $\varphi_1 \circ \varphi_2 \in Möb^0(2)$.*

Proof This is a simple algebra exercise. Write

$$\varphi_1(z) = \lim_{w\to z} \frac{aw+b}{cw+d}$$

and

$$\varphi_2(z) = \lim_{w\to z} \frac{a'w+b'}{c'w+d'}.$$

Then

$$\begin{aligned}(\varphi_1 \circ \varphi_2)(z) &= \lim_{w\to z} \frac{a\varphi_2(w)+b}{c\varphi_2(w)+d} \\ &= \lim_{w\to z} \frac{a\lim_{w'\to w}\frac{a'w'+b'}{c'w'+d'}+b}{c\lim_{w'\to w}\frac{a'w'+b'}{c'w'+d'}+d}.\end{aligned}$$

However, all of our functions are continuous, so we may pull out the limit as $w' \to w$. This allows us to simplify to

$$\begin{aligned}(\varphi_1 \circ \varphi_2)(z) &= \lim_{w\to z}\lim_{w'\to w} \frac{a\frac{a'w'+b'}{c'w'+d'}+b}{c\frac{a'w'+b'}{c'w'+d'}+d} \\ &= \lim_{w\to z} \frac{a\frac{a'w+b'}{c'w+d'}+b}{c\frac{a'w+b'}{c'w+d'}+d} \\ &= \lim_{w\to z} \frac{a(a'w+b')+b(c'w+d')}{c(a'w+b')+d(c'w+d')} \\ &= \lim_{w\to z} \frac{(aa'+bc')w+(ab'+bd')}{(a'c+c'd)w+(b'c+dd')}.\end{aligned}$$

We see that the expression at the end is of the right form once we check that

$$\begin{aligned}(aa'+bc')(b'c+dd')-(ab'+bd')(a'c+c'd)&\\ &=aa'b'c+aa'dd'+bb'cc'+bc'dd'\\ &-aa'b'c-ab'c'd-a'bcd'-bc'dd'\\ &=aa'dd'+bb'cc'-ab'c'd-a'bcd'\\ &=(ad-bc)(a'd'-b'c')\neq 0,\end{aligned}$$

whence it is a linear fractional transformation. □

The trick of rearranging the limit whenever we encounter compositions of linear fractional transformations will always work, but I maintain that it is something that only needs to be seen once. As such, henceforth, we shall leave off writing the limit entirely. Instead, by slight abuse of notation, we will simply write

$$\begin{aligned}\varphi : \mathbb{C}P^1 &\to \mathbb{C}P^1\\ z &\mapsto \frac{az+b}{cz+d},\end{aligned}$$

without worrying about the exceptional points.

Lemma 2.2 *For any $\varphi \in M\ddot{o}b^0(2)$, there exists $\varphi^{-1} \in M\ddot{o}b^0(2)$ with the property that $\varphi(\varphi^{-1}(z)) = \varphi^{-1}(\varphi(z)) = z$ for all $z \in \mathbb{C}P^1$.*

Proof Write

$$\begin{aligned}\varphi(z) &= \frac{az+b}{cz+d}\\ \varphi^{-1}(z) &= \frac{dw-b}{-cw+a}.\end{aligned}$$

It is clear that φ^{-1} is a linear fractional transformation since $da-(-b)(-c) = ad-bc \neq 0$. It remains to show that the correct composition law holds. I leave this as an exercise to the reader. (See Exercise 2.2.2.) □

While there is nothing particularly difficult about the proof of Lemma 2.2, where the function φ^{-1} came from is likely a little mysterious. In principle, we could have deduced it from the composition law we derived in Lemma 2.1. However, this would be messy. A more elegant description will present itself once we connect linear fractional transformations and matrices.

Theorem 2.2 *The set of linear fractional transformations $M\ddot{o}b^0(2)$ is a group if we take the group operation to be composition and $\imath(z) = (1\cdot z+0)/(0\cdot z+1) = z$.*

Proof The group operation is closed by Lemma 2.1. It is associative since function composition is always associative. It is clear that $\imath$ is a linear fractional transformation, hence it is the identity. Inverses exist by Lemma 2.2. □

This result helps shed light on why linear fractional transformations should be important—they form a group of transformations on a space, and so if we study what type of properties such transformations preserve, we will be studying a new kind of geometry. While this group might seem unfamiliar, I claim that it is intimately tied to another common group.

Definition 2.5 The *general linear group on* $\mathbb{C}^2$, denoted by $GL(2, \mathbb{C})$, consists of all 2×2 matrices with complex coefficients and non-zero determinant. That is, if

$$M = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in GL(2, \mathbb{C}),$$

then $\det M = ad - bc \neq 0$.

Remark 2.4 The term "linear" comes from "linear transformation." This group is just the collection of linear transformations from $\mathbb{C}^2$ to $\mathbb{C}^2$ that are invertible, represented as matrices.

Theorem 2.3 *If we take the operation to be matrix multiplication, then $GL(2, \mathbb{C})$ is a group.*

Proof Write

$$M = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \quad M' = \begin{pmatrix} a' & b' \\ c' & d' \end{pmatrix}$$

and note that

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} a' & b' \\ c' & d' \end{pmatrix} = \begin{pmatrix} aa' + bc' & ab' + bd' \\ ca' + dc' & cb' + dd' \end{pmatrix},$$

and since

$$\begin{aligned}(aa' + bc')(cb' + dd') - (ab' + bd')(ca' + dc')& \\ &= aa'b'c + aa'dd' + bb'cc' + bc'dd' \\ &- aa'b'c - ab'c'd - a'bcd' - bc'dd' \\ &= aa'dd' + bb'cc' - ab'c'd - a'bcd' \\ &= (ad - bc)(a'd' - b'c') \neq 0,\end{aligned}$$

we see that $MM' \in GL(2, \mathbb{C})$. It is easy to check from the above that if we take

$$I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

then it will satisfy the properties of an identity. The existence of inverses is similarly easy to check:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix} = \begin{pmatrix} ad - bc & 0 \\ 0 & ad - bc \end{pmatrix} = \begin{pmatrix} d & -b \\ -c & a \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix},$$

hence

$$\frac{1}{ad - bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

is an inverse of our matrix. The only thing remaining is to prove that matrix multiplication is associative. This is a standard linear algebra exercise. (One way to see it is that multiplication of matrices corresponds to composition of linear transformations. Since composition of linear transformations is just function composition, it is associative.) □

Now, there is something deeply surprising here, which the careful reader might have picked up on: the calculations that we used to compute the product of matrices look oddly similar to the calculations that we used to compute the composition of linear fraction transformations! In fact, we have unwittingly proved a very interesting result.

Theorem 2.4 *There exists a surjective map*

$$\Psi : GL(2, \mathbb{C}) \to M\ddot{o}b^0(2)$$
$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \mapsto \left(z \mapsto \frac{az + b}{cz + d}\right)$$

with the property that $\Psi(M_1 M_2) = \Psi(M_1) \circ \Psi(M_2)$ *for all* $M_1, M_2 \in GL(2, \mathbb{C})$.

Proof I leave this as an exercise to the reader. (See Exercise 2.2.4.) □

This correspondence gives a very handy computational tool. Rather than being forced to think about function composition to work out what linear fractional transformation $\varphi_1 \circ \varphi_2$ gives, we can instead pass from φ_1 and φ_2 to their corresponding matrices, multiply those, and voilà! The result gives the coefficients of $\varphi_1 \circ \varphi_2$. For example, if we take

$$\varphi_1(z) = \frac{2z + i}{iz - 1} \qquad \varphi_2(z) = \frac{iz - 1 - i}{z},$$

then we can multiply the corresponding matrices and get the composition

$$\begin{pmatrix} 2 & i \\ i & -1 \end{pmatrix}\begin{pmatrix} i & -1 - i \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 3i & -2 - 2i \\ -2 & 1 - i \end{pmatrix}$$
$$(\varphi_1 \circ \varphi_2)(z) = \frac{3iz - 2 - 2i}{-2z + 1 - i}.$$

However, we have to be a little careful. Certainly, every single linear fraction transformation can be represented by a matrix in $GL(2, \mathbb{C})$. However, this representation is not unique, since for any $\lambda \in \mathbb{C}^\times$,

$$\frac{az + b}{cz + d} = \frac{\lambda az + \lambda b}{\lambda cz + \lambda d},$$

hence

$$\Psi\left(\begin{pmatrix} a & b \\ c & d \end{pmatrix}\right) = \Psi\left(\begin{pmatrix} \lambda a & \lambda b \\ \lambda c & \lambda d \end{pmatrix}\right).$$

Thankfully, there are standard ways to fix this issue. The intuitive idea is that you create a new group from $GL(2, \mathbb{C})$, but one in which matrices that differ by multiplication by a non-zero constant are treated as being one and the same. This new group is called $PGL(2, \mathbb{C})$, and in a certain precise sense, it is exactly the same as $\text{Möb}^0(2)$! (See Exercise 2.4.6.)

In any case, this map and its computational aid merit additional discussion. Why? Because it has greater repercussions than just for this one particular group; it is a much more broadly applicable phenomenon.

Definition 2.6 Let $(G, *, \imath_G)$, $(H, \circ, \imath_H)$ be two groups. A function $f : G \to H$ is called a *group homomorphism* if for all $g_1, g_2 \in G$, $f(g_1 * g_2) = f(g_1) \circ f(g_2)$. It is called a *group isomorphism* if additionally f is bijective. If there exists a group isomorphism between G and H, we say that they are *isomorphic*.

It is easy to show that a group homomorphism is a group isomorphism if and only if it has an inverse and that inverse is also a group homomorphism. (See Exercise 2.4.2.) We have just demonstrated an example of a group homomorphism: namely, the map $\Psi : GL(2, \mathbb{C}) \to \text{Möb}^0(2)$. We also hinted at a group isomorphism: a map $PGL(2, \mathbb{C}) \to \text{Möb}^0(2)$. But there are many, many other examples. For instance,

1. $\mathbb{R}^+$ (considered as a group under multiplication) is isomorphic to $\mathbb{R}$ (considered as a group under addition) via the map $\ln : \mathbb{R}^+ \to \mathbb{R}$. (See Exercise 1.4.7.)
2. For any two groups G, H, the map $\varphi : G \to H$ defined by $\varphi(g) = \imath_H$ is a group homomorphism.
3. If H is a subgroup of G, then the inclusion map $H \to G$ defined simply by $h \mapsto h$ is a group homomorphism. (See Exercise 2.4.3.)

We will see far more examples in later chapters, as well as in the exercises. The importance of group isomorphisms is perhaps a little easier to understand: two groups are isomorphic if and only if they are essentially "the same." I mean this in the following sense: if $f : G \to H$ is a group isomorphism, then for any $a, b, c \in G$, $ab = c$ if and only if $f(a)f(b) = f(c)$. This means that if we were to write out the multiplication tables of both G and H, they would look the same—we would just be relabeling the various group elements via f. Thus, important properties of groups are preserved by group isomorphisms. For example, if G and H are isomorphic, then one is abelian if and only if the other is abelian. (See Exercise 2.4.7.)

However, our example of $GL(2, \mathbb{C}) \to \text{Mob}^0(2)$ shows that group homomorphisms are important even if they are not isomorphisms. Yes, in general, group homomorphisms will not preserve all nice group properties like abelian-ness. However, if $f : G \to H$ is a group homomorphism, you can still learn something about the multiplication table of H from the multiplication table of G, and vice versa. This leads us to the following general principle.

Philosophical Principle

If you wish to study a mathematical object (such as the Euclidean plane or groups), don't just study it in isolation. Instead, identify what are the transformations between this kind of mathematical object that preserve something important about it. (Such as how isometries preserve distance, or how group homomorphisms preserve multiplication.)

▶ **Example** *Let* $\varphi(z) = \frac{z+2}{iz+1+i}$*. Find the subset of linear fractional transformations* ψ *such that* $\varphi \circ \psi$ *is a translation.*

We want that $(\varphi \circ \psi)(z) = z + b$ for some $b \in \mathbb{C}$. Passing to the corresponding matrices, we have

$$\begin{pmatrix} 1 & 2 \\ i & 1+i \end{pmatrix} M = \begin{pmatrix} 1 & b \\ 0 & 1 \end{pmatrix},$$

or, taking an inverse,

$$\begin{aligned} M = \begin{pmatrix} 1 & 2 \\ i & 1+i \end{pmatrix}^{-1} \begin{pmatrix} 1 & b \\ 0 & 1 \end{pmatrix} &= \frac{1}{1-i} \begin{pmatrix} 1+i & -2 \\ -i & 1 \end{pmatrix} \begin{pmatrix} 1 & b \\ 0 & 1 \end{pmatrix} \\ &= \frac{1}{1-i} \begin{pmatrix} 1+i & -2+(1+i)b \\ -i & 1-ib \end{pmatrix}. \end{aligned}$$

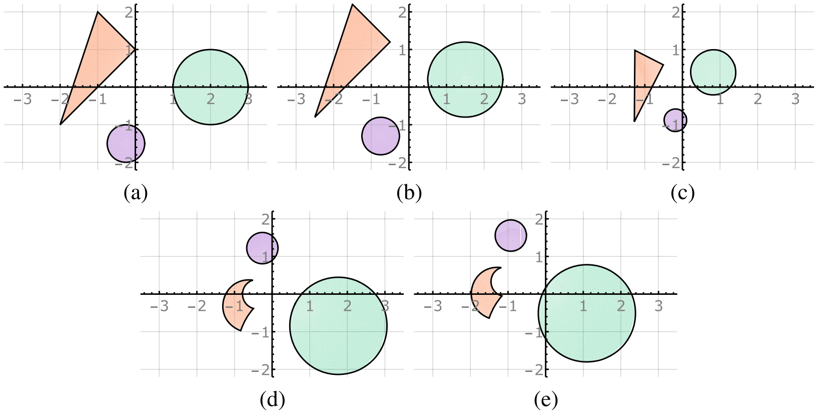
Since we can freely multiply by scalars without changing the original linear fractional transformation, we can just ignore the factor of $(1-i)^{-1}$ in front and conclude that

$$\left\{ z \mapsto \frac{(1+i)z - 2 + (1+i)b}{-iz + 1 - ib} \,\middle|\, b \in \mathbb{C} \right\}$$

is the desired family of linear fractional transformations.

2.3 Circle Inversions

The fact that $\text{Möb}^0(2)$ is a group suggests that we might be able to use similar reasoning to our approach in studying $\text{Sim}(\mathbb{C})$ in Chapter 1. Namely, we will first break it apart into more simple transformations. We will understand those simple transformations as thoroughly as we can, and use the fact that $\text{Möb}^0(2)$ is a group to show that various properties preserved by simple transformations are preserved by



(a) (b) (c)

(d) (e)

Fig. 2.4 An illustration of the decomposition in Theorem 2.5. (a) shows an initial configuration; (b) shows a translation of (a); (c) shows a rotation and scaling of (b); (d) shows the image of (c) under the map $z \mapsto 1/z$; (e) shows a translation of (d).

more complicated ones. What are these simple transformations? This is answered by the following theorem.

Theorem 2.5 (Decomposition Theorem for LFTs on $\mathbb{C}P^1$) *Any element in Möb*$^0(2)$ *can be written as a composition of rotations, translations, dilations, and the map* $z \mapsto 1/z$.

Remark 2.5 A concrete example of this decomposition is shown in Figure 2.4.

Proof Choose any $\varphi \in \text{Möb}^0(2)$, and write

$$\varphi(z) = \frac{az+b}{cz+d}.$$

This theorem is most convenient to prove in terms of $GL(2,\mathbb{C})$ via Theorem 2.4, so we shall actually consider the related matrix

$$M = \begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

We shall prove that M can be written as a product of matrices of the forms

$$\begin{pmatrix} 1 & b \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} a & 0 \\ 0 & d \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

as, by inspection, such matrices correspond to translations, rotations/dilations, and $z \mapsto 1/z$, respectively. We are going to have two different cases: either $c = 0$ or $c \neq 0$. If $c = 0$, then

$$M = \begin{pmatrix} a & b \\ 0 & d \end{pmatrix} = \begin{pmatrix} a & 0 \\ 0 & d \end{pmatrix} \begin{pmatrix} 1 & \frac{b}{a} \\ 0 & 1 \end{pmatrix},$$

which proves what we wanted. If $c \neq 0$, then we first note that

$$\begin{pmatrix} 1 & -\frac{a}{c} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} 0 & b - \frac{ad}{c} \\ c & d \end{pmatrix},$$

and

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 0 & b - \frac{ad}{c} \\ c & d \end{pmatrix} = \begin{pmatrix} c & d \\ 0 & b - \frac{ad}{c} \end{pmatrix} = \begin{pmatrix} c & 0 \\ 0 & -\frac{ad-bc}{c} \end{pmatrix} \begin{pmatrix} 1 & \frac{d}{c} \\ 0 & 1 \end{pmatrix}.$$

Thus,

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & -\frac{a}{c} \\ 0 & 1 \end{pmatrix} M = \begin{pmatrix} c & 0 \\ 0 & -\frac{ad-bc}{c} \end{pmatrix} \begin{pmatrix} 1 & \frac{d}{c} \\ 0 & 1 \end{pmatrix},$$

from which we get, by multiplying by inverse matrices,

$$M = \begin{pmatrix} 1 & \frac{a}{c} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} c & 0 \\ 0 & -\frac{ad-bc}{c} \end{pmatrix} \begin{pmatrix} 1 & \frac{d}{c} \\ 0 & 1 \end{pmatrix}.$$

Passing to the corresponding linear fraction transformations, we precisely get the desired result. □

Remark 2.6 To students of linear algebra, our proof may seem vaguely familiar: it is essentially Gaussian row reduction with some minor alterations.

We studied translations, rotations, and dilations extensively in Chapter 1, but the map $\varphi(z) = 1/z$ is new. What does it do to $\mathbb{C}P^1$? Well, there are two points where its action is completely obvious:

$$\varphi(0) = \infty \qquad\qquad \varphi(\infty) = 0.$$

That is, the map $z \mapsto 1/z$ interchanges 0 and the point at infinity. What about every other point? Any other $z \in \mathbb{C}P^1$ we can write as $z = re^{i\theta}$ for some $r > 0$, and then we check that

$$\varphi(re^{i\theta}) = \frac{1}{re^{i\theta}} = \frac{1}{r}e^{-i\theta}.$$

Therefore, we see that $z \mapsto 1/z$ does two things: first, it moves points that are distance r away from the origin to points that are distance $1/r$ away from the origin; second, it does a reflection around the real axis. This makes it deeply tied to circle inversions.

Definition 2.7 Let C be a circle in $\mathbb{C}$ with center z_0 and radius R. A *reflection through C*, also known as a *circle inversion*, is a transformation Φ on $\mathbb{C}P^1$ defined as follows: $\Phi(z_0) = \infty$, $\Phi(\infty) = z_0$, and for every other point $z \in \mathbb{C}P^1$, take the

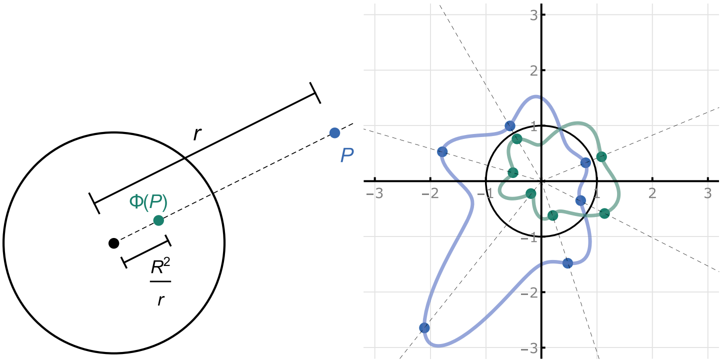


Fig. 2.5 On the left, an illustration of the effect of a circle inversion on a single point: P is distance r from the center, and it is sent to $\Phi(P)$, which is R^2/r from the center, where R is the radius of the circle. On the right, the effect of an inversion through the unit circle. The loop in blue is the original; the loop in green is its image after the inversion.

ray from z_0 to z, measure the distance r between them, and send z to the point along that same ray that is distance R^2/r away from z_0. (See Figure 2.5 for an illustration.)

It might not be clear why we refer to such a transformation as a "reflection"—we will show later that you can get reflections through lines as limits of reflections through circles in some sense. Indeed, reflections through circles and reflections through lines share various similarities. For example, just as reflections through lines fix a particular line and interchange the two areas on either side of it, reflections through circles do the same but with circles. Furthermore, they do this exchange in such a way that they are their own inverses.

Theorem 2.6 *Let Φ be a reflection through a circle C. Then for all $z \in C$, $\Phi(z) = z$, all points inside C are sent to points outside C, and all points outside C are sent to points inside C.*

Remark 2.7 This effect can be seen in Figures 2.6 and 2.7.

Proof Let C have center z_0 and radius R. Choose any point z on C—by definition, z is distance R away from z_0. The inversion Φ will send z to a point distance $R^2/R = R$ away from z_0 along the same ray, which is to say that $\Phi(z) = z$. If $z = z_0$, then z is inside C, but $\Phi(z) = \infty$, which is outside C. If $z = \infty$, then z is outside C, but $\Phi(z) = z_0$, which is inside C. For all other $z \in \mathbb{C}P^1$, let $r > 0$ be its distance away from z_0. If $r < R$, then z is inside C, but $\Phi(z)$ will be distance $R^2/r > R$ away from z_0, hence outside C. If $r > R$, then z is outside C, but $\Phi(z)$ will be distance $R^2/r < R$ away from z_0, hence inside C. □

Theorem 2.7 *Let Φ be a reflection through a circle C. Then Φ is its own inverse.*

Proof Let C have center z_0 and radius R. Then

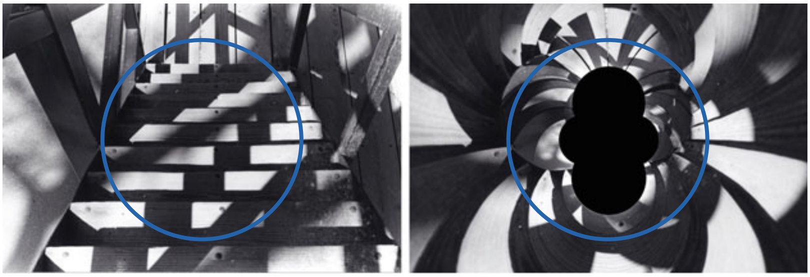


Fig. 2.6 On the left is the original image. On the right is its image under the reflection through the blue circle.

$$\Phi(\Phi(z_0)) = \Phi(\infty) = z_0 \qquad\qquad \Phi(\Phi(\infty)) = \Phi(z_0) = \infty.$$

For any other point z, z is some distance $r > 0$ away from z_0, so $\Phi(z)$ is distance R^2/r away from z_0 along the same ray. Ergo, $\Phi(\Phi(z))$ is distance $R^2/(R^2/r) = r$ away from z_0 along the same ray, which is to say that $\Phi(\Phi(z)) = z$. Thus, we see that $\Phi(\Phi(z)) = z$ for all $z \in \mathbb{C}P^1$ and we have proved our claim. □

The map $z \mapsto 1/z$ is not quite a circle reflection: it is a circle reflection composed with a line reflection. Well, technically this cannot possibly be true. After all, any line reflection is only defined on $\mathbb{C}$, whereas we need it to be defined on all of $\mathbb{C}P^1$. There is an easy fix for this.

Definition 2.8 A *line reflection* on $\mathbb{C}P^1$ is a transformation $\Phi : \mathbb{C}P^1 \to \mathbb{C}P^1$ defined as follows: restricted to $\mathbb{C}$, Φ is a reflection across some line l, and $\Phi(\infty) = \infty$. More generally, for any similarity $\Phi : \mathbb{C} \to \mathbb{C}$, we can extend it to a transformation $\mathbb{C}P^1 \to \mathbb{C}P^1$ by defining $\Phi(\infty) = \infty$.

Theorem 2.8 *Define $\varphi(z) = 1/z$ as a function on $\mathbb{C}P^1$. Then φ is a composition of inversion through the unit circle and a reflection across the real axis.*

Proof Let Φ be a reflection through the unit circle, and $\phi(z) = \overline{z}$ be the reflection across the real axis. Then $\Phi(0) = \infty$, and $(\phi \circ \Phi)(0) = \infty = \varphi(0)$. Similarly, $\Phi(\infty) = 0$, hence $(\phi \circ \Phi)(\infty) = 0 = \varphi(\infty)$. For every other z, we can write it as $re^{i\theta}$ for some $r > 0$. By the definition of Φ, we have

$$\Phi(re^{i\theta}) = \frac{1}{r}e^{i\theta} \qquad\qquad (\phi \circ \Phi)\left(re^{i\theta}\right) = \frac{1}{r}e^{-i\theta} = \varphi(re^{i\theta}).$$

Thus, $\phi \circ \Phi = \varphi$. □

Conversely, any circle reflection can be expressed as a composition of a linear fractional transformation and a line reflection.

Theorem 2.9 *Let Φ be a reflection through a circle C. Then*

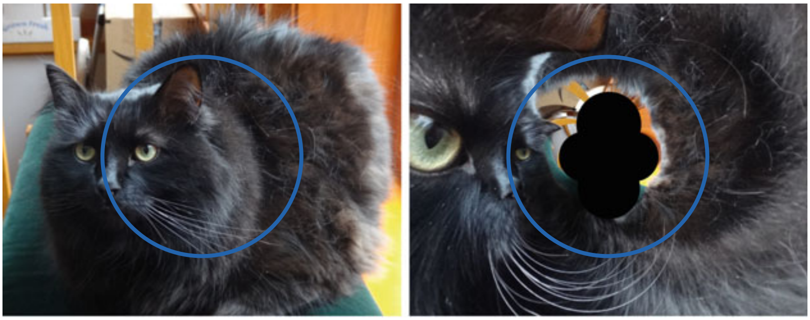


Fig. 2.7 On the left is the best cat. On the right is his image under the reflection through the blue circle.

$$\Phi(z) = \frac{z_0\overline{z} + R^2 - |z_0|^2}{\overline{z} - \overline{z_0}}$$

for all $z \in \mathbb{C}P^1$.

Proof Any circle C with center z_0 and radius R is the image of the unit circle after a dilation $\varphi_1(z) = Rz$ and a translation $\varphi_2(z) = z + z_0$. If $\varphi_3(z) = 1/\overline{z}$, we claim that

$$\varphi_2 \circ \varphi_1 \circ \varphi_3 \circ (\varphi_2 \circ \varphi_1)^{-1} = \Phi.$$

Intuitively, the idea is that we first change coordinates so that the circle C turns into the unit circle; then we do φ_3, which is a reflection through the unit circle; finally, we change coordinates back. This should be the same as reflecting through C. Now,

$$\begin{aligned}
\left(\varphi_2 \circ \varphi_1 \circ \varphi_3 \circ (\varphi_2 \circ \varphi_1)^{-1}\right)(z) &= \left(\varphi_2 \circ \varphi_1 \circ \varphi_3 \circ \varphi_1^{-1} \circ \varphi_2^{-1}\right)(z) \\
&= \left(\varphi_2 \circ \varphi_1 \circ \varphi_3 \circ \varphi_1^{-1}\right)(z - z_0) \\
&= (\varphi_2 \circ \varphi_1 \circ \varphi_3)\left(\frac{z - z_0}{R}\right) \\
&= (\varphi_2 \circ \varphi_1)\left(\frac{R}{\overline{z} - \overline{z_0}}\right) \\
&= \varphi_2\left(\frac{R^2}{\overline{z} - \overline{z_0}}\right) = \frac{R^2}{\overline{z} - \overline{z_0}} + z_0 \\
&= \frac{z_0\overline{z} + R^2 - |z_0|^2}{\overline{z} - \overline{z_0}},
\end{aligned}$$

and from this it is easy to see that $z_0 \mapsto \infty$ and $\infty \mapsto z_0$, as expected. For all other $z \in \mathbb{C}P^1$, we can write $z = z_0 + re^{i\theta}$ for some $r > 0$ and it is an easy computation that

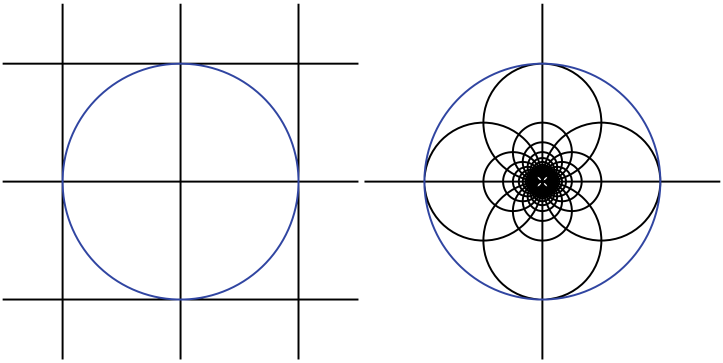


Fig. 2.8 On the left is the unit square grid. On the right is its image under the map $z \mapsto 1/z$.

$$\left(\varphi_2 \circ \varphi_1 \circ \varphi_3 \circ \varphi_1^{-1} \circ \varphi_2^{-1}\right)\left(z_0 + re^{i\theta}\right) = z_0 + \frac{R^2}{r}e^{i\theta},$$

which is precisely what the action of Φ should do. We are done. □

▶ **Example** *Let* Φ_1, Φ_2 *be the circle reflections through the circles centered at* 0 *with radii* 1 *and* 2, *respectively. What is* $\Phi_2 \circ \Phi_1$?

First, we note that $(\Phi_2 \circ \Phi_1)(0) = \Phi_2(\infty) = 0$ and $(\Phi_2 \circ \Phi_1)(\infty) = \Phi_2(0) = \infty$. Any other point $z \in \mathbb{C}P^1$, we can write as $re^{i\theta}$, and we see that

$$(\Phi_2 \circ \Phi_1)(re^{i\theta}) = \Phi_2\left(\frac{1}{r}e^{i\theta}\right) = \frac{2^2}{1/r}e^{i\theta} = 4re^{i\theta}.$$

From this, we get that $\Phi_2(\Phi_1(z)) = 4z$, which is a dilation.

2.4 Generalized Circles

From the various illustrations that we have given so far, we see that circle inversions are not isometries nor even similarities—correspondingly, neither is $z \mapsto 1/z$. Recall that we showed that one of the defining properties of similarities was that they sent lines to lines and circles to circles. However, examples such as in Figure 2.8 show that while $z \mapsto 1/z$ sometimes sends lines to lines, it does not always.

However, studying these illustrations carefully, one notices something surprising: while $z \mapsto 1/z$ doesn't always send lines to lines, it looks like when it doesn't, it sends them to circles! This is indeed true, but to prove it, it shall be convenient to introduce the concept of a generalized circle.

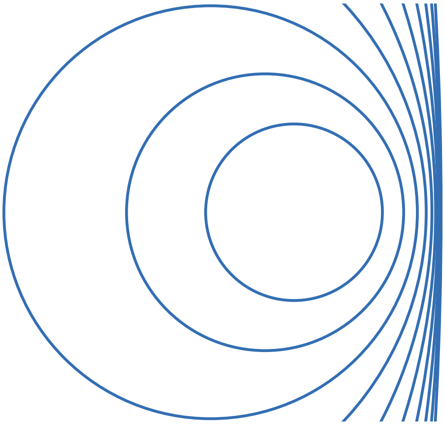


Fig. 2.9 A family of circles that seem to approach a line.

Definition 2.9 A *(generalized) circle* in $\mathbb{C}P^1$ is either a circle in $\mathbb{C}$ or a line in $\mathbb{C}$ union $\{\infty\}$. We often call lines *circles through infinity*.

This definition can be motivated in various ways. One possible way is to observe that as the radius of a circle increases, in some sense, it can start to approach a line. Such a statement is technically meaningless without specifying a mode of convergence, but the intuition is clear from illustrations such as Figure 2.9. One could also motivate this definition in terms of the Riemann sphere—it is possible to show that both lines and circles on the sphere correspond to circles on the sphere via stereographic projection. In any case, in order to prove that $z \mapsto 1/z$ preserves generalized circles, we shall want an algebraic description of them that unifies the descriptions of circles and lines.

Theorem 2.10 (Algebraic Description of Generalized Circles) *Generalized circles are precisely those curves in $\mathbb{C}P^1$ that are solutions to equations of the form $Az\overline{z} + Bz + \overline{B}z + C = 0$, where $A, C \in \mathbb{R}$, $B \in \mathbb{C}$, and*

$$\det \begin{pmatrix} A & B \\ \overline{B} & C \end{pmatrix} = AC - B\overline{B} < 0.$$

Remark 2.8 There is an obvious question: how do we define whether or not ∞ is a solution to such an equation? In general, questions like this involve appeals to projective geometry. For our purposes, we approach as follows: if we divide by $z\overline{z}$ on both sides, we get

$$A + \frac{B}{\overline{z}} + \frac{\overline{B}}{z} + \frac{C}{z\overline{z}} = 0.$$

In light of this, we will define ∞ as being a solution to this equation if

$$\lim_{z\to\infty} A + \frac{B}{\overline{z}} + \frac{\overline{B}}{z} + \frac{C}{z\overline{z}} = 0.$$

It isn't hard to see that this is true if and only if $A = 0$.

Proof We know that any circle with radius R and center z_0 is the set of solutions to $|z - z_0| = R$. But

$$|z - z_0|^2 = (z - z_0)(\overline{z} - \overline{z_0}) = |z|^2 - \overline{z_0}z - z_0\overline{z} + |z_0|^2,$$

so we see that any circle is a solution to

$$|z|^2 - \overline{z_0}z - z_0\overline{z_0} + |z_0|^2 - R^2 = 0,$$

which is of the desired form, since

$$\det\begin{pmatrix} 1 & -\overline{z_0} \\ -z_0 & |z_0|^2 - R^2 \end{pmatrix} = -R^2 < 0.$$

Any line can be obtained as the set of solutions to $|z - z_0| = |z - z_1|$ for some $z_0 \neq z_1$. (See Exercise 2.2.5.) Equivalently,

$$\begin{aligned} |z - z_0|^2 &= |z - z_1|^2 \\ |z|^2 - \overline{z_0}z - z_0\overline{z} + |z_0|^2 &= |z|^2 - \overline{z_1}z - z_1\overline{z} + |z_1|^2 \\ (\overline{z_1} - \overline{z_0})\, z + (z_1 - z_0)\, \overline{z} + |z_0|^2 - |z_1|^2 &= 0, \end{aligned}$$

and as

$$\det\begin{pmatrix} 0 & \overline{z_1} - \overline{z_0} \\ z_1 - z_0 & |z_0|^2 - |z_1|^2 \end{pmatrix} = -\,|z_1 - z_0|^2 < 0,$$

we see that this is also an equation of the desired form. Conversely, for any equation

$$Az\overline{z} + Bz + \overline{B}z + C = 0,$$

if $A \neq 0$ we can divide by it to get a new equation

$$z\overline{z} + \frac{B}{A}z + \frac{\overline{B}}{A}z + \frac{C}{A} = 0.$$

This is the equation of a circle with center z_0 and radius R, where

$$z_0 = -\frac{\overline{B}}{A} \qquad\qquad R = \sqrt{\frac{B\overline{B}}{A^2} - \frac{C}{A}} = \frac{\sqrt{B\overline{B} - AC}}{|A|} > 0.$$

If $A = 0$, then it is the equation of a line $|z - z_0| = |z - z_1|$ where

$$z_1 - z_0 = \overline{B}$$
$$|z_0|^2 - |z_1|^2 = C.$$

Any such equation has solutions: for example, take

$$z_0 = \frac{|B|^2 + C}{2\Im(B)} i \qquad z_1 = \overline{B} + \frac{|B|^2 + C}{2\Im(B)} i.$$

(See Exercise 2.2.5.) We have thereby shown that any equation of the given form corresponds to either a circle or a line. □

We can now prove that $z \mapsto 1/z$ preserves generalized circles.

Lemma 2.3 *Let $\varphi(z) = 1/z$ and let γ be a generalized circle in $\mathbb{C}P^1$. Then $\varphi(\gamma)$ is a generalized circle in $\mathbb{C}P^1$.*

Proof By the algebraic description of generalized circles, we know that γ is the set

$$\left\{ z \in \mathbb{C}P^1 \middle| Az\overline{z} + Bz + \overline{B}\overline{z} + C = 0 \right\}$$

for some $A, C \in \mathbb{R}$ and $B \in \mathbb{C}$ such that $AC - B\overline{B} < 0$. The image of this set is the set

$$\left\{ \frac{1}{z} \in \mathbb{C}P^1 \middle| Az\overline{z} + Bz + \overline{B}\overline{z} + C = 0 \right\} = \left\{ z \in \mathbb{C}P^1 \middle| A\frac{1}{z\overline{z}} + B\frac{1}{z} + \overline{B}\frac{1}{\overline{z}} + C = 0 \right\}$$
$$= \left\{ z \in \mathbb{C}P^1 \middle| Cz\overline{z} + \overline{B}z + B\overline{z} + A = 0 \right\}.$$

Since $CA - \overline{B}B < 0$, we see that $\varphi(\gamma)$ is the set of solutions to a new equation describing a generalized circle. □

Of course, this result is merely a stepping stone to what we want to show: all linear fractional transformations preserve generalized circles.

Theorem 2.11 *Let $\varphi \in M\ddot{o}b^0(2)$ and let γ be a generalized circle in $\mathbb{C}P^1$. Then $\varphi(\gamma)$ is a generalized circle in $\mathbb{C}P^1$.*

Proof By the decomposition theorem for linear fractional transformations on $\mathbb{C}P^1$, we know that φ is a composition of translations, rotations, dilations, and $z \mapsto 1/z$. The first three preserve generalized circles by Theorem 1.6. By Lemma 2.3, we know that $z \mapsto 1/z$ does as well. Therefore, so does the composition of them all. □

This result is of enormous importance for a number of reasons. First, this basic property of linear fractional transformations will be most useful when we shall investigate how to give elegant proofs of various classical theorems in Euclidean geometry about lines and circles. Second, it will allow us to give simple definitions of things like angles and orientation without resorting to multivariable calculus or complex analysis.

▶ **Example** *What is the image of the curve described by the equation* $2z\bar{z} + (3+i)z + (3-i)\bar{z} + 3 = 0$ *under the map* $z \mapsto (z+i)?(iz)$?
Since the original curve was the set of solutions to

$$\left\{ z \in \mathbb{C}P^1 \middle| 2z\bar{z} + (3+i)z + (3-i)\bar{z} + 3 = 0 \right\},$$

the image will be the set

$$\begin{aligned} &\left\{ \varphi(z) \in \mathbb{C}P^1 \middle| 2z\bar{z} + (3+i)z + (3-i)\bar{z} + 3 = 0 \right\} \\ &= \left\{ z \in \mathbb{C}P^1 \middle| 2\varphi^{-1}(z)\overline{\varphi^{-1}(z)} + (3+i)\varphi^{-1}(z) + (3-i)\overline{\varphi^{-1}(z)} + 3 = 0 \right\}. \end{aligned}$$

It isn't hard to compute that

$$\varphi^{-1}(z) = \frac{i}{iz-1},$$

so the following equations are all equivalent.

$$\begin{aligned} 2\varphi^{-1}(z)\overline{\varphi^{-1}(z)} + (3+i)\varphi^{-1}(z) + (3-i)\overline{\varphi^{-1}(z)} + 3 &= 0 \\ 2\frac{i}{iz-1}\frac{-i}{-i\bar{z}-1} + 2\Re\left((3+i)\frac{i}{iz-1}\right) + 3 &= 0 \\ 3\left|iz-1\right|^2 + 2\Re\left((3+i)i(-i\bar{z}-1)\right) + 2 &= 0. \end{aligned}$$

If we write $z = x + iy$, the above can be expanded to

$$\begin{aligned} 3\left|iz-1\right|^2 + 2\Re\left((3+i)i(-i\bar{z}-1)\right) + 2 &= 3\left(x^2 + (y+1)^2\right) + 2(x - 3y - 3) + 2 \\ &= 3x^2 + 2x + 3y^2 - 1 = 0. \end{aligned}$$

To make further progress, we complete the square.

$$\begin{aligned} 3x^2 + 2x + 3y^2 - 1 &= 3\left(x^2 + \frac{2}{3}x + \frac{1}{9}\right) + 3y^2 - 1 - \frac{1}{3} \\ &= 3\left(x - \frac{1}{3}\right)^2 + 3y^2 - \frac{4}{3} = 0. \end{aligned}$$

Finally, dividing by 3 on both sides and rearranging,

$$\left(x - \frac{1}{3}\right)^2 + y^2 = \frac{4}{9}$$

which is the equation of a circle with center $(1/3, 0)$ and radius $2/3$.

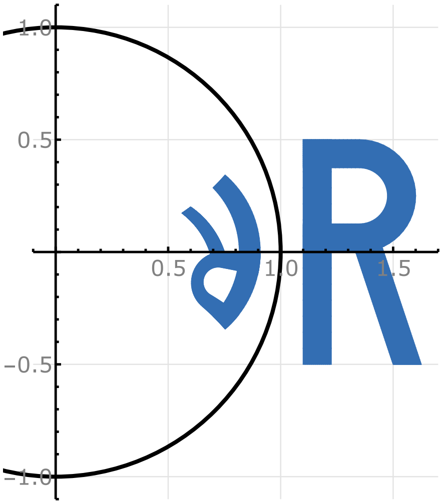


Fig. 2.10 An “R” and its image under the map $z \mapsto 1/z$. Note that while it is rotated and distorted, it is not flipped.

2.5 Oriented Circles

We defined orientation in Chapter 1 in a way that made use of the fact that similarities preserve circles. At that time, the concept of orientation made perfect intuitive sense: things that were orientation-reversing behaved like mirrors, while orientation-preserving transformations didn’t. But, of course, one can have a curved mirror, and that too should be “orientation-reversing” in some sense, as in Figure 2.10. On the other hand, we know that the map $z \mapsto 1/z$ is a composition of a reflection through a circle and a reflection through a line, and so it should be orientation-preserving.

We could introduce ideas from multivariable calculus in order to define a notion of orientation that would apply to all real differentiable maps. This would certainly include linear fractional transformations in particular. We proceed more simply: we shall take Definition 1.4—which applied to similarities—and alter it just enough for it to apply to maps that preserve generalized circles. Before we do that though, we will first have to define oriented circles.

Definition 2.10 An *oriented circle* C in $\mathbb{C}P^1$ is a generalized circle together with a direction in which it is traversed if considered as a path. We will write $-C$ for the generalized circle, but with the direction of travel reversed, and we shall say that this oriented circle has the *opposite orientation*. Any generalized circle splits $\mathbb{C}P^1$ into two connected regions; the region to the left of the path as it is traversed shall be termed the *interior*, written as $\text{Int}(C)$; the region to the right of the path shall be termed the *exterior*, written as $\text{Ext}(C)$. We therefore have the relations $\text{Int}(C) = \text{Ext}(-C)$ and $\text{Ext}(C) = \text{Int}(-C)$.

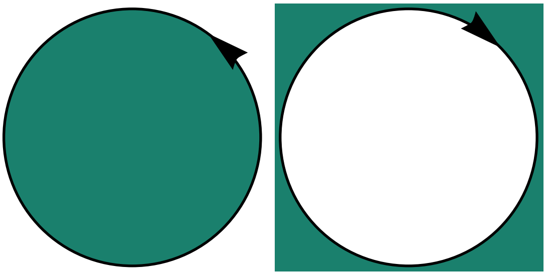


Fig. 2.11 Two oriented circles, with their interiors shaded.

The intuitive picture behind this definition is shown in Figure 2.11. While I generally strive for mathematical rigor, this is one case where I think a slightly informal—but deeply intuitive—definition is entirely justifiable. The reader who finds this to be unacceptably sloppy should bear in mind, though, that for circles and lines, we can fix particular kinds of paths, such as those of the form $t \mapsto Re^{\pm it}$ for circles, which allows us to unambiguously define what we mean by direction and what we mean by "to the left of".

Using the language of oriented circles, we can give a definition of orientation-preserving and orientation-reversing maps that will apply to transformations that preserve generalized circles.

Definition 2.11 Let $\varphi : \mathbb{C}P^1 \to \mathbb{C}P^1$ be a transformation that

1. maps oriented circles to oriented circles and
2. for any oriented circle C, the image of $\text{Int}(C)$ is either $\text{Int}(\varphi(C))$ or $\text{Ext}(\varphi(C))$.

We say that φ is *orientation-preserving* if for any oriented circle C, $\varphi(\text{Int}(C)) = \text{Int}(\varphi(C))$. We say that φ is *orientation-reversing* if for any oriented circle C, $\varphi(\text{Int}(C)) = \text{Ext}(\varphi(C))$.

Remark 2.9 The second restriction in Definition 2.11 may potentially feel a little artificial. It can be replaced with the following, simpler requirement: φ must be continuous with a continuous inverse.

While this definition is perhaps a little harder to digest than Definition 1.4, I claim that it is nothing more than a generalization. Indeed, this new definition reduces to the old one in the case where φ is a similarity. This will be easiest to prove using the following lemma.

Lemma 2.4 *Consider functions $\varphi_1, \varphi_2 : \mathbb{C}P^1 \to \mathbb{C}P^1$. All of the following are true.*

1. *If φ_1, φ_2 are orientation-preserving, then $\varphi_1 \circ \varphi_2$ is orientation-preserving.*
2. *If φ_1, φ_2 is orientation-reversing, then $\varphi_1 \circ \varphi_2$ is orientation-preserving.*
3. *If one of φ_1, φ_2 is orientation-preserving and the other is orientation-reversing, then $\varphi_1 \circ \varphi_2$ is orientation-reversing.*

Proof Choose any generalized circle C. If φ_1, φ_2 are both orientation-preserving,

$$(\varphi_1 \circ \varphi_2)\,(\mathrm{Int}(C)) = \varphi_1\,(\mathrm{Int}(\varphi_2(C))) = \mathrm{Int}\,((\varphi_1 \circ \varphi_2)(C))\,.$$

I leave the other cases as an exercise for the reader. (See Exercise 2.2.6.) □

Theorem 2.12 *Let $\varphi : \mathbb{C}P^1 \to \mathbb{C}P^1$ be a similarity (with the usual extension that $\varphi(\infty) = \infty$). Then φ is orientation-preserving/-reversing in the sense of Definition 2.11 if and only if it is orientation-preserving/-reversing in the sense of Definition 1.4.*

Proof By the decomposition theorem for complex affine maps (Theorem 1.1), the classification of orientation-preserving and orientation-reversing similarities (Theorem 1.7), and Lemma 2.4, we know that it suffices to prove this result for maps like $z \mapsto z + b$, $z \mapsto rz$, $z \mapsto e^{i\theta}z$, and $z \mapsto \overline{z}$. We already know that the first three are all orientation-preserving in the sense of Definition 1.4, and the fourth is orientation-reversing. We will be done if we show that the same is true using Definition 2.11. Choose any oriented circle C. There are three cases.

1. C is a circle with center z_0 and radius R, traversed counter-clockwise: its interior is the set of all points z such that $|z - z_0| < R$.
2. C is a circle with center z_0 and radius R, traversed clockwise: its interior is the set of all points z such that $|z - z_0| > R$.
3. C is a line passing through the point z_0 and traversed in the direction $e^{i\theta}$: its interior is the set of all points z such that $z = z_0 + e^{i\theta}t + -ie^{i\theta}s$ for some $t \in \mathbb{R}$ and some $s > 0$.

In the first case, $z \mapsto z + b, z \mapsto rz, z \mapsto e^{i\theta}z$ all move C to a circle traversed counter-clockwise, and the interior is the set of points $|z - \varphi(z)| < R$—therefore, they are all orientation-preserving, since this set is the image of the interior of C. However, $z \mapsto \overline{z}$ produces a circle traversed clockwise, and which therefore has interior $|z - \overline{z_0}| > R$. The image of $|z - z_0| < R$ under $z \mapsto z_0$ is $|z - \overline{z_0}| < R$ which is the exterior of the image of C—therefore, $z \mapsto \overline{z}$ is orientation-reversing. The other cases are similar, and so I leave them as an exercise to the reader. (See Exercise 2.2.7.) □

As we expect, all linear fractional transformations on $\mathbb{C}P^1$ are orientation-preserving. To prove this, we need an important lemma.

Lemma 2.5 *Under stereographic projection, the map $\varphi(z) = 1/z$ corresponds to a rotation of the unit sphere by π radians around the real axis. That is,*

$$\begin{aligned} \overline{stereo}^{-1} \circ \varphi \circ \overline{stereo} : S^2 &\to S^2 \\ (x, y, z) &\mapsto (x, -y, -z). \end{aligned}$$

Remark 2.10 This correspondence is likely easier to see via an illustration of a disk on the sphere being rotated, as in Figure 2.12.

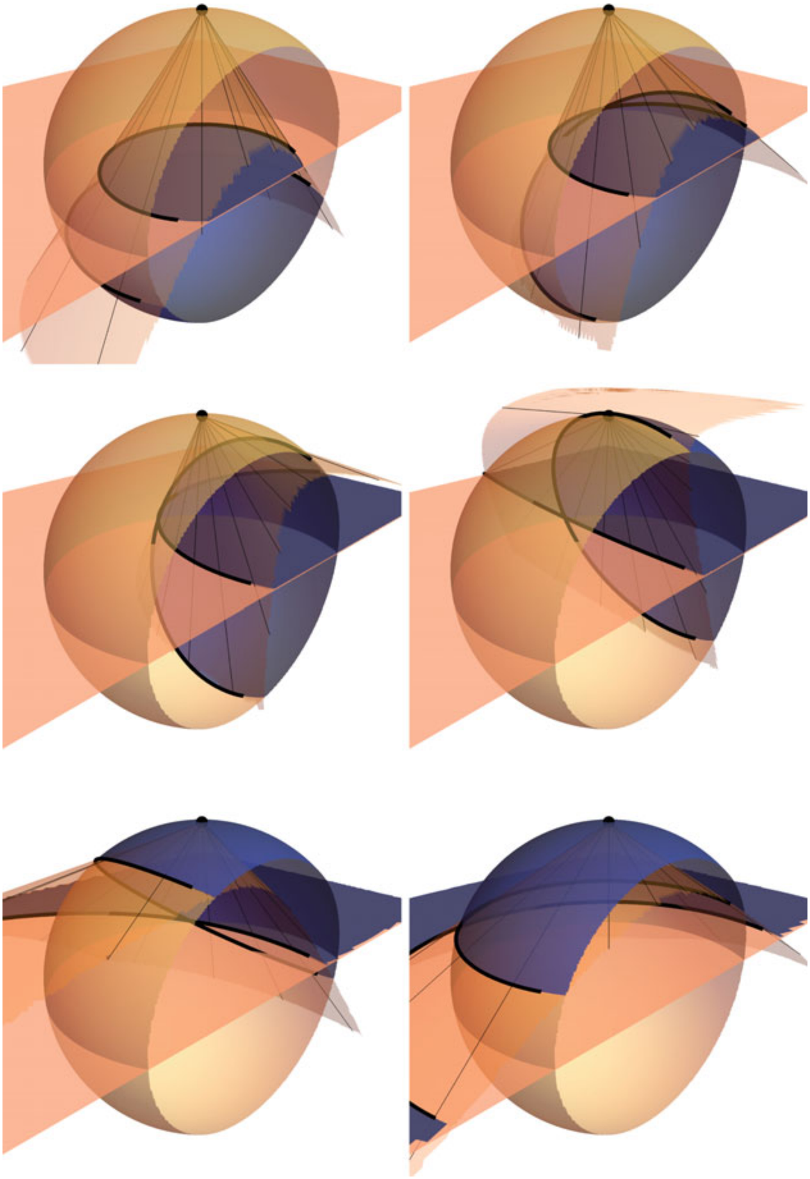


Fig. 2.12 The stereographic projection of an oriented circle onto the sphere, as the sphere is rotated. Everything has been sliced by a plane for easier viewing of the interior.

Proof This is equivalent to proving that $\varphi(\overline{\text{stereo}}(x, y, z)) = \overline{\text{stereo}}((x, -y, -z))$. Since $\overline{\text{stereo}}(0, 0, 1) = \infty$ and $\overline{\text{stereo}}(0, 0, -1) = 0$, if $z = 1$, then

$$\varphi(\overline{\text{stereo}}(x, y, z)) = \varphi(\infty) = 0 = \overline{\text{stereo}}((x, -y, -z));$$

similarly, if $z = -1$, then

$$\varphi(\overline{\text{stereo}}(x, y, z)) = \varphi(0) = \infty = \overline{\text{stereo}}((x, -y, -z)).$$

If $z \neq \pm 1$, then

$$\begin{aligned}\varphi(\overline{\text{stereo}}(x, y, z)) = \varphi\left(\frac{x + iy}{1 - z}\right) &= \frac{1 - z}{x + iy} = \frac{(1 - z)(x - iy)}{x^2 + y^2} \\ &= \frac{(1 - z)(x - iy)}{1 - z^2} = \frac{x - iy}{1 + z}.\end{aligned}$$

But this is the same as $\overline{\text{stereo}}((x, -y, -z))$. □

Theorem 2.13 *Let $\varphi \in Möb^0(2)$. Then φ is orientation-preserving.*

Proof In light of the decomposition theorem for linear fractional transformations on $\mathbb{C}P^1$, Lemma 2.4, and Theorem 2.12, it suffices to prove that $\varphi(z) = 1/z$ is orientation-preserving. This is most easily seen via Lemma 2.5—any oriented circle C in $\mathbb{C}P^1$ has an image which is some curve γ on S^2. If the interior of C is on the left of the curve, then the interior of γ will be on the right. Rotating the sphere will move γ to some new curve, but the interior will still remain on the right-hand side. But then after reversing the projection, we get a new oriented circle $\varphi(C)$ whose interior is on the left-hand side. □

▶ **Example** *Does there exist $\varphi \in$ Möb$^0(2)$ such that for some $t > 1$, $\varphi(0) = 0$, $\varphi(1) = 1$, $\varphi(\infty) = t$, and $\Im(\varphi(i)) < 0$?*
Suppose that there was such a φ. Consider $\mathbb{R} \cup \{\infty\}$—this is a generalized circle that passes through $0, 1, \infty$. Its image under φ must be a generalized circle passing through $0, 1, t$—which is to say that it must be $\mathbb{R} \cup \{\infty\}$ again. In fact, if we give $\mathbb{R} \cup \{\infty\}$ an orientation by saying that we must traverse it from left to right (i.e. from 0 to 1 to ∞), then its image under φ must also be traversed in that same direction (i.e. from 0 to 1 to t). But i is in the interior of $\mathbb{R} \cup \{\infty\}$ given this orientation, and $\varphi(i)$ is not! This is a clear contradiction, ergo there is no such $\varphi \in$ Möb$^0(2)$.

2.6 Angles, Revisited

We previously showed in Chapter 1 that all similarities preserve angles between lines. Illustrations like Figures 2.8 and 2.10 certainly suggest that in some sense, maps like $z \mapsto 1/z$ still preserve angles, even though they don't preserve lines in

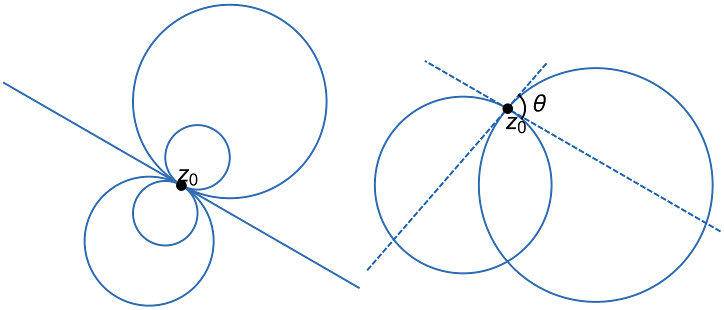


Fig. 2.13 On the left is a family of generalized circles all tangent at one point z_0. On the right, two generalized circles intersect at z_0 at an angle θ defined by their tangent lines.

general. We can make this precise as follows: let $\gamma_1, \gamma_2 : [0, 1] \to \mathbb{C}$ be two curves in the complex plane that intersect at some point z_0. We define the angle between γ_1, γ_2 at z_0 to be the angle θ between their tangent lines at z_0, as in Figure 2.13. We would like to say that a map φ is preserves the angle between γ_1 and γ_2 if the angle between the tangent lines of $\varphi \circ \gamma_1$ and $\varphi \circ \gamma_2$ at $\varphi(z_0)$ is also α. Of course, for this definition to make sense, you need to make sure that the images of γ_1 and γ_2 under φ are curves with well-defined tangent lines. We will eliminate this worry simply by always choosing our curves γ_1, γ_2 to be generalized circles.

There is another, more thorny issue with this definition: it forces the point of intersection to lie inside of $\mathbb{C}$ rather than $\mathbb{C}P^1$. This is inconvenient and against the general philosophy that ∞ is just like any other point in $\mathbb{C}P^1$. Thankfully, there is an obvious way to define angles at infinity by exploiting a property that is true of circles: they are either tangent or they intersect in two distinct points, in which case the angles of intersection are the same at both points. (See Exercises 2.2.8 and 2.2.9.)

Definition 2.12 Let C_1, C_2 be two generalized circles that intersect at ∞. If C_1 and C_2 are tangent, define the angle of intersection to be 0. Otherwise, there is another intersection point z_0—define the angle of intersection at ∞ to be the angle of intersection at z_0.

This definition is going to make everything wonderfully convenient for us—in particular, with this definition, we will be able to show that all elements in Möb$^0(2)$ are angle-preserving. To get to that point, though, we will need a collection of lemmas. We begin with the observation that if we choose any generalized circle C and a point z_0 on C then there will be an infinite family of generalized circles tangent to C at that point. Note that the definition of angle that we have chosen doesn't care about which circle in this family we select.

Lemma 2.6 *Let C_1, C_2, C_3 be generalized circles that all intersect at a point $z_0 \in \mathbb{C}P^1$. If C_1 is tangent to C_2, then the angle between C_1 and C_3 is the angle between C_2 and C_3.*

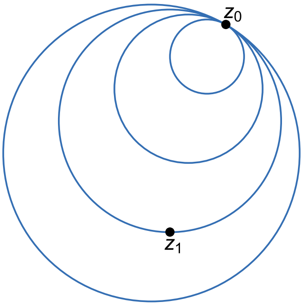


Fig. 2.14 A collection of circles tangent at z_0. One of them passes through a second point z_1.

Proof Let l_1, l_2, l_3 be the tangent lines to C_1, C_2, C_3 at z_0. Clearly, $l_1 = l_2$. But this means that in both cases, the angle is simply the angle between l_1 and l_3. □

That we can make this replacement makes life much easier because we can choose a member of a family of tangent circles that has convenient properties such as passing through a particular point.

Lemma 2.7 *Let C be a generalized circle and z_0 a point on C. For any $z_1 \in \mathbb{C}P^1$, there exists a generalized circle C_1 that is tangent to C at z_0 and which passes through z_1.*

Proof This statement is visually obvious, as illustrated in Figure 2.14. I leave the proof as an exercise for the reader. (See Exercise 2.2.10.) □

The last component that we need to proceed with a proof is to understand how linear fractional transformations interact with our chosen definition of angle.

Lemma 2.8 *Let C_1, C_2, C_3 be generalized circles that all intersect at a point $z_0 \in \mathbb{C}$. Let $\varphi \in Möb^0(2)$. If C_1 is tangent to C_2, then $\varphi(C_1)$ is tangent to $\varphi(C_2)$ and the angle between $\varphi(C_1)$ and $\varphi(C_3)$ is the angle between $\varphi(C_2)$ and $\varphi(C_3)$.*

Proof We know that $\varphi(C_1), \varphi(C_2), \varphi(C_3)$ will be generalized circles intersecting at $\varphi(z_0)$. If $C_1 = C_2$ then $\varphi(C_1) = \varphi(C_2)$; alternatively, if the intersection between C_1 and C_2 is just z_0, then the intersection between $\varphi(C_1)$ and $\varphi(C_2)$ must be $\varphi(z_0)$ since φ is bijective. Therefore, in either case, $\varphi(C_1)$ is tangent to $\varphi(C_2)$ at $\varphi(z_0)$. By Lemma 2.6, we know that the angle between $\varphi(C_1)$ and $\varphi(C_3)$ is the angle between $\varphi(C_2)$ and $\varphi(C_3)$. □

With this in mind, it will now be comparatively easy to prove that elements of Möb^0(2) are angle-preserving.

Theorem 2.14 *Let $\varphi \in Möb^0(2)$. For any two generalized circles C_1, C_2 that intersect at a point $z_0 \in \mathbb{C}P^1$, the angle of intersection of C_1 and C_2 is equal to the angle of intersection of $\varphi(C_1)$ and $\varphi(C_2)$.*

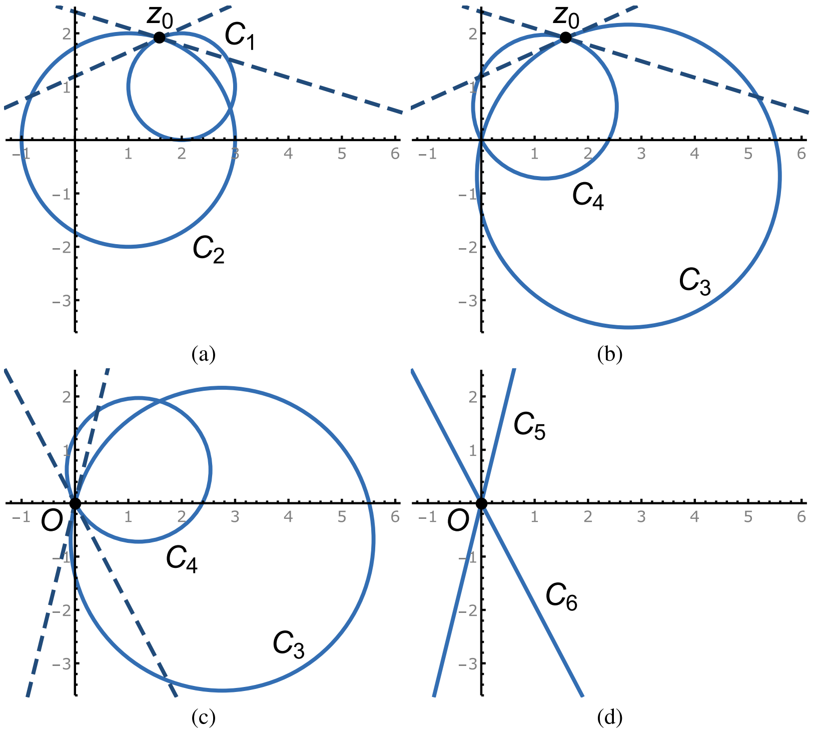


Fig. 2.15 A visual sketch of the proof of Theorem 2.14. (a) shows the initial configuration with C_1, C_2 intersecting z_0; in (b), we exchange both circles with tangent ones C_3, C_4 that pass through the origin O; in (c) and (d), we replace those circles with their tangent lines C_5, C_6.

Proof It is easy to see that if this is true for $\varphi_1, \varphi_2 \in \text{Möb}^0(2)$ then it must be true for $\varphi_1 \circ \varphi_2$. Since we already know that translations, rotations, and dilations are all angle-preserving, it shall suffice to prove that $z \mapsto 1/z$ is angle-preserving. Since we know that $z \mapsto \overline{z}$ is angle-preserving, it will suffice to prove that $z \mapsto 1/\overline{z}$, the reflection through the unit circle, is angle-preserving.

Choose any two generalized circles C_1, C_2 with a common point of intersection z_0. By Lemmas 2.6 and 2.7, we can choose some generalized circles C_3, C_4 such that

1. C_3, C_4 are tangent at z_0 to C_1 and C_2, respectively;
2. C_3 and C_4 pass through 0; and
3. the angle of intersection between C_1, C_2 is the same as the angle of intersection between C_3 and C_4.

I refer the reader to Figure 2.15 for a diagram of this new configuration. By Lemma 2.8, we know that if the angle of intersection between C_3 and C_4 is preserved, then it is preserved between C_1 and C_2. On the other hand, we know that 0 is another common intersection point of C_1 and C_2, and therefore the angle of intersection at that point must also be the same. Furthermore, if that angle of intersection is

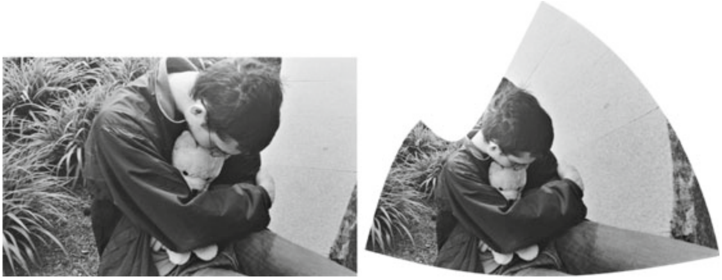


Fig. 2.16 An angle-preserving transformation that does not preserve generalized circles.

preserved, then so is the angle at z_0. The final reduction is as follows: we can replace C_3 and C_4 with two generalized circles C_5 and C_6 that are tangent to them at 0 and which both pass through ∞—that is, C_5 and C_6 are both lines passing through 0. Of course, the angle between these two lines must be the same as the angle between C_5 and C_6, and if this angle is preserved by $z \mapsto 1/\overline{z}$, then the original angle is also preserved. However, it is easy to see that the reflection through the unit circle simply fixes lines through the origin and so this final angle is preserved. □

Remark 2.11 This approach does not generalize to discussing angles of intersection between arbitrary curves. Linear fractional transformations preserve those too, but proving it requires some knowledge of the derivatives of maps $\mathbb{R}^n \to \mathbb{R}^k$ or, even better, complex analysis. For the latter approach, I refer the interested reader to Needham's *Visual Complex Analysis* [11].

While the fact that elements of $\text{Möb}^0(2)$ are angle-preserving is a special property, the reader should not be mistaken in thinking that these are the only types of transformations that have this property—there is a very large family of functions of interest in complex analysis that all possess this same quality. An example of such a function is depicted in Figure 2.16.

2.7 The Cross-Ratio

When we introduced maps $z \mapsto az + b$, it was in the context of transformations that preserve distances or ratios of distances. A reasonable question to ask is whether there is some similar type of quantity that is preserved by linear fractional transformations. And, indeed, there is! To motivate the definition, let us recall that we defined a similarity as being a map Φ such that for any triple of points z_1, z_2, z_3,

$$\frac{|z_1 - z_2|}{|z_1 - z_3|} = \frac{|\Phi(z_1) - \Phi(z_2)|}{|\Phi(z_1) - \Phi(z_3)|}.$$

However, we could also write

$$\frac{|z_1 - z_2|^2}{|z_1 - z_3|^2} = \frac{(z_1 - z_2)(\overline{z_1} - \overline{z_2})}{(z_1 - z_3)(\overline{z_1} - \overline{z_3})} = \left(\frac{z_1 - z_2}{z_1 - z_3}\right)\overline{\left(\frac{z_1 - z_2}{z_1 - z_3}\right)}$$

and with this inspiration we might notice that actually for any triple of points z_1, z_2, z_3 it will be true that

$$\frac{z_1 - z_2}{z_1 - z_3} = \frac{\Phi(z_1) - \Phi(z_2)}{\Phi(z_1) - \Phi(z_3)}$$

for any orientation-preserving similarity Φ, and

$$\frac{z_1 - z_2}{z_1 - z_3} = \overline{\left(\frac{\Phi(z_1) - \Phi(z_2)}{\Phi(z_1) - \Phi(z_3)}\right)}$$

for any orientation-reversing similarity Φ. It is easy enough to check that this is not always true for elements of $\text{Möb}^0(2)$, but there is an easy generalization that does work.

Definition 2.13 For any four distinct points $z_1, z_2, z_3, z_4 \in \mathbb{C}P^1$, their *cross-ratio* is defined to be the complex number

$$[z_1, z_2; z_3, z_4] = \begin{cases} \frac{z_2 - z_4}{z_2 - z_3} & \text{if } z_1 = \infty \\ \frac{z_1 - z_3}{z_1 - z_4} & \text{if } z_2 = \infty \\ \frac{z_2 - z_4}{z_1 - z_4} & \text{if } z_3 = \infty \\ \frac{z_1 - z_3}{z_2 - z_3} & \text{if } z_4 = \infty \\ \frac{(z_1 - z_3)(z_2 - z_4)}{(z_2 - z_3)(z_1 - z_4)} & \text{otherwise.} \end{cases}$$

Remark 2.12 If the piecewise definition is unappealing, one could always define this in terms of a limit as

$$[z_1, z_2; z_3, z_4] = \lim_{(w_1, w_2, w_3, w_4) \to (z_1, z_2, z_3, z_4)} \frac{(w_1 - w_3)(w_2 - w_4)}{(w_2 - w_3)(w_1 - w_4)}.$$

Alternatively, one way to remember what the cross-ratio is is to write down

$$\frac{(z_1 - z_3)(z_2 - z_4)}{(z_2 - z_3)(z_1 - z_4)}$$

but if any of z_1, z_2, z_3, z_4 is ∞, simply remove the factors in the numerator and denominator that contain it.

Theorem 2.15 *Elements of $\text{Möb}^0(2)$ preserve the cross-ratio, in the sense that if z_1, z_2, z_3, z_4 are four distinct points in $\mathbb{C}P^1$ and $\varphi \in \text{Möb}^0(2)$, then*

$$[z_1, z_2; z_3, z_4] = [\varphi(z_1), \varphi(z_2); \varphi(z_3), \varphi(z_4)] .$$

Proof First, we note that all orientation-preserving similarities preserve the cross-ratio. This is because

$$\frac{(z_1 - z_3)(z_2 - z_4)}{(z_2 - z_3)(z_1 - z_4)} = \frac{z_1 - z_3}{z_2 - z_3} \cdot \frac{z_2 - z_4}{z_1 - z_4},$$

and we already saw that orientation-preserving similarities preserve expressions of these forms. Thus, it will suffice to prove that $\varphi(z) = 1/z$ preserves the cross-ratio. This is an easy computation:

$$\begin{aligned}[\varphi(z_1), \varphi(z_2); \varphi(z_3), \varphi(z_4)] &= \frac{(1/z_1 - 1/z_3)(1/z_2 - 1/z_4)}{(1/z_2 - 1/z_3)(1/z_1 - 1/z_4)} \\ &= \frac{(z_3 - z_1)(z_4 - z_2)}{(z_3 - z_2)(z_4 - z_1)} \\ &= \frac{(z_1 - z_3)(z_2 - z_4)}{(z_2 - z_3)(z_1 - z_4)} \\ &= [z_1, z_2; z_3, z_4].\end{aligned}$$

Technically, this computation is correct on the nose only if none of z_1, z_2, z_3, z_4 are either 0 or ∞. However, since we can define both φ and the cross-ratio in terms of a limit, this is irrelevant. □

This proof has a string of important corollaries.

Corollary 2.1 *For any triple of distinct points $z_1, z_2, z_3 \in \mathbb{C}P^1$, there is exactly one $\varphi \in$ Möb$^0(2)$ such that $\varphi(z_1) = 0$, $\varphi(z_2) = 1$, $\varphi(z_3) = \infty$. Specifically,*

$$\varphi(z) = \frac{z_2 - z_3}{z_2 - z_1} \cdot \frac{z - z_1}{z - z_3}.$$

Proof Suppose there is such a transformation φ. Choose any point $z \neq z_1, z_2, z_3$. By Theorem 2.15, we know that

$$[\varphi(z), 1; 0, \infty] = [\varphi(z), \varphi(z_2); \varphi(z_1), \varphi(z_3)] = [z, z_2; z_1, z_3].$$

However,

$$[\varphi(z), 1; 0, \infty] = \frac{\varphi(z) - 0}{1 - 0} = \varphi(z)$$

and

$$[z, z_2; z_1, z_3] = \frac{(z - z_1)(z_2 - z_3)}{(z_2 - z_1)(z - z_3)} = \frac{z_2 - z_3}{z_2 - z_1} \cdot \frac{z - z_1}{z - z_3},$$

and therefore

$$\varphi(z) = \frac{z_2 - z_3}{z_2 - z_1} \cdot \frac{z - z_1}{z - z_3}.$$

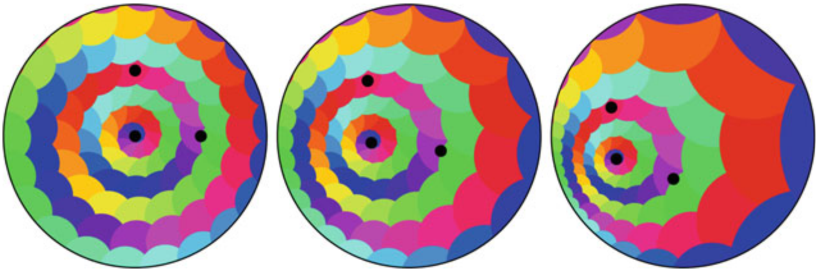


Fig. 2.17 An illustration of how linear fractional transformations allow us to take a triple of points and move it to any other triple.

It is easy to check that this is an element of Möb$^0(2)$ with the desired properties. □

Corollary 2.2 *For any triple of distinct points $w_1, w_2, w_3 \in \mathbb{C}P^1$, there is exactly one $\varphi \in$ Möb$^0(2)$ such that $\varphi(0) = w_1$, $\varphi(1) = w_2$, $\varphi(\infty) = w_3$. Specifically,*

$$\varphi(z) = \frac{(w_1 - w_2)w_3 z + (w_2 - w_3)w_1}{(w_1 - w_2)z + w_2 - w_3}.$$

Proof If $\varphi(0) = w_1, \varphi(1) = w_2$, and $\varphi(\infty) = w_3$,then $\varphi^{-1}(w_1) = 0, \varphi^{-1}(w_2) = 1$, and $\varphi^{-1}(w_3) = \infty$. We know that there is exactly one transformation with that property, so

$$\varphi^{-1}(z) = \frac{w_2 - w_3}{w_2 - w_1} \cdot \frac{z - w_1}{z - w_3}.$$

Using the techniques we developed for computing inverses of linear fractional transformations, it is not too difficult to show that

$$\varphi(z) = \frac{(w_1 - w_2)w_3 z + (w_2 - w_3)w_1}{(w_1 - w_2)z + w_2 - w_3}$$

as was claimed. □

Corollary 2.3 *For any pair of distinct triples of points z_1, z_2, z_3 and w_1, w_2, w_3 in $\mathbb{C}P^1$, there is exactly one $\varphi \in$ Möb$^0(2)$ with the property that $\varphi(z_i) = w_i$ for $i = 1, 2, 3$.*

Remark 2.13 An example of linear fractional transformations moving triples of points is illustrated in Figure 2.17.

Proof Choose the unique elements $\varphi_1, \varphi_2 \in$ Möb$^0(2)$ with the properties that $\varphi_1(z_1) = 0$, $\varphi_1(z_2) = 1$, $\varphi_1(z_3) = \infty$ and $\varphi_2(0) = w_1$, $\varphi_2(1) = w_2$, $\varphi_2(\infty) = w_3$. Then $\varphi = \varphi_2 \circ \varphi_1$ has the property that $\varphi(z_i) = w_i$ for $i = 1, 2, 3$. Now, suppose that $\tilde{\varphi} \in$ Möb$^0(2)$ satisfies $\tilde{\varphi}(z_i) = w_i$ for $i = 1, 2, 3$. Define $\tilde{\varphi}_1 = \varphi_2^{-1} \circ \tilde{\varphi}$. Then $\tilde{\varphi}_1(z_1) = 0$, $\tilde{\varphi}_1(z_2) = 1$, and $\tilde{\varphi}_1(z_3) = \infty$, so $\tilde{\varphi}_1 = \varphi_1$. Therefore, $\tilde{\varphi} = \varphi_2 \circ \varphi_1 = \varphi$. □

Corollary 2.4 *Let $\Phi : \mathbb{C}P^1 \to \mathbb{C}P^1$ be any map that preserves the cross-ratio. Then $\Phi \in Möb^0(2)$.*

Proof First, we note that it must be that Φ is injective—otherwise

$$[\Phi(z_1), \Phi(z_2); \Phi(z_3), \Phi(z_4)]$$

won't even be defined in general. Thus, $w_1 = \Phi(0)$, $w_2 = \Phi(1)$, and $w_3 = \Phi(\infty)$ are distinct points. Therefore, there is some $\varphi \in \text{Möb}^0(2)$ such that $\varphi(w_1) = 0$, $\varphi(w_2) = 1$, and $\varphi(w_3) = \infty$. This implies that $\tilde{\Phi} = \varphi \circ \Phi$ is a transformation that preserves the cross-ratio and has the property that $\tilde{\Phi}(z) = z$ for $z = 0, 1, \infty$. Choose any $z \in \mathbb{C}P^1$ other than $0, 1, \infty$, and note that it must be true that

$$z = [z, 1; 0, \infty] = [\tilde{\Phi}(z), 1; 0, \infty] = \tilde{\Phi}(z).$$

This implies that $\Phi = \varphi^{-1} \in \text{Möb}^0(2)$, as desired. □

In short, linear fractional transformations on $\mathbb{C}P^1$ are exactly the transformations on $\mathbb{C}P^1$ that preserve the cross-ratio; furthermore, they give exactly enough freedom to move any three distinct points to any other set of three distinct points. The first statement gives us a broad philosophical idea of why linear fractional transformations should be important or natural; we will see in the next chapter that the second statement is extremely useful for writing proofs about Euclidean geometry.

► **Example** *Show there exist $z_1, z_2, z_3, z_4 \in \mathbb{C}$ such that $[z_1, z_2; z_3, z_4] = \lambda$ if and only if $\lambda \neq 0, 1, \infty$.*
Since linear fractional transformations preserve the cross-ratio and can move any triple of points to any other triple, we may assume without loss of generality that $z_2 = 1$, $z_3 = 0$, $z_4 = \infty$, hence the condition is satisfied if and only if

$$\lambda = [z_1, z_2; z_3, z_4] = \frac{z_1 - 0}{1 - 0} = z_1,$$

for some $z_1 \neq 0, 1, \infty$.

2.8 The Group of Möbius Transformations

Before we end this chapter, I want to finally explain why we have been using the notation $\text{Möb}^0(2)$ to stand for the linear fractional transformations on $\mathbb{C}P^1$. The notation is reminiscent of our notation for similarities from Chapter 1, and so the reader might correctly guess that $\text{Möb}^0(2)$ is the collection of orientation-preserving transformations of some larger group. This larger group is the collection of Möbius transformations.

Definition 2.14 The *group of Möbius transformations* in $\mathbb{C}P^1$, denoted by Möb(2), is the collection of all transformations $\Phi : \mathbb{C}P^1 \to \mathbb{C}P^1$ such that Φ can be written as compositions of circle and line inversions.

This definition departs slightly from some common conventions in the literature. Most often, the term "Möbius transformation" is used to denote what I have termed "linear fractional transformation on $\mathbb{C}P^1$" (although that term is also used). However, it is possible to talk about, say, inversions through a sphere, or even an n-sphere, and to consider the group of transformations that can be obtained as compositions of such inversions. This is relevant for higher-dimensional hyperbolic geometry, for example. In that context, that group is usually called the group of Möbius transformations. Of course, that larger group contains elements that are not orientation-preserving, and so this conflicts with the somewhat more traditional usage. For convenience, I have opted to call the group obtained by allowing compositions of n-sphere inversions Möb(n), or the group of Möbius transformations in n-dimensional space. Of course, we should check that this really is a group.

Theorem 2.16 *With function composition as the operation, Möb*(2) *is a group.*

Proof We know that function composition is associative. Furthermore, we know that the identity function ι has the properties of a group identity. We only need to check that compositions of elements in Möb(2) are still elements in Möb(2) and that they have inverses. Both assertions are easy to verify—for any $\Phi_1, \Phi_2 \in$ Möb(2), write

$$\Phi_1 = \varphi_1 \circ \varphi_2 \circ \ldots \varphi_m$$
$$\Phi_2 = \psi_1 \circ \psi_2 \circ \ldots \psi_n,$$

where all of the φ_i's and ψ_j's are inversions. Then

$$\Phi_1 \circ \Phi_2 = \varphi_1 \circ \ldots \varphi_m \circ \psi_1 \circ \ldots \psi_n \in \text{Möb}(2)$$

and

$$\begin{aligned}\Phi_1^{-1} &= (\varphi_1 \circ \varphi_2 \circ \ldots \varphi_m)^{-1} \\ &= \varphi_m^{-1} \circ \ldots \varphi_2^{-1} \circ \varphi_1^{-1} \in \text{Möb}(2).\end{aligned}$$

This concludes the proof. □

We can now justify using the notation Möb$^0(2)$ to denote the collection of linear fractional transformations on $\mathbb{C}P^1$.

Theorem 2.17 *The set Möb*(2) *can be partitioned into a subset of orientation-preserving transformations and a subset of orientation-reversing transformations. The set Möb*$^0(2)$ *is precisely the set of orientation-preserving Möbius transformations. Any orientation-reversing element can be uniquely written as* $\varphi \circ$ *conj for some* $\varphi \in$ *Möb*$^0(2)$*, where*

$$\begin{aligned} conj : \mathbb{C}P^1 &\to \mathbb{C}P^1 \\ z &\mapsto \overline{z}. \end{aligned}$$

Proof Since any circle or line reflection is orientation-reversing, any element $\Phi \in \text{Möb}(2)$ is orientation-reversing if it is a composition of an odd number of reflections, and orientation-preserving otherwise. Next, we'll show that $\text{Möb}^0(2)$ sits inside of $\text{Möb}(2)$. It shall suffice to prove this for translations, rotations, dilations, and the map $z \mapsto \overline{z}$. By Theorem 1.9, we know that every isometry can be written as a composition of line reflections—in particular, this applies to $z \mapsto z+b$ and $z \mapsto e^{i\theta}z$. We already saw that $z \mapsto \overline{z}$ is a composition of a circle reflection and a line reflection, so this leaves the dilations. For any $\lambda > 0$, consider the inversions through the circles $|z| = \sqrt{\lambda}$ and $|z| = 1$—call these φ_1 and φ_2 respectively. Note that

$$(\varphi_1 \circ \varphi_2)(0) = 0 \qquad\qquad (\varphi_1 \circ \varphi_2)(\infty) = \infty$$

and for any $z = re^{i\theta}$ with $r > 0$,

$$(\varphi_1 \circ \varphi_2)(re^{i\theta}) = \varphi_1\left(\frac{1}{r}e^{i\theta}\right) = \lambda re^{i\theta},$$

whence we have $(\varphi_1 \circ \varphi_2)(z) = \lambda z$. It remains to show that all orientation-preserving maps in Möb(2) are linear fractional transformations. By Theorem 2.9, we know that every circle reflection is a composition of a linear fractional transformation and $z \mapsto \overline{z}$; we know from Chapter 1 that all line reflections are compositions of a linear fractional transformation and $z \mapsto \overline{z}$ as well. Notice that if

$$\varphi(z) = \frac{az+b}{cz+d}$$

then

$$\begin{aligned} \varphi(\overline{z}) &= \frac{a\overline{z}+b}{c\overline{z}+d} \\ \overline{\varphi(z)} &= \frac{\overline{az}+\overline{b}}{\overline{cz}+\overline{d}}. \end{aligned}$$

This means that if we define (by slight abuse of notation)

$$\begin{aligned} &\text{conj} : \text{Möb}^0(2) \to \text{Möb}^0(2) \\ &\left(z \mapsto \frac{az+b}{cz+d}\right) \mapsto \left(z \mapsto \frac{\overline{a}z+\overline{b}}{\overline{c}z+\overline{d}}\right) \end{aligned}$$

then for any $\varphi \in \text{Möb}^0(2)$,

$$\text{conj} \circ \varphi = \text{conj}(\varphi) \circ \text{conj}.$$

Thus, if we write

$$\Phi = \varphi_1 \circ \mathrm{conj} \circ \varphi_2 \circ \mathrm{conj} \circ \dots \varphi_n \circ \mathrm{conj}$$

for some $\varphi_1, \varphi_2, \dots \varphi_n \in \mathrm{Möb}^0(2)$, then we can rewrite it as

$$\begin{aligned}\Phi &= \varphi_1 \circ \mathrm{conj}(\varphi_2) \circ \mathrm{conj} \circ \mathrm{conj} \circ \varphi_3 \circ \dots \varphi_n \circ \mathrm{conj}\\ &= \begin{cases} \varphi_1 \circ \mathrm{conj}(\varphi_2) \circ \varphi_3 \circ \dots \varphi_n \circ \mathrm{conj} & \text{if } n \text{ is odd}\\ \varphi_1 \circ \mathrm{conj}(\varphi_2) \circ \varphi_3 \circ \dots \mathrm{conj}(\varphi_n) & \text{if } n \text{ is even.}\end{cases}\end{aligned}$$

Therefore, every orientation-preserving map in Möb(2) is a linear fractional transformation, and any orientation-reversing map is a composition of a linear fractional transformation and conj. □

This result allows us to characterize Möb(2) in a different way.

Corollary 2.5 *The set of Möbius transformations on $\mathbb{C}P^1$ is exactly the set of maps $\Phi : \mathbb{C}P^1 \to \mathbb{C}P^1$ that either preserve the cross-ratio or conj $\circ\, \Phi$ preserves the cross-ratio.*

Proof This is an immediate consequence of Corollary 2.4 and Theorem 2.17. □

To sum up, we know that if $\Phi \in \mathrm{Möb}(2)$, then for any distinct quadruple of points $z_1, z_2, z_3, z_4 \in \mathbb{C}P^1$,

$$[\Phi(z_1), \Phi(z_2); \Phi(z_3), \Phi(z_4)] = \begin{cases} [z_1, z_2; z_3, z_4] & \text{if } \Phi \in \mathrm{Möb}^0(2)\\ \overline{[z_1, z_2; z_3, z_4]} & \text{otherwise.}\end{cases}$$

This has the obvious corollary that any element of Möb(2) will preserve the quantity

$$|[z_1, z_2; z_3, z_4]| = \frac{|z_1 - z_3||z_2 - z_4|}{|z_2 - z_3||z_1 - z_4|} = \frac{d_{\mathrm{Euclid}}(z_1, z_3) d_{\mathrm{Euclid}}(z_2, z_4)}{d_{\mathrm{Euclid}}(z_2, z_3) d_{\mathrm{Euclid}}(z_1, z_4)},$$

which is also sometimes referred to as the cross-ratio. One may well ask whether Möbius transformations are the only kind of transformations that preserve this quantity, and indeed they are.

Theorem 2.18 (Cross-Ratio Characterization of the Möbius Group) *The set of Möbius transformations on $\mathbb{C}P^1$ is exactly the set of transformations $\Phi : \mathbb{C}P^1 \to \mathbb{C}P^1$ such that for all distinct quadruples of points z_1, z_2, z_3, z_4,*

$$\frac{d_{Euclid}(z_1, z_3) d_{Euclid}(z_2, z_4)}{d_{Euclid}(z_2, z_3) d_{Euclid}(z_1, z_4)} = \frac{d_{Euclid}(\Phi(z_1), \Phi(z_3)) d_{Euclid}(\Phi(z_2), \Phi(z_4))}{d_{Euclid}(\Phi(z_2), \Phi(z_3)) d_{Euclid}(\Phi(z_1), \Phi(z_4))}.$$

Proof We just showed that all Möbius transformations have this property. On the other hand, let Φ be a transformation that preserves this cross-ratio. By composing with Möbius transformations if necessary, we can assume without loss of generality that $\Phi(0) = 0$, $\Phi(1) = 1$, and $\Phi(\infty) = \infty$. But this means that for any $z \in \mathbb{C}P^1\backslash\{0, 1\infty\}$,

$$|z| = |[z, 1; 0, \infty]| = |[\Phi(z), 1; 0, \infty]| = |\Phi(z)|$$
$$|z - 1| = |[z, 0; 1, \infty]| = |[\Phi(z), 0; 1, \infty]| = |\Phi(z) - 1|.$$

From this, it is easy to deduce that for any such z, either $\Phi(z) = z$ or $\Phi(z) = \overline{z}$. Therefore, by composing with conj if necessary, we can assume that $\Phi(i) = i$. But this means there is an additional restriction

$$|z - i| = |[z, 0; i, \infty]| = |[\Phi(z), 0; i, \infty]| = |\Phi(z) - i|.$$

This forces $\Phi(z) = z$ for all $z \in \mathbb{C}P^1$, and we see that $\Phi \in$ Möb(2). □

While I titled this chapter "Inversive Geometry," I never explained what this actually is. With these final theorems, however, one can give a fairly straightforward definition: inversive geometry is the study of what is preserved by circle reflections or, equivalently, the types of transformations that preserve the cross-ratio.

Problems

2.1 COMPUTATIONAL EXERCISES

1. For each of the following, compute the image under $\varphi : \mathbb{C}P^1 \to \mathbb{C}P^1$.

   a) Line $y = x$, $\varphi(z) = \frac{z}{1-z}$.
   b) Line $x = 4$, $\varphi(z) = \frac{(-3+i)\bar{z}-1-2i}{iz-1}$.
   c) Circle $|z| = 1$, $\varphi(z) = \frac{(1+2i)z+1-2i}{i-(1-i)z}$.
   d) Circle $|z-3|^2 = 2$, $\varphi(z) = \frac{(1+2i)\bar{z}-1+4i}{i\bar{z}+1-i}$.

2. a) Find the angle between the line $y = x+4$ and the circle $|z-2+3/2i|^2 = 1/2$.
   b) Find the images of the above-mentioned line and circle under the map

   $$z \mapsto -\frac{iz+2+2i}{(1+i)z+4+i}.$$

   What is the angle between the images? Does it match your result from the previous part?

3. For each of the following pairs of triples (z_1, z_2, z_3), (w_1, w_2, w_3), find an element $\varphi \in \text{Möb}^0(2)$ such that $\varphi(z_i) = w_i$ for $i = 1, 2, 3$.

   a) $(0, 1+i, 1-i)$, $(0, 1, \infty)$.
   b) $(0, 1, \infty)$, $(i, -i, 1)$.
   c) $(0, 1+i, 1-i)$, $(i, -i, 1)$.

2.2 PROOFS

1. Prove that if $a, b, c, d \in \mathbb{C}$ and $ad - bc = 0$, then $(az+b)/(cz+d)$ is either undefined, or is some constant that does not depend on z.
2. Prove that if $a, b, c, d \in \mathbb{C}$, $ad - bc \neq 0$, $\varphi(z) = (az+b)/(cz+d)$, and $\varphi^{-1}(z) = (dz-b)/(-cz+a)$, then $\varphi(\varphi^{-1}(z)) = \varphi^{-1}(\varphi(z)) = z$ for all $z \in \mathbb{C}P^1$.
3. Finish the proof of Theorem 2.1.
4. Prove Theorem 2.4. *(Hint: Study carefully the proofs of Lemma 2.1 and Theorem 2.3.)*
5. a) Prove that $\mathbb{R}$ is the set of complex points z such that $|z-i| = |z+i|$.
   b) Prove that any line l in $\mathbb{C}$ can be obtained as the set of solutions to $|z-z_0| = |z-z_1|$ for some complex numbers $z_0 \neq z_1$. *(Hint: You may want to use an isometry to reduce to the case $l = \mathbb{R}$.)*

c) Prove that if

$$z_0 = \frac{|B|^2 + C}{2\Im(B)} i \qquad\qquad z_1 = \overline{B} + \frac{|B|^2 + C}{2\Im(B)} i,$$

then the line $|z - z_0| = |z - z_1|$ is the set of solutions to the equation $Bz + \overline{B}z + C = 0$.

6. Finish the proof of Lemma 2.4.
7. Finish the proof of Theorem 2.12.
8. Our goal is to prove that any two generalized circles intersect at either no, one, two, or all points.

   a) Prove that for any two lines L_1, L_2, if they share two points of intersection, then $L_1 = L_2$.
   b) Prove that for any generalized circles C_1, C_2, if they share three points of intersection, then $C_1 = C_2$. *(Hint: Let z be one of the points of intersection. Consider taking the image under a circle inversion centered at z. What will the image of C_1 and C_2 be?)*
   c) Prove that for any two generalized circles C_1, C_2, if C_1 and C_2 are not tangent, then either they don't intersect, or they intersect in two points.

9. a) Let C_1 be the unit circle, and let C_2 be a circle with center $z > 0$ which intersects C_1 in exactly two points. Prove that the angles of intersection at these two points are the same. *(Hint: What is the effect on C_1 and C_2 if we take the image under the map $z \mapsto \overline{z}$? What does it do to the two points of intersection?)*
   b) Let C_1, C_2 be two circles that intersect in exactly two points. Prove that the angles of intersection at these two points are the same. *Hint: You may want to use a similarity to reduce to the case where C_1 is the unit circle and the center of C_2 lies on the positive x-axis.*
   c) Let C_1 be the unit circle, and let C_2 be a line $x = x_0$ for some $x_0 > 0$ which intersects C_1 in exactly two points. Prove that the angles of intersection at these two points are the same. *(Hint: What is the effect on C_1 and C_2 if we take the image under the map $z \mapsto \overline{z}$? What does it do to the two points of intersection?)*
   d) Let C_1 be a circle and let C_2 be a line that intersect in exactly two points. Prove that the angles of intersection at these two points are the same. *(Hint: You may want to use a similarity to reduce to the case where C_1 is the unit circle and C_2 is a vertical line.)*
   e) Let C_1, C_2 be two generalized circles that intersect at exactly two points $z_1, z_2 \in \mathbb{C}$. Prove that the angles of intersection at these two points are the same.

10. a) Let L be a line and let $z_0 = \infty$. For any $z_1 \in \mathbb{C}$, prove that there exists a line L' that is tangent to L at z_0 and which passes through z_1.
    b) Prove that for any two distinct points $z_0, z_1 \in \mathbb{C}$, there exists a circle centered at z_0 that goes through z_1.
    c) Let C be a generalized circle and let z_0 be a point on C. Let $z_1 \in \mathbb{C}$ be another distinct point. Prove that there exists a generalized circle C_1 that is tangent to C at z_0 and which passes through z_1. *(Hint: Considering taking the inversion through the circle centered at z_0 and passing through z_1. What are the images of C, C_1, z_0, z_1?)*
    d) Finish the proof of Lemma 2.7.

11. Let z_1, z_2, z_3, z_4 be four distinct points in $\mathbb{C}P^1$. Prove each of the following assertions.

    a) $[z_1, z_2; z_4, z_3] = 1/[z_1, z_2; z_3, z_4]$.
    b) $[z_3, z_4; z_1, z_2] = [z_1, z_2; z_3, z_4]$.
    c) $[z_1, z_3; z_2, z_4] = 1 - [z_1, z_2; z_3, z_4]$.

12. Let z_1, z_2, z_3, z_4 be four distinct points in $\mathbb{C}P^1$. Let $[z_1, z_2; z_3, z_4] = \lambda$. Prove each of the following assertions. *(Hint: You may want to use the result of Exercise 2.2.11.)*

    a) $[z_1, z_2; z_3, z_4] = [z_2, z_1; z_4, z_3] = [z_3, z_4; z_1, z_2] = [z_4, z_3; z_2, z_1] = \lambda$.
    b) $[z_1, z_2; z_4, z_3] = [z_2, z_1; z_3, z_4] = [z_3, z_4; z_2, z_1] = [z_4, z_3; z_1, z_2] = \frac{1}{\lambda}$.
    c) $[z_1, z_3; z_2, z_4] = [z_2, z_4; z_1, z_3] = [z_3, z_1; z_4, z_2] = [z_4, z_2; z_3, z_1] = 1 - \lambda$.
    d) $[z_1, z_3; z_4, z_2] = [z_2, z_4; z_3, z_1] = [z_3, z_1; z_2, z_4] = [z_4, z_2; z_1, z_3] = \frac{1}{1-\lambda}$.
    e) $[z_1, z_4; z_2, z_3] = [z_2, z_3; z_1, z_4] = [z_3, z_2; z_4, z_1] = [z_4, z_1; z_3, z_2] = \frac{\lambda-1}{\lambda}$.
    f) $[z_1, z_4; z_3, z_2] = [z_2, z_3; z_4, z_1] = [z_3, z_2; z_1, z_4] = [z_4, z_1; z_2, z_3] = \frac{\lambda}{\lambda-1}$.

13. Prove that $[z_1, z_2; z_3, z_4] = [z_5, z_2; z_3, z_4]$ if and only if $z_1 = z_5$. *(Hint: You may want to use the fact that linear fractional transformations allow you to move any three points to any other three points and do not change the cross-ratio.)*
14. Prove that for any pair of distinct quadruples of points $z_1, \ldots, z_4$ and $w_1, \ldots, w_4$ with the property that $[z_1, z_2; z_3, z_4] = [w_1, w_2; w_3, w_4]$, there exists a unique element $\varphi \in \text{Möb}^0(2)$ such that $\varphi(z_i) = w_i$ for $i = 1, 2, 3, 4$. *(Hint: You may want to use the result of Exercise 2.2.13.)*

2.3 PROOFS (Calculus)

1. Assuming that a, b, c, d are complex numbers, $c \neq 0$, and $ad - bc \neq 0$, prove that
$$\lim_{z \to -d/c} \left| \frac{az + b}{cz + d} \right| = \infty.$$
If you have not seen limits of complex numbers before, you can interpret this by doing the following change of variables: write $z = -d/c + re^{i\theta}$ and show that
$$\lim_{r \to 0} \left| \frac{az + b}{cz + d} \right| = \infty$$
regardless of the choice of θ.
2. Assuming that a, b, c, d are complex numbers, $c \neq 0$, and $ad - bc \neq 0$, prove that
$$\lim_{z \to \infty} \frac{az + b}{cz + d} = \frac{a}{c}.$$
If you have not seen limits of complex numbers before, you can interpret this as follows: prove that
$$\lim_{r \to \infty} \frac{are^{i\theta} + b}{cre^{i\theta} + d} = \frac{a}{c}$$
regardless of the choice of θ.

2.4 PROOFS (Group Theory)

1. Let G, H be groups with identities ι_G and ι_H, and operations $*$ and $\circ$. Let $\varphi : G \to H$ be a group homomorphism.

   a) Prove that $\varphi(\iota_G) = \iota_H$.
   b) Prove that $\varphi(a^{-1}) = \varphi(a)^{-1}$.

2. Let G, H be groups and let $f : G \to H$ be a group homomorphism. Prove that f is a group isomorphism if and only if there exists a group homomorphism $g : H \to G$ such that $(f \circ g)(x) = x$ and $(g \circ f)(y) = y$ for all $x \in H$ and $y \in G$.
3. Let G be a subgroup of H. Prove that the obvious inclusion map
$$\varphi : G \to H$$
$$a \mapsto a$$
is a group homomorphism.

4. Consider $\mathbb{Z}$ and $G = \{\text{even, odd}\}$ as groups under addition. Prove that the map
$$\begin{aligned}\varphi : \mathbb{Z} &\to G\\ n &\mapsto \begin{cases}\text{even} & \text{if } x \text{ is even}\\ \text{odd} & \text{otherwise}\end{cases}\end{aligned}$$
is a group homomorphism.
5. Define $PGL(2, \mathbb{C})$ to be the set of equivalence classes of elements in $GL(2, \mathbb{C})$, where two matrices M_1, M_2 are considered to be equivalent if there exists $\lambda \in \mathbb{C}$ such that $M_1 = \lambda M_2$.

   a) Let $M_1, M_1', M_2, M_2' \in GL(2, \mathbb{C})$ such that M_1, M_1' are equivalent and M_2, M_2' are equivalent. Prove that $M_1 M_2$, $M_1' M_2'$ are equivalent.
   b) Using the above, prove that $PGL(2, \mathbb{C})$ is a group. *(Hint: The main problem is showing that it has a well-defined multiplication on it. The previous part suggested how to do this.)*
   c) Prove that the quotient map
   $$\begin{aligned}\varphi : GL(2, \mathbb{C}) &\to PGL(2, \mathbb{C})\\ M &\mapsto M\end{aligned}$$
   is a group homomorphism.

6. Prove that Möb$^0(2)$ is isomorphic to $PGL(2, \mathbb{C})$. *(Hint: Alter the group homomorphism in Theorem 2.4 slightly using your result from Exercise 2.2.5.)*
7. Let G, H be isomorphic groups. Prove that G is abelian if and only if H is abelian.
8. Given a set $X = \{x_1, x_2, \dots x_n\}$ of n elements, a *permutation* is a bijection function $\sigma : X \to X$. (Intuitively, σ simply permutes the elements of X.) The *symmetric group on n elements* is the set S_n of all permutations $\sigma : X \to X$. Prove that S_n is a group with function composition as the operation.
9. Choose four arbitrary complex numbers z_1, z_2, z_3, z_4—which ones we select is virtually irrelevant, but for concreteness let's suppose that they are $3, 1, 0, \infty$. Consider the set $X = \{z_1, z_2, z_3, z_4\}$ and the symmetric group S_4 that shuffles around its elements. Let K be the subset of S_4 consisting of permutations σ such that $[\sigma(z_1), \sigma(z_2); \sigma(z_3), \sigma(z_4)] = [z_1, z_2, z_3, z_4]$. Write down the elements of K and show that it is a subgroup of S_4. *(Hint: You may want to do Exercise 2.2.12 first.)*
10. In Exercise 2.4.9, to what extent do the four complex numbers that we choose matter? Could we select them in such a way that the resulting subgroup of permutations is smaller than K?

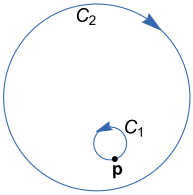
3 Applications of Inversive Geometry

In which we look at how ancient questions can be answered using modern tools.

In Chapters 1 and 2, we studied the properties of linear fractional transformations; I motivated this as a means of understanding certain kinds of geometries. Now is a good time to make good on this promise: we are going to see how convenient inversive geometry is when attacking various problems that would have given the ancient Greeks and later geometers trouble.

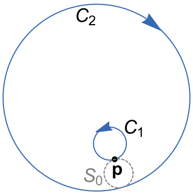
3.1 Steiner's Porism

We begin by considering a comparatively modern problem posed by Jakob Steiner in the nineteenth century, but which could just as easily have been asked by any of the ancient Greeks. The basic setup is as follows: consider two oriented circles C_1, C_2 in the plane which are not tangent and whose interiors do not intersect. Classically, one of these circles is contained inside the other one (in which case we take the orientation of the inner one to be counter-clockwise, and the orientation of the outer one to be clockwise). However, we shall see that we can take one of the oriented circles to be a line without changing anything substantive. In any case, choose a point p on C_1.

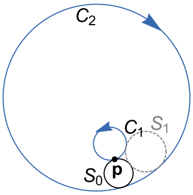


A. Sheydvasser, *Linear Fractional Transformations*, Undergraduate Texts in Mathematics, https://doi.org/10.1007/978-3-031-25002-6_3

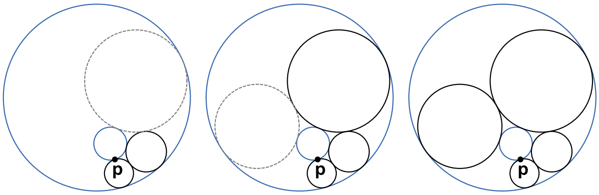
There is a unique oriented circle S_0 that is tangent to C_1 at p, tangent to C_2, and whose interior does not intersect the interiors of C_1 or C_2—this might not be obvious at this point, but we will prove it.



Choose an oriented circle S_1 which is tangent to C_1, C_2, and S_0—there are two possible choices.



Then there is a unique oriented circle S_2 which is tangent to S_1, C_1, C_2. In fact, we can keep going inductively, adding an oriented circle S_n tangent to S_{n-1}, C_1, C_2 at each step.



If we require that none of the interiors of these circles intersect, then this process will eventually halt, producing what is called a *Steiner chain*. If the final circle is tangent to S_0, then the chain is said to *closed*; otherwise, it is *open*. Above, we have drawn an example of an open Steiner chain. Figure 3.1 shows some closed Steiner chains differing only in the choice of the initial point.

There are already various things that we might want to prove rigorously here, to show that Steiner chains are well defined. We might want to prove that there is a unique oriented circle through p tangent to both C_1 and C_2 which doesn't intersect their interiors. We might want to prove that once we choose a direction, there is a unique tangent circle that we can put in at each step of the algorithm. We might want to prove that this procedure always halts after finitely many steps. All of these are worthy considerations, but these are not Steiner's porism. Suppose that we fix our two circles C_1, C_2, but we change the point p on C_1. This will give us a new Steiner chain. What properties does it share with the old chain? Remarkably, if you try out some examples, you will quickly discover that it appears that:

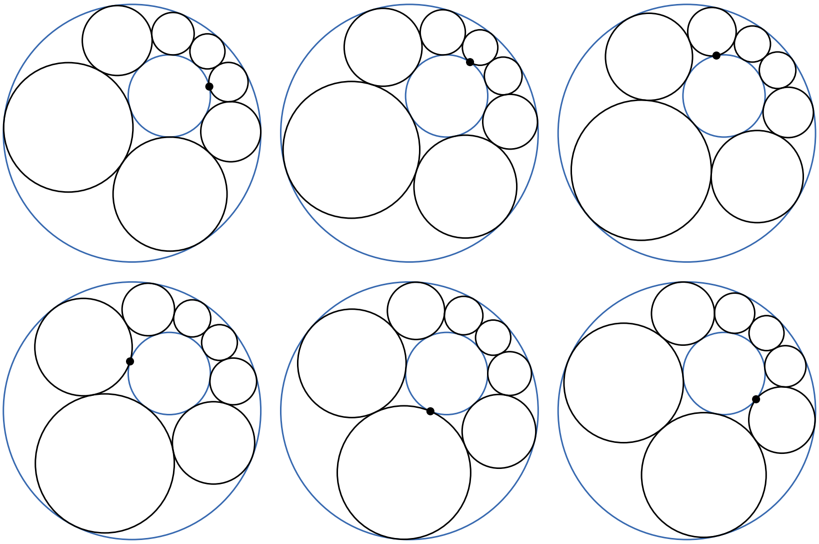


Fig. 3.1 A collection of closed Steiner chains differing only in the choice of the initial point.

1. the number of circles in the chain does not depend on the choice of point p and
2. either all chains are open or all chains are closed, regardless of p.

This is Steiner's porism, and it is the main result that we will try to prove in this section. All of our proofs will be agnostic about the exact configurations of our circles and points—we shall always require just that there are two generalized circles C_1, C_2 which are not tangent and whose interiors do not intersect. However, we will see that one can always assume without loss of generality that C_1 is a circle inside a circle C_2, and, in fact, one can request something even stronger. Before we get into that, though, let's take some time to properly convince ourselves that Steiner chains are well defined, via a sequence of lemmas.

Lemma 3.1 *Let C_1, C_2 be two oriented circles that are not tangent and whose interiors do not intersect. For any point p on C_1, there exists a unique oriented circle C_3 tangent to C_1 at p, tangent to C_2, and whose interior does not intersect the interiors of C_1 or C_2.*

Proof Recall that linear fractional transformations preserve both angles and generalized circles but allow us to move any three points in $\mathbb{C}P^1$ to any other three points; therefore, if we choose two points p', p'' on C_1 aside from p, there exists a unique $\varphi \in \text{Möb}^0(2)$ such that $\varphi(p) = \infty$, $\varphi(p') = 0$, $\varphi(p'') = 1$, and the image under φ of C_1 is the real line. By rotating if necessary, we can assume that the interior of the image of C_1 lies below the real line and therefore C_2 is a circle lying above the real line with a counter-clockwise orientation. Note that if the statement of the lemma is true of this new configuration, then it has to be true of the old configuration.

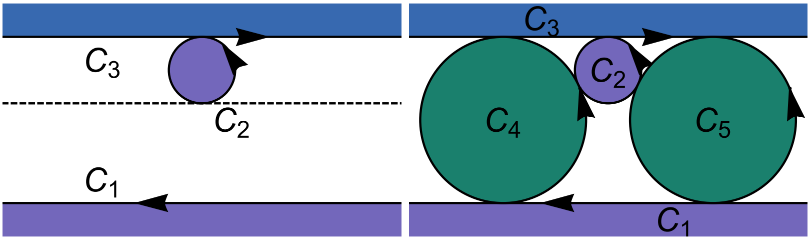


Fig. 3.2 On the left, an illustration of the circle configuration constructed in Lemma 3.1; C_1, C_2 are shown in purple, and C_3 is shown in blue. On the right, an illustration of the configuration constructed in Lemma 3.2; C_1, C_2 are shown in purple, C_3 is shown in blue, and C_4, C_5 are shown in green.

Therefore, we can assume without loss of generality that we started with this configuration. If C_3 is tangent to C_1 at $p = \infty$, then C_3 must be a line; more precisely, a line parallel to the real line. It is easy to see that there are two lines tangent to C_2 which are parallel to the real line. However, only one of them can be given an orientation such that its interior does not intersect the interiors of C_1 or C_2. This is illustrated in Figure 3.2, where the original two generalized circles are drawn with their interiors in purple, and the only oriented circle satisfying the desired conditions is drawn in blue. The other tangent line is dashed. □

This is a good place for an important aside: note that the crucial idea behind the proof of the previous lemma was to move the given configuration of circles to a standard one. In so doing, we simplify: we start with a configuration that might be difficult to analyze and work with, but then by using a linear fractional transformation, we can reduce to a particular case that is easy to think about instead. This fundamental insight is captured in the following philosophical statement, which we will be using implicitly throughout this chapter.

Philosophical Principle

Given a difficult geometry problem involving circles, lines, and angles (but not necessarily distances), see if you can use a Möbius transformation to transform a complicated configuration into a simple one where the answer is easy to see.

Lemma 3.2 *Let C_1, C_2 be two oriented circles which are not tangent and whose interiors do not intersect. Let C_3 be an oriented circle tangent to both of them, but such that its interior does not intersect theirs. There exist exactly two oriented circles C_4, C_5 that are tangent to C_1, C_2, C_3 with an orientation such none of the interiors of C_1, C_2, C_3, C_4, C_5 intersect.*

Proof Let p be the point at which C_3 is tangent to C_1. Using a linear fractional transformation, we can move p to ∞. By applying rotations, translations, and dilations, we may assume that C_1 is the real line whose interior is below the real axis; C_2 is a circle above the real axis with a counter-clockwise orientation with center $(0, y_0)$ and radius 1; and C_3 is a line $y = y_0 + 1$ tangent to C_2 whose interior lies above it. It is plainly obvious that any circle with center $(x_0, (y_0+1)/2)$ and radius $(y_0+1)/2$ will be tangent to both C_1 and C_3; by changing x_0, we can arrange for this circle to be tangent to C_2 in one of three ways, two of which are illustrated in Figure 3.2. (The third way would have the interior of this circle intersect one the interiors of at least one of C_1, C_2, C_3.) This assertion can be proved formally, although at present we lack the tools to prove it elegantly—this will be rectified later in this chapter. However, it can still be done by solving the set of equations

$$(x - x_0)^2 + \left(y - \frac{y_0+1}{2}\right)^2 = \left(\frac{y_0+1}{2}\right)^2$$
$$x^2 + (y - y_0)^2 = 1$$

where the solutions are points (x, y) that lie both on C_2 and our new circle. This is something of an algebraic mess, but it resolves to

$$x = \frac{-2x_0^2\,(y_0-1) \pm (y_0-1)\sqrt{x_0^4\left(-x_0^2+2y_0+2\right)} + 2x_0^4}{x_0\left(4x_0^2 + (y_0-1)^2\right)}$$
$$y = \frac{x_0^2\,(3y_0+1) \pm 2\sqrt{x_0^4\left(-x_0^2+2y_0+2\right)} + (y_0-1)^2\,(y_0+1)}{4x_0^2 + (y_0-1)\,2}.$$

The intersection point should be unique, which means that we need $x_0^2(-x_0^2+2y_0+2) = 0$. If $x_0 = 0$, then the interior of our new circle intersects either C_2 or C_1 and C_3. The other possibility is that $x_0 = \pm\sqrt{2(y_0+1)}$, whence our two solutions. Giving our new solutions the counter-clockwise orientation, we are done. □

Thus, we conclude that our intuitive definition of Steiner chains is perfectly valid regardless of the circles C_1, C_2 that we choose. It remains to attack Steiner's porism itself. The methods we use are essentially the same as for the lemma.

Theorem 3.1 (Steiner's Porism) *Let C_1, C_2 be two oriented circles that are not tangent and whose interiors do not intersect. All Steiner chains starting with these two circles have the same number of circles and are either all open or all closed.*

Proof The key observation is that if we apply an inversive transformation to a Steiner chain, then the result will also be a Steiner chain. Furthermore, this new Steiner chain will have the same number of circles as the original and will be closed if and only if the original was closed. Thus, we can use inversive geometry to reduce the general case to a simple case that is easy to work with. The visual picture for this is given in Figure 3.3. First, we note that we can always assume that both C_1 and C_2 are circles—if not, simply choose a point z_0 that does not lie on either C_1 or C_2, and apply the

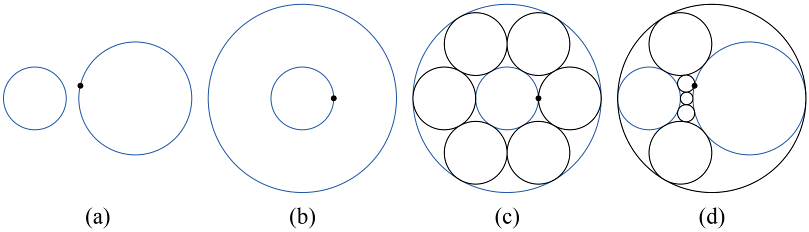


Fig. 3.3 A visual sketch of the proof of Steiner's porism. (a) shows the starting configuration of two circles; (b) shows the image of this configuration, put into standard position; (c) shows a completed Steiner chain in this standard position; (d) shows how this Steiner chain lifts back to the original configuration.

transformation $z \mapsto 1/(z - z_0)$. Second, we can assume that C_1 is contained inside C_2—if not, simply take a circle inversion through C_2. This reduces to the classical setting in which Steiner's porism is usually presented. By applying a sequence of translations, rotations, and dilations, we can assume that C_2 is the unit circle and that the center of C_1 lies on the real axis. The final step is to apply a linear fractional transformation that fixes C_2 but moves C_1 to a circle that is concentric with C_2; the only difficulty is proving that such a transformation exists.

There are a number of different ways to prove this; we will take a hybrid geometric/analytic approach. First, note that the real line passes through the centers of both C_1 and C_2, and therefore is perpendicular to both of them. Find any other generalized circle C_3 that is perpendicular to both C_1 and C_2; the most convenient choice is to have C_3 perpendicular to the real line as well. Here is one way to show that such a circle exists: for any real number r, there exists a unique generalized circle perpendicular to both the real line and C_2 that passes through r. (See Exercise 3.2.1) If we take r to be inside C_1, then this circle intersects C_1 at some angle θ which ranges from 0 to π. Since this function $r \mapsto \theta$ is continuous, by the intermediate value theorem, there is some r where $\theta = \pi/2$ exactly—that is, the circle we have chosen is perpendicular. Let p be the point where the real line intersects with C_3 and let $\varphi \in \text{Möb}^0(2)$ such that $\varphi(-1) = -1$, $\varphi(1) = 1$, and $\varphi(p) = 0$. The image of the real line under φ is itself. The image of C_2 is a circle that is perpendicular to the real line at -1 and 1—however, there is only one such circle, and that is C_2, the unit circle. The image of C_3 is a circle that is perpendicular to the real line and C_2 and passes through 0—there is only one such circle, and that is the vertical line $x = 0$. Thus, C_1 is a circle that is perpendicular to $y = 0$ and $x = 0$—this is true if and only if the center of C_1 is 0. (See Exercise 3.2.2) Thus, we have shown that Steiner's porism is true in general as long as it is true of concentric circles. However, for concentric circles, it is obvious that Steiner's porism is true—first of all, we can use a rotation to move the starting point for the chain onto the positive real axis; secondly, we can use a reflection if necessary to switch between the two possible choices of circle in the second step of the construction. Therefore, all Steiner chains have the same number of circles, and they are either all closed or all open. □

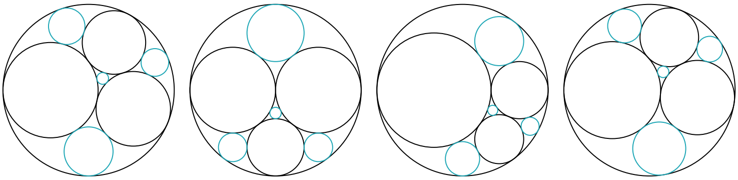


Fig. 3.4 For any three mutually tangent circles in these drawings, you can see that there are precisely two circles tangent to all three of them.

3.2 Apollonian Gaskets

In the third century BCE, Apollonius of Perga asked the following question.

▶ **Question** Consider three circles in the Euclidean plane. How many circles exist that are tangent to each of these three circles simultaneously? How might one find such circles?

From the writings of other Greeks, we know that Apollonius gave a solution to this problem; sadly, the details of how he did it are lost to history. On the other hand, we might wonder how we might attack this using inversive geometry by reducing the general problem to a simpler one using linear fractional transformations to move the circles into a standard configuration. However, there is still some splitting into cases that must happen, because it matters if the three initial circles intersect or not. We shall consider a special case of Apollonius' problem which has been of the greatest interest to modern mathematicians.

▶ **Question** Consider three circles in the Euclidean plane such that any two circles are tangent to one another. How many circles exist that are tangent to each of these three circles simultaneously?

We will see that the answer to this question is quite simple: there are always exactly two such (generalized) circles. The proof of this is not at all complicated and comes from the following slight generalization.

Theorem 3.2 *Let C_1, C_2, C_3 be three distinct, mutually tangent generalized circles in $\mathbb{C}P^1$. Then there exist exactly two generalized circles C_4, C_4' that are tangent to each of C_1, C_2, C_3.*

Remark 3.1 Some concrete examples of this result are shown in Figure 3.4.

Proof Let z_1, z_2, z_3 be the points where C_1, C_2, C_3 are tangent. Use a linear fractional transformation to send $z_1 \mapsto \infty$, $z_2 \mapsto 0$. What does our configuration of generalized circles now look like? Well, two of them pass through ∞, so they must be lines—since they are tangent at ∞, they must be parallel lines. Furthermore, one of them passes through 0. Using a rotation and dilation if necessary, we can move this configuration further so that one circle is the real line and another is $y = 1$. The final generalized circle is a circle that is tangent to $y = 0$ at 0 and is tangent to $y = 1$. What is this circle? Well, since it is tangent to $y = 0$ at 0, it must be of the form

$$x^2 + (y - y_0)^2 = y_0^2$$

for some $y_0 \in \mathbb{R}$. There is only one such circle that is tangent to $y = 1$, and that is $x^2 + (y - 1/2)^2 = 1/4$. Note as well that, more generally, any circle that is tangent to both $y = 0$ and $y = 1$ must be a translation of this circle—this is because we can always translate the point of intersection at $y = 0$ to 0 if need be. Therefore, any circle that is tangent to the images of C_1, C_2 must be of the form

$$(x - x_0)^2 + (y - 1/2)^2 = 1/4.$$

There are exactly two such circles that are tangent to the image of C_3, and those occur if $x_0 = -1$ or $x_0 = 1$. However, since we moved C_1, C_2, C_3 by linear fractional transformations, the number of tangent circles could not have changed. □

There is a useful corollary of this result that applies to Descartes configurations.

Definition 3.1 We say that four generalized circles are a *Descartes configuration* if any three of them are mutually tangent to one another.

Corollary 3.1 (Existence and Uniqueness of Descartes Swaps) *For any Descartes configuration C_1, C_2, C_3, C_4, there is a unique circle C_4' such that C_1, C_2, C_3, C_4' is also a Descartes configuration.*

Proof Note that, by definition, C_1, C_2, C_3 are three mutually tangent circles and so, by Theorem 3.2, we know that there exist exactly two circles that are mutually tangent to each of C_1, C_2, C_3. One of them must be C_4; the other one is C_4'. It is easy to see from the definition that C_1, C_2, C_3, C_4' is a Descartes configuration. □

Exchanging C_4 for C_4' is often called a *Descartes swap*. For any Descartes configuration, there are exactly four corresponding Descartes swaps, coming from the four different choices of circle that we can swap out. One might well ask whether there is some kind of geometric interpretation that we can give to these swaps, and indeed there is.

Theorem 3.3 (Geometric Interpretation of Descartes Swaps) *For any Descartes configuration C_1, C_2, C_3, C_4, the Descartes swap*

$$(C_1, C_2, C_3, C_4) \mapsto (C_1, C_2, C_3, C_4')$$

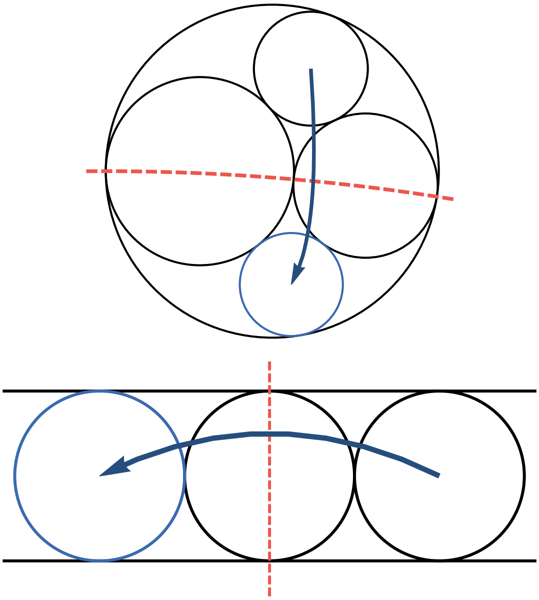


Fig. 3.5 Two examples of Descartes swaps.

is given by an inversion through the unique circle that passes through the intersection points of C_1, C_2, C_3. *Furthermore, this aforementioned circle is orthogonal to* C_1, C_2, C_3.

Proof The easiest way to prove this is to reduce to the standard Descartes configuration that we had previously, where C_1, C_2 are the lines $y = 0$ and $y = 1$, C_3 is the circle with center $i/2$ and radius $1/2$, and C_4 is the circle with center $1 + i/2$ and radius $1/2$. Then C_4' is the circle with center $-1 + i/2$ and radius $1/2$; we see that that is exactly the image of the reflection through the line $x = 0$, as illustrated in Figure 3.5. This reflection indeed passes through the intersection points of C_1, C_2, C_3, and is orthogonal to those three circles. Finally, this reflection does not move C_1, C_2, C_3, so indeed it moves the Descartes configuration C_1, C_2, C_3, C_4 to the Descartes configuration C_1, C_2, C_3, C_4', as desired. □

With this result in mind, we make a definition.

Definition 3.2 For any Descartes configuration C_1, C_2, C_3, C_4, the collection of circles that are orthogonal to triples of C_1, C_2, C_3, C_4 are known as the *dual circles*. The collection of all transformations in $\text{Möb}(2)$ that can be written as a composition of reflections through these dual circles is known as the *Apollonian group*. (The reader should check for themselves that this is indeed a group—see Exercise 3.3.1.)

We will give a more algebraic description of the Apollonian group later in this chapter. For now, we content ourselves with exploring its geometric significance,

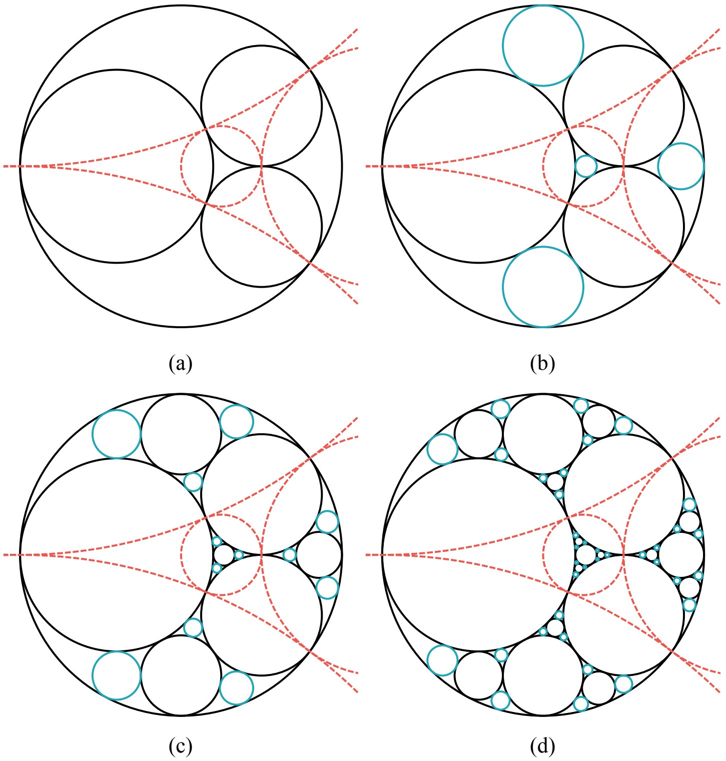


Fig. 3.6 The iterative construction of an Apollonian gasket. We start with a Descartes configuration in (a), drawing the dual circles in red. In (b), we add all Descartes swaps of the initial configuration through the dual circles. In (c), we add all Descartes swaps of the new circles, and in (d) we repeat this process again.

which comes from Apollonian gaskets. Notice that since Descartes swaps move Descartes configurations to Descartes configurations, we could iterate the process, doing it over and over again. If we do this ad infinitum, the result is what is known as an Apollonian gasket, which is shown in Figure 3.6.

Definition 3.3 Let C_1, C_2, C_3, C_4 be a Descartes configuration. The *Apollonian gasket with starting configuration* C_1, C_2, C_3, C_4 is the smallest set $\mathcal{S}$ of generalized circles in $\mathbb{C}P^1$ such that

1. $\mathcal{S}$ contains C_1, C_2, C_3, C_4 and
2. if C_1', C_2', C_3', C_4' are all circles in $\mathcal{S}$ that form a Descartes configuration, then all of the Descartes swaps of these circles are also in $\mathcal{S}$.

Remark 3.2 Such configurations are also sometimes called Leibniz packings as they were first described by the mathematician Gottfried Leibniz in a letter to de Brosses in the seventeenth century.

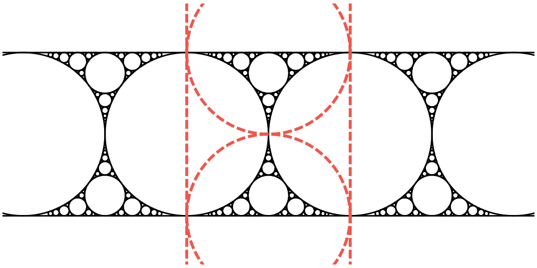


Fig. 3.7 The standard Apollonian gasket.

Strictly speaking, this is defined in a way that is a little different than how we described it initially, but it is not hard to see that these descriptions are equivalent. (See Exercise 3.2.4.) However, our definition is easier to work with. We begin by noting that Apollonian gaskets are all essentially the same. Call the Descartes configuration with circles $y = 0$, $y = 1$, $x^2 + (y/2)^2 = 1/4$, and $(x - 1)^2 + (y/2)^2 = 1/4$ the *standard configuration*—the corresponding Apollonian gasket is illustrated in Figure 3.7.

Theorem 3.4 *Let C_1, C_2, C_3, C_4 and D_1, D_2, D_3, D_4 be two Descartes configurations. Let A_1, A_2 be the corresponding Apollonian gaskets. If $\varphi \in M\ddot{o}b^0(2)$ is such that $\varphi(C_i) = D_i$ for $i = 1, 2, 3, 4$, then $\varphi(A_1) = A_2$.*

Proof It is not hard to see that linear fractional transformations preserve Descartes swaps since they preserve tangencies. Therefore, $\varphi(A_1)$ is a set of generalized circles that contains D_1, D_2, D_3, D_4 and such that for any quadruple in the set, all of their Descartes swaps are also in the set. The set $\varphi(A_1)$ must be the smallest set with this property—if it were not, then $\varphi^{-1}(A_2)$ would be a proper subset of A_1 closed under Descartes swaps and containing C_1, C_2, C_3, C_4. However, this would violate the definition of A_1 as the smallest set with that property. We conclude that $\varphi(A_1) = A_2$. □

Corollary 3.2 *Let A be an Apollonian gasket with initial configuration C_1, C_2, C_3, C_4. There exists $\varphi \in M\ddot{o}b^0(2)$ such that $\varphi(A)$ is the Apollonian gasket of the standard configuration.*

Proof We already know that any Descartes configuration can be reduced to the standard configuration by linear fractional transformations. The rest follows from Theorem 3.4. □

3.3 Inversive Coordinates

Thus far, our computations of the images of generalized circles under linear fractional transformations have been inefficient: we have had to compute equations describing those circles, calculate how those equations transform, and then finally work out what is the new corresponding generalized circle. While this is certainly not impossible to do, there is a substantially faster way via inversive coordinates, which will aid us in our investigations into Steiner's porism and Apollonian gaskets.

Definition 3.4 For any oriented circle C, let $\kappa(C)$ be the *bend* of C—that is,

$$\kappa(C) = \begin{cases} \frac{1}{R} & \text{if } C \text{ is a circle with radius } R\text{, oriented counter-clockwise} \\ -\frac{1}{R} & \text{if } C \text{ is a circle with radius } R\text{, oriented clockwise} \\ 0 & \text{if } C \text{ is a line.} \end{cases}$$

The *co-bend* of C—denoted $\kappa'(C)$—is the bend of the image of C under the map $z \mapsto -1/z$. The *bend-center* of C is denoted by $\xi(C)$ and

$$\xi(C) = \begin{cases} \kappa(C)z_0 & \text{if } C \text{ is a circle with center } z_0 \\ ie^{i\theta} & \text{if } C \text{ is a line traversed in the direction } e^{i\theta}. \end{cases}$$

Together, $(\kappa(C), \kappa'(C), \xi(C))$ are the *inversive coordinates* for C.

Remark 3.3 It is more common in the literature to refer to the bend, co-bend, and bend-center as the *curvature*, the *co-curvature*, and the *curvature-center*—see [17], for example. Furthermore, it is more typical to see the co-bend defined in terms of $1/z$ rather than $-1/z$. My reason for departing from these conventions is very simple: one can generalize all these definitions for higher dimensional spaces and study, for instance, oriented spheres in $\mathbb{R}^3 \cup \{\infty\}$. (I did exactly that in my thesis [14].) In that case, you want to define inversive coordinates exactly as I have here—e.g. the bend should be $\pm 1/R$ where R is the radius. However, the usual definition of curvature for a sphere is $\pm 1/R^2$. Similarly, while $1/z$ is orientation-preserving as a transformation on the plane, as a transformation on $\mathbb{R}^3$, it is orientation-reversing! On the other hand, $-1/z$ is still a perfectly good orientation-preserving transformation.

The inversive coordinates of an oriented circle specify it uniquely.

Lemma 3.3 *The map*

$$\begin{aligned} inv : \{\textit{oriented circles in } \mathbb{C}P^1\} &\to \mathbb{R}^4 \\ C &\mapsto (\kappa(C), \kappa'(C), \Re(\xi(C)), \Im(\xi(C))) \end{aligned}$$

is injective and $inv(-C) = -inv(C)$ *for all oriented circles* C.

Proof Suppose that $\text{inv}(C_1) = \text{inv}(C_2)$. Since $\kappa(C_1) = \kappa(C_2)$, either they are both circles or both lines. If they are both circles, then they have the same radius, center, and orientation since $\kappa(C_1) = \kappa(C_2)$ and $\xi(C_1) = \xi(C_2)$. If they are both lines, the fact that $\xi(C_1) = \xi(C_2)$ tells us that they are parallel. Since they are parallel, we can find a line through the origin that is perpendicular to both of them. Now, what is the image under the transformation $z \mapsto -1/z$? Well, the line through the origin will still be a line through the origin, which will still be orthogonal to the images of C_1 and C_2, which will now both be circles passing through 0—indeed, since C_1 and C_2 were tangent at ∞, these circles must be tangent at 0. Since both of these lines are orthogonal to a line through the origin, their centers must lie on this line. Since $\kappa'(C_1) = \kappa'(C_2)$, the radii of these circles must be the same, so they must either coincide or be reflections across a line normal to both of them; furthermore, they must have the same orientation. However, the second case is impossible due to orientation considerations: under the transformation $z \mapsto -1/z$, the two circles will map to lines pointing in opposite directions. Therefore, $C_1 = C_2$. Proving that $\mathrm{inv}(-C) = -\mathrm{inv}(C)$ is left as an exercise to the reader. (See Exercise 3.2.5.) □

A curious fact is that the inversive coordinates of any oriented circle always lie on the surface of a hyperboloid.

Theorem 3.5 *Let C be an oriented circle with inversive coordinates (κ, κ', ξ). Then $-\kappa\kappa' + |\xi|^2 = 1$.*

Proof If C is a line, then $\kappa = 0$ and ξ is a unit vector, and so the claim follows immediately. Otherwise, we note that since $\mathrm{inv}(-C) = -\mathrm{inv}(C)$, we may assume without loss of generality that C is a circle with positive orientation—in that case, we know that if C has center z_0 and radius R, then $\kappa = 1/R$, $\xi = z_0/R$. Note that if we rotate C around the origin by θ radians, then this won't change the bend or co-bend, and will merely change ξ to $e^{i\theta}\xi$—thus, this will not change the value of $-\kappa\kappa' + |\xi|^2$. Thus, we may assume that $z_0 \geq 0$. If $z = 0$, it is easy to check that $\kappa' = -R$, so it remains to consider the case where $z_0 > 0$. We know that our circle together with its interior will be the set of points satisfying $|z - z_0| \leq R^2$; its image under the map $z \mapsto -z^{-1}$ will therefore be the set of points satisfying $| - z^{-1} - z_0| \leq R$. Equivalently, this is the set of points $1 + 2z_0\Re(z) + |z|^2 z_0^2 \leq |z|^2 R^2$. If $z_0 = R$, then this is the equation of a half-plane, hence $\kappa' = 0$. However, then $\xi = z_0/R = 1$, and so $-\kappa\kappa' + |\xi|^2 = 1$. If $z_0 \neq R$, then instead we can complete the square, yielding

$$(z_0^2 - R^2)\left(|z|^2 + \frac{2z_0}{z_0^2 - R^2}\Re(z) + \frac{z_0^2}{\left(z_0^2 - R^2\right)^2}\right) + 1 \leq \frac{z_0^2}{z_0^2 - R^2}$$

$$(z_0^2 - R^2)\left|z + \frac{z_0}{z_0^2 - R^2}\right|^2 \leq \frac{R^2}{z_0^2 - R^2}$$

whence

$$\begin{cases} \left|z + \frac{z_0}{z_0^2 - R^2}\right|^2 \leq \frac{R^2}{\left(z_0^2 - R^2\right)^2} & \text{if } z_0 > R \\ \left|z + \frac{z_0}{z_0^2 - R^2}\right|^2 \geq \frac{R^2}{\left(z_0^2 - R^2\right)^2} & \text{if } z_0 < R. \end{cases}$$

Therefore, $\kappa' = (z_0^2 - R^2)/R$ and consequently

$$-\kappa\kappa' + |\xi|^2 = -\frac{1}{R} \cdot \frac{z_0^2 - R^2}{R} + \frac{z_0^2}{R^2} = 1,$$

exactly as claimed. □

There is another way to write down inversive coordinates that is a little more useful for our purposes. Specifically, suppose that we have an oriented circle C and we write the matrix

$$M(C) = \begin{pmatrix} \kappa'(C) & \xi(C) \\ \overline{\xi(C)} & \kappa(C) \end{pmatrix}.$$

This captures the same information. Moreover, it is easy to see that $\det(M(C)) = \kappa(C)\kappa'(C) - |\xi(C)|^2$. As a consequence, we have the following observation.

Theorem 3.6 *There exists a bijective map*

$$inv : \{oriented\ circles\ in\ \mathbb{C}P^1\} \to \left\{ M \in Mat(2, \mathbb{C}) \middle| M = \overline{M}^T,\ \det(M) = -1 \right\}$$

$$C \mapsto \begin{pmatrix} \kappa'(C) & \xi(C) \\ \overline{\xi(C)} & \kappa(C) \end{pmatrix}$$

where $Mat(2, \mathbb{C})$ *denotes the set of all* 2×2 *matrices with complex coefficients and* $\overline{M}^T$ *denotes the conjugate transpose—i.e.*

$$M = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \Rightarrow \overline{M}^T = \begin{pmatrix} \overline{a} & \overline{c} \\ \overline{b} & \overline{d} \end{pmatrix}.$$

Remark 3.4 It is technically an abuse of notation to call this map inv as well, but I think it is acceptable since it will always be clear from context whether we are thinking of inversive coordinates as vectors or matrices.

Proof Notice that $M = \overline{M}^T$ if and only if

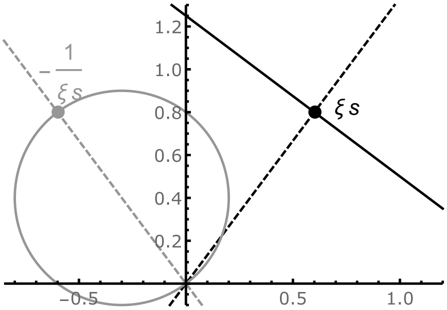
$$M = \begin{pmatrix} a & x + iy \\ x - iy & b \end{pmatrix}$$

for some $a, b, x, y \in \mathbb{R}$; by this and Theorem 3.5, we see that the defined map is well defined. That it is injective follows immediately from Lemma 3.3. It remains to prove surjectivity, which isn't too hard. Consider a matrix

$$M = \begin{pmatrix} a & x + iy \\ x - iy & b \end{pmatrix}.$$

If $b = 0$, then we know that $|x + iy| = 1$, so $x + iy = e^{i\theta}$ for $\theta \in \mathbb{R}$. Let C be the line traversed in the direction $-ie^{i\theta}$ and passing through the point $ae^{i\theta}/2$. Then one can check that $\text{inv}(C) = M$. If $b \neq 0$, then there exists a circle C with bend b and center $(x + iy)/b$. The co-bend κ' of this circle must satisfy $-b\xi + |x + iy|^2 = 1$, which means that $\xi = a$. Thus, $\text{inv}(C) = M$, and we are done. □

▶ **Example** *Let C be a line with inversive coordinates $(0, \kappa', \xi)$. Find a parametric equation for C in terms of κ' and ξ.*
We know that C is traversed in the direction $-i\xi$ by the definition of the bend-center, so it suffices to find a single point on C. Since we know that ξ points in the direction of the interior of C, we know that $s\xi \in C$ for some real s. The image of C under $z \mapsto -1/z$ has to contain both 0 and $-1/(s\xi) = -\overline{\xi}/s$, as these are the images of ∞ and $s\xi$. The line L through the origin in the direction ξ passes through both of these points and is orthogonal to C at the intersection point $t\xi$; its image under $z \mapsto -1/z$ is also a line orthogonal to C at 0 and $-\overline{\xi}/s$. This is easiest to see with a diagram.



We conclude that if $s \neq 0$, then the image of C is a circle and 0 and $-\overline{\xi}/s$ are diametrically opposed. Therefore, the radius of C is $|\overline{\xi}/s|/2 = 1/(2|s|)$. Reasoning out where the interior must be, we see that actually $\kappa' = 2s$; in fact, this still holds true even if $\kappa' = 0$. We conclude that the point on C is $s\xi = \kappa'\xi/2$ and so $z = (\kappa'\xi)/2 - it\xi$ is a parametric equation for C, where $t \in \mathbb{R}$.

▶ **Example** *Suppose that C is a circle with positive orientation and tangent to the real line at the origin. Determine the possible inversive coordinates of C.*
If C is tangent to the real line at the origin, then its center must be of the form ti for some real $t \neq 0$, and its radius must be $|t|$. Therefore, the bend is $1/|t|$ and the bend-center is $\operatorname{sgn}(t)i$, where $\operatorname{sgn}(t)$ is the sign of t; that is, it is 1 if $t > 0$ and -1 if $t < 0$. It remains to determine the co-bend κ'. However, we know that

$$1 = -\kappa\kappa' + |\xi|^2 = -|t|\kappa' + 1,$$

so $\kappa' = 0$. We could also have seen this geometrically: since C passes through 0, its image under $z \mapsto 1/z$ is a line. Either way, the inversive coordinates are $(t, 0, \pm i)$ for some $t > 0$.

3.4 The Special Linear Group

Before we proceed to show how to use inversive coordinates for fast computations, we need to introduce another player, which might initially seem unrelated, but will eventually be very useful.

Definition 3.5 The *special linear group on* $\mathbb{C}^2$, denoted by $SL(2, \mathbb{C})$, consists of all 2×2 matrices with complex coefficients with determinant 1. That is, if

$$M = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL(2, \mathbb{C}),$$

then $\det M = ad - bc = 1$.

There are a few things we will need to know about the special linear group in order to proceed. To start with, it is actually a group.

Theorem 3.7 *The set $SL(2, \mathbb{C})$ is a group if we take matrix multiplication to be the operation.*

Proof In the course of the proof of Theorem 2.3, we showed that matrix multiplication is associative, the identity matrix satisfies the properties of an identity, and that $\det(M_1 M_2) = \det(M_1)\det(M_2)$ for any matrices M_1, M_2. Therefore, if $M_1, M_2 \in SL(2, \mathbb{C})$, then $\det(M_1 M_2) = \det(M_1)\det(M_2) = 1$, which is to say that $M_1 M_2 \in SL(2, \mathbb{C})$. The only thing that remains is to check that $SL(2, \mathbb{C})$ contains inverses. Indeed, if

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL(2, \mathbb{C}),$$

then

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \begin{pmatrix} d & -b \\ -c & a \end{pmatrix} \in SL(2, \mathbb{C})$$

since the determinant is $da - (-b)(-c) = ad - bc = 1$. □

Theorem 3.8 *Every $\varphi \in \text{Möb}^0(2)$ can be written in the form*

$$\varphi(z) = \frac{az + b}{cz + d}$$

for some $a, b, c, d \in \mathbb{C}$ such that

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL(2, \mathbb{C}).$$

Proof We know that any linear fractional transformation can be written in the form

$$\varphi(z) = \frac{\tilde{a}z + \tilde{b}}{\tilde{c}z + \tilde{d}}$$

such that

$$\tilde{M} = \begin{pmatrix} \tilde{a} & \tilde{b} \\ \tilde{c} & \tilde{d} \end{pmatrix} \in GL(2, \mathbb{C}).$$

However, since $\det(\tilde{M}) \neq 0$, there exists some $\lambda \in \mathbb{C}^\times$ such that $\lambda^2 = \det(\tilde{M})$. Define $a = \tilde{a}/\lambda$, $b = \tilde{b}/\lambda$, $c = \tilde{c}/\lambda$, $d = \tilde{d}/\lambda$. Then

$$\varphi(z) = \frac{\tilde{a}z + \tilde{b}}{\tilde{c}z + \tilde{d}} = \frac{az + b}{cz + d},$$

but on the other hand if

$$M = \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} \tilde{a} & \tilde{b} \\ \tilde{c} & \tilde{d} \end{pmatrix} \begin{pmatrix} 1/\lambda & 0 \\ 0 & 1/\lambda \end{pmatrix}$$

then $\det(M) = \det(\tilde{M})/\lambda^2 = 1$, so $M \in SL(2, \mathbb{C})$. □

There are many cases where it is more convenient to associate a matrix in $SL(2, \mathbb{C})$ to an element of φ rather than the more general case of a matrix in $GL(2, \mathbb{C})$, even if it requires a little extra work to renormalize the determinant to 1. One reason why this is nice is that there are infinitely matrices in $GL(2, \mathbb{C})$ that correspond to any single $\varphi \in \text{Möb}^0(2)$, but there are only two matrices in $SL(2, \mathbb{C})$ that correspond to any given φ.

Theorem 3.9 *Suppose $M_1, M_2 \in SL(2, \mathbb{C})$ are both matrices that correspond to the same linear fractional transformation under the map Ψ defined in Theorem 2.4. Then $M_1 = \pm M_2$.*

Proof Consider the matrix $M = M_1 M_2^{-1} \in SL(2, \mathbb{C})$—write this as

$$M = \begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

Since $\Psi(M_1) = \Psi(M_2)$, we know that $\Psi(M_1 M_2^{-1}) = \Psi(M_1)\Psi(M_2)^{-1} = \iota$, the identity function. Therefore, we know that

$$\frac{az + b}{cz + d} = z$$

for all $z \in \mathbb{C}P^1$. This immediately implies that $c = 0$—this is because $\iota(\infty) = \infty$ but $\Psi(M)(\infty) = a/c$ if $c \neq 0$. It also means that $b = 0$ since $\iota(0) = 0$ and $\Psi(M)(0) = b/d$. Ergo,

$$M = \begin{pmatrix} a & 0 \\ 0 & d \end{pmatrix} \in SL(2, \mathbb{C}),$$

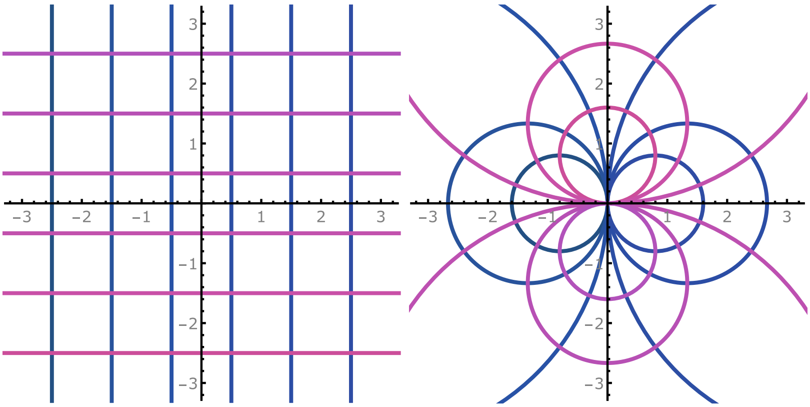


Fig. 3.8 The illustration on the right is the image of the circles on the left under the action of $\begin{pmatrix} 0 & 2 \\ -1/2 & 0 \end{pmatrix} \in SL(2, \mathbb{C})$.

but since $\det(M) = ad = 1$, we can simply take $a = 1/d$. This means that $\Psi(M)(z) = a^2 z$, and this is the same as $\imath$ if and only if $a = \pm 1$. Thus, $M = \pm I$, where I is the identity matrix. But this means that $M_1 M_2^{-1} = \pm I$, or $M_1 = \pm M_2$. □

Another benefit of $SL(2, \mathbb{C})$ is that there is a simple way that it moves around inversive coordinates. Specifically, choose any $\gamma \in SL(2, \mathbb{C})$ and any matrix $M \in \mathrm{Mat}(2, \mathbb{C})$ such that $\det(M) = -1$ and $\overline{M}^T = M$. If we consider $N = \gamma M \overline{\gamma}^T$, then we notice that

1. $\det(N) = \det(\gamma) \det(M) \det(\overline{\gamma}^T) = -1$ and
2. $\overline{N}^T = \overline{\overline{\gamma}^T}^T \, \overline{M}^T \, \overline{\gamma}^T = \gamma M \overline{\gamma}^T = N$,

where we have used the fact that the conjugate transpose does not change the determinant and that it reverses multiplication. (See Exercise 3.2.6.) This justifies the following definition.

Definition 3.6 Let C be an oriented circle and $\gamma \in SL(2, \mathbb{C})$. By $\gamma.C$, we shall denote the unique oriented circle C' such that $\mathrm{inv}(C') = \gamma\, \mathrm{inv}(C) \overline{\gamma}^T$.

This is a particular example of something known as a group action—a way that we can use a group to move around elements of some sets. To help illustrate what is going on, Figure 3.8 shows the effect of this particular action on a collection of lines, and how they get mapped to circles.

As most groups we have dealt with have been groups of transformations, group actions are not exactly new to us. (Although a formal definition and further examples are relegated to the exercises—see Exercise 3.3.4.) However, this is the first example that we have seen where there are ostensibly two different group actions on the same space: on the one hand, we know that elements of $SL(2, \mathbb{C})$ correspond to Möbius transformations which we know move around oriented circles; on the other hand, we just came up with this new action by which we can move around oriented circles by thinking about their inversive coordinates instead. As it happens, both of these actions are secretly one and the same.

Theorem 3.10 (Equivalence of Matrix and Linear Fractional Linear Actions) *Let C be an oriented circle and let $\gamma \in SL(2,\mathbb{C})$. Then $\gamma.C = \Psi(\gamma)(C)$, where Ψ is the usual map from $GL(2,\mathbb{C})$ to Möb$^0(2)$.*

Proof To simplify the proof, we first note that every element in $SL(2,\mathbb{C})$ that can be written as a product of matrices of the following types:

$$\begin{pmatrix} u & 0 \\ 0 & u^{-1} \end{pmatrix}, \begin{pmatrix} 1 & \tau \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

A proof of this can be easily adapted from the proof of Theorem 2.5, and is left as an exercise to the reader. (See Exercise 3.2.7.) Notice that if we can prove the theorem for these basic types of matrices, then we will in fact have proved it for all elements of $SL(2,\mathbb{C})$. Why is this? Well, it is easy to check that $\gamma_1.(\gamma_2.C) = (\gamma_1\gamma_2).C$. (See Exercise 3.3.7.) We already know that $\Psi(\gamma_1\gamma_2) = \Psi(\gamma_1)\circ\Psi(\gamma_2)$. Therefore, if we can write $\gamma = \gamma_1\gamma_2\ldots\gamma_n$ for some $\gamma_i \in SL(2,\mathbb{C})$ for which we know that $\gamma_i.C = \Psi(\gamma_i)(C)$, then it follows that

$$\begin{aligned}\gamma.C &= \gamma_1.\,(\gamma_2.\,(\ldots\gamma_n.C)\ldots) \\ &= \gamma_1.\,(\gamma_2.\,(\ldots\Psi(\gamma_n)(C))\ldots) \\ &= (\Psi(\gamma_1)\circ\ldots\circ\Psi(\gamma_n))\,(C) \\ &= \Psi(\gamma_1\gamma_2\ldots\gamma_n)(C) = \Psi(\gamma)(C).\end{aligned}$$

Now, let's look at what each of these basic matrices do, in turn. Let C be an oriented circle with inversive coordinates (κ,κ',ξ). First,

$$\begin{aligned}\begin{pmatrix} re^{i\theta} & 0 \\ 0 & \frac{1}{r}e^{-i\theta} \end{pmatrix}\begin{pmatrix} \kappa' & \xi \\ \overline{\xi} & \kappa \end{pmatrix}\begin{pmatrix} re^{-i\theta} & 0 \\ 0 & \frac{1}{r}e^{i\theta} \end{pmatrix} &= \begin{pmatrix} re^{i\theta}\kappa' & re^{i\theta}\xi \\ \frac{1}{r}e^{-i\theta}\overline{\xi} & \frac{1}{r}e^{-i\theta}\kappa \end{pmatrix}\begin{pmatrix} re^{-i\theta} & 0 \\ 0 & \frac{1}{r}e^{i\theta} \end{pmatrix} \\ &= \begin{pmatrix} r^2\kappa' & e^{2i\theta}\xi \\ e^{-2i\theta}\overline{\xi} & \frac{\kappa}{r^2} \end{pmatrix},\end{aligned}$$

so if

$$\gamma = \begin{pmatrix} re^{i\theta} & 0 \\ 0 & \frac{1}{r}e^{-i\theta} \end{pmatrix},$$

then $\operatorname{inv}(\gamma.C) = (\kappa/r^2, r^2\kappa', e^{2i\theta}\xi)$. On the other hand, $\Psi(\gamma)$ is the transformation $z \mapsto r^2e^{2i\theta}z$. Well, if C is a circle with bend κ and center ξ/κ, then its image under γ will be a circle with bend κ/r^2 and center $r^2e^{2i\theta}/\kappa$. If C is a line, then its image will be rotated by $e^{2i\theta}$ and so will its bend-center; the co-bend will be scaled by r^2. Therefore, $\gamma.C = \Psi(\gamma).C$ in this case. Next,

$$\begin{pmatrix}1&\tau\\0&1\end{pmatrix}\begin{pmatrix}\kappa'&\xi\\\overline{\xi}&\kappa\end{pmatrix}\begin{pmatrix}1&0\\\overline{\tau}&1\end{pmatrix}=\begin{pmatrix}\kappa'+\tau\overline{\xi}&\xi+\tau\kappa\\\overline{\xi}&\kappa\end{pmatrix}\begin{pmatrix}1&0\\\overline{\tau}&1\end{pmatrix}$$
$$=\begin{pmatrix}\kappa'+2\Re(\tau\overline{\xi})+|\tau|^2\kappa&\xi+\kappa\tau\\\overline{\xi}+\kappa\overline{\tau}&\kappa\end{pmatrix},$$

so if

$$\gamma=\begin{pmatrix}1&\tau\\0&1\end{pmatrix},$$

then $\mathrm{inv}(\gamma.C)=(\kappa,\kappa'+2\Re(\tau\overline{\xi})+|\tau|^2\kappa,\xi+\kappa\tau)$. On the other hand, $\Psi(\gamma)$ is the transformation $z\mapsto z+\tau$. If C is a circle with bend κ and center ξ/κ, then its image will be a circle with bend κ and center $\xi/\kappa+\tau=(\xi+\kappa\tau)/\kappa$. If C is a line in the direction $-i\xi$ and passing through the point $\kappa'\xi/2$, then its image will be a line in the direction $-i\xi$ and passing through the point $\kappa'\xi/2+\tau$. More relevantly, this line will pass through the point $\kappa'\xi/2+\Re(\tau\overline{\xi})\xi=(\kappa'+2\Re(\tau\overline{\xi}))\xi/2$ since $\Re(\tau\overline{\xi})\xi$ is the projection of τ onto the ray in the direction of ξ. In any case, we see that $\gamma.C=\Psi(\gamma).C$. Finally,

$$\begin{pmatrix}0&1\\-1&0\end{pmatrix}\begin{pmatrix}\kappa'&\xi\\\overline{\xi}&\kappa\end{pmatrix}\begin{pmatrix}0&-1\\1&0\end{pmatrix}=\begin{pmatrix}\overline{\xi}&\kappa\\-\kappa'&-\xi\end{pmatrix}\begin{pmatrix}0&-1\\1&0\end{pmatrix}$$
$$=\begin{pmatrix}\kappa&-\overline{\xi}\\-\xi&\kappa'\end{pmatrix},$$

so if

$$\gamma=\begin{pmatrix}0&1\\-1&0\end{pmatrix},$$

then $\mathrm{inv}(\gamma.C)=(\kappa',\kappa,-\overline{\xi})$. On the other hand, $\Psi(\gamma)$ is the transformation $z\mapsto -1/z$, which we know exchanges κ and κ' simply by the definition of the bend and co-bend. Since $-\kappa\kappa'+|\xi|^2=1$, we only need to determine the direction of the bend-center; it is not hard to see that if the bend-center of C is ξ, then the image of C must have bend-center in the direction of $-\overline{\xi}$. Thus, $\gamma.C=\Psi(\gamma).C$, and we are done. □

This is fantastic news: it means that if we know the inversive coordinates of an oriented circle, then it is easy to compute its image under any Möbius transformation.

▶ **Example** *Prove that if* $\gamma\in SL(2,\mathbb{C})$ *has real coefficients and* C *is the real line oriented so that* i *is in its interior, then* $\gamma.C=C$.
It is easy to check that the inversive coordinates of C are $(0,0,i)$. Therefore, the inversive coordinates of the image are given by

$$\gamma \begin{pmatrix} 0 & i \\ -i & 0 \end{pmatrix} \overline{\gamma}^T .$$

Writing

$$\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

for some $a, b, c, d \in \mathbb{R}$ such that $ad - bc = 1$, we calculate directly that this is the same as

$$\begin{pmatrix} 0 & (ad-bc)i \\ -(ad-bc)i & 0 \end{pmatrix} = \begin{pmatrix} 0 & i \\ -i & 0 \end{pmatrix},$$

which we note are just the inversive coordinates of C.

3.5 Inversive Distance

We need one final computational tool before we can revisit the examples we looked at earlier, and that is the inversive distance of oriented circles.

Definition 3.7 Let C_1, C_2 be oriented circles with inversive coordinates $(\kappa_1, \kappa_1', \xi_1)$ and $(\kappa_2, \kappa_2', \xi_2)$. Their *inversive distance* is

$$\langle C_1, C_2 \rangle_I = \frac{\kappa_1 \kappa_2' + \kappa_2 \kappa_1'}{2} - \Re(\xi_1 \overline{\xi_2}).$$

Confusingly, inversive distance can be negative. Another term that sometimes appears in the literature for this metric is “Pedoe product.” However, this seems to be a case of Stigler’s law of eponymy[1], since the notion of inversive distance was already discussed by Coxeter in 1966 [2] and is likely much older, whereas Daniel Pedoe didn’t write about it until 1970.

There are many equivalent ways to write the inversive distance. For example, if we write $\xi_1 = x_1 + y_1 i$ and $\xi_2 = x_2 + y_2 i$, then we can write the inversive distance as

$$\langle C_1, C_2 \rangle_I = \begin{pmatrix} \kappa_1 & \kappa_1' & x_1 & y_1 \end{pmatrix} \begin{pmatrix} 0 & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix} \begin{pmatrix} \kappa_2 \\ \kappa_2' \\ x_2 \\ y_2 \end{pmatrix},$$

which makes it easy to see that $\langle C_1, C_2 \rangle_I = \langle C_2, C_1 \rangle_I$ and that $\langle C_1, -C_2 \rangle = -\langle C_1, C_2 \rangle$. Another way to write the inversive distance is

[1] This was an observation by statistician Stephen Stigler: no scientific discovery is named after its discoverer. Stigler attributed this law to sociologist Robert Merton, but it is likely far older.

$$\langle C_1, C_2\rangle_I = \frac{1}{2}\mathrm{tr}\left(\begin{pmatrix}\kappa_1' & \xi_1\\ \overline{\xi_1} & \kappa_1\end{pmatrix}\begin{pmatrix}\kappa_2' & \xi_2\\ \overline{\xi_2} & \kappa_2\end{pmatrix}^{-1}\right)$$

where $\mathrm{tr}(M)$ denotes the trace of M—that is, if

$$M = \begin{pmatrix} a_{1,1} & a_{1,2} & \ldots & a_{1,n}\\ a_{2,1} & a_{2,2} & \ldots & a_{2,n}\\ \vdots & \vdots & \ddots & \vdots\\ a_{n,1} & a_{n,2} & \ldots & a_{n,n}\end{pmatrix}$$

then $\mathrm{tr}(M) = a_{1,1} + a_{2,2} + \ldots + a_{n,n}$.

Theorem 3.11 (Invariance of Inversive Distance) *Let C_1, C_2 be oriented circles and let $\gamma \in SL(2, \mathbb{C})$. Then $\langle C_1, C_2\rangle_I = \langle \gamma.C_1, \gamma.C_2\rangle_I$.*

Proof Notice that

$$\begin{aligned}\langle \gamma.C_1, \gamma.C_2\rangle_I &= \frac{1}{2}\mathrm{tr}\left(\gamma\begin{pmatrix}\kappa_1' & \xi_1\\ \overline{\xi_1} & \kappa_1\end{pmatrix}\overline{\gamma}^T\left(\gamma\begin{pmatrix}\kappa_2' & \xi_2\\ \overline{\xi_2} & \kappa_2\end{pmatrix}\overline{\gamma}^T\right)^{-1}\right)\\ &= \frac{1}{2}\mathrm{tr}\left(\gamma\begin{pmatrix}\kappa_1' & \xi_1\\ \overline{\xi_1} & \kappa_1\end{pmatrix}\overline{\gamma}^T\left(\overline{\gamma}^T\right)^{-1}\begin{pmatrix}\kappa_2' & \xi_2\\ \overline{\xi_2} & \kappa_2\end{pmatrix}^{-1}\gamma^{-1}\right)\\ &= \frac{1}{2}\mathrm{tr}\left(\gamma\begin{pmatrix}\kappa_1' & \xi_1\\ \overline{\xi_1} & \kappa_1\end{pmatrix}\begin{pmatrix}\kappa_2' & \xi_2\\ \overline{\xi_2} & \kappa_2\end{pmatrix}^{-1}\gamma^{-1}\right).\end{aligned}$$

Here, we use a general property of the trace: for any $n \times n$ matrix M and $G \in GL(n, \mathbb{C})$, $\mathrm{tr}(M) = \mathrm{tr}(GMG^{-1})$. (See Exercise 3.2.8.) Therefore,

$$\langle \gamma.C_1, \gamma.C_2\rangle_I = \frac{1}{2}\mathrm{tr}\left(\begin{pmatrix}\kappa_1' & \xi_1\\ \overline{\xi_1} & \kappa_1\end{pmatrix}\begin{pmatrix}\kappa_2' & \xi_2\\ \overline{\xi_2} & \kappa_2\end{pmatrix}^{-1}\right) = \langle C_1, C_2\rangle_I$$

as desired. □

The fact that the inversive distance is invariant under Möbius transformations certainly signals its importance. Even so, we would like to have a more geometric interpretation for what it actually means. As it happens, this is also possible.

Theorem 3.12 (Geometric Interpretation of Inversive Distance) *Let C_1, C_2 be oriented circles that are not lines, with positive orientation and radii r_1 and r_2. Let d be the distance between their centers. Then*

$$\langle C_1, C_2\rangle_I = \frac{d^2 - r_1^2 - r_2^2}{2r_1r_2}.$$

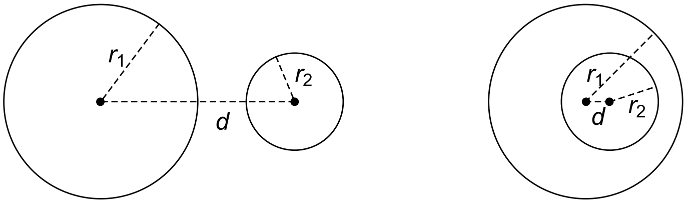


Fig. 3.9 Two possible ways that two circles can fail to intersect.

Proof Let $(\kappa_1, \kappa_1', \xi_1)$, $(\kappa_2, \kappa_2', \xi_2)$ be the inversive coordinates of C_1 and C_2. Then $r_1 = 1/\kappa_1$, $r_2 = 1/\kappa_2$, and $d = |\xi_1/\kappa_1 - \xi_2/\kappa_2|$. Therefore,

$$\begin{aligned}\frac{d^2 - r_1^2 - r_2^2}{2r_1r_2} &= \frac{|\xi_1/\kappa_1 - \xi_2/\kappa_2|^2 - 1/\kappa_1^2 - 1/\kappa_2^2}{2/(\kappa_1\kappa_2)} \\ &= \frac{|\kappa_2\xi_1 - \kappa_1\xi_1|^2 - \kappa_1^2 - \kappa_2^2}{2\kappa_1\kappa_2} \\ &= \frac{-\kappa_1^2 - \kappa_2^2 + \kappa_2^2|\xi_1|^2 + \kappa_1^2|\xi_2|^2 - 2\kappa_1\kappa_2\Re(\xi_1\overline{\xi_2})}{2\kappa_1\kappa_2} \\ &= \frac{\kappa_1^2(|\xi_2|^2 - 1) + \kappa_2^2(|\xi_1|^2 - 1)}{2\kappa_1\kappa_2} - \Re(\xi_1\overline{\xi_2}) \\ &= \frac{\kappa_1^2(\kappa_2\kappa_2') + \kappa_2^2(\kappa_1\kappa_1')}{2\kappa_1\kappa_2} - \Re(\xi_1\overline{\xi_2}) \\ &= \frac{\kappa_1\kappa_2' + \kappa_2\kappa_1'}{2} - \Re(\xi_1\overline{\xi_2}) = \langle C_1, C_2\rangle_I,\end{aligned}$$

as was claimed. □

Corollary 3.3 *Two oriented circles C_1, C_2 intersect if and only if $|\langle C_1, C_2\rangle_I| \leq 1$.*

Proof The inversive distance is invariant under Möbius transformations, so we can assume without loss of generality that neither of C_1, C_2 is a line. Furthermore, since $|\langle C_1, -C_2\rangle| = |\langle C_1, C_2\rangle| = |\langle -C_1, C_2\rangle|$, we may assume that both C_1 and C_2 have positive orientation. Since $\langle C_1, C_2\rangle_I = \langle C_2, C_1\rangle$, we can assume that the radius of C_1 is at least as large as the radius of C_2. In that case, notice that these two circles intersect if and only if $r_1 - r_2 \leq d \leq r_1 + r_2$. The two different ways that circles can fail to intersect are shown in Figure 3.9 to help illustrate this. In any case, this implies that

$$\frac{d^2 - r_1^2 - r_2^2}{2r_1r_2} \leq \frac{(r_1 + r_2)^2 - r_1^2 - r_2^2}{2r_1r_2} = \frac{r_1^2 + 2r_1r_2 + r_2^2 - r_1^2 - r_2^2}{2r_1r_2} = 1$$

and

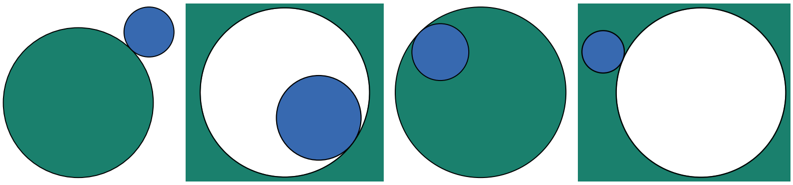


Fig. 3.10 Two pairs of externally tangent circles and two pairs of internally tangent circles.

$$\frac{d^2 - r_1^2 - r_2^2}{2r_1r_2} \geq \frac{(r_1 - r_2)^2 - r_1^2 - r_2^2}{2r_1r_2} = \frac{r_1^2 - 2r_1r_2 + r_2^2 - r_1^2 - r_2^2}{2r_1r_2} = -1,$$

which can be summarized, by appealing to the geometric interpretation of the inversive distance, as $|\langle C_1, C_2\rangle_I| \leq 1$. □

We can say more.

Theorem 3.13 (Inversive Distance Angle Formula) *Let C_1, C_2 be two oriented circles. They intersect if and only if $|\langle C_1, C_2\rangle_I| \leq 1$ and if they do, then $|\langle C_1, C_2\rangle_I| = |\cos(\phi)|$, where ϕ is the angle between them.*

Proof Since angles, interiors, and the inversive distance are all preserved by Möbius transformations, we can reduce to the simple case where C_1 is the real line traversed from left to right and C_2 is a line through the origin. In order for the angle between C_1 and C_2 to be ϕ, C_2 must be traversed in the direction $e^{\pm i\phi}$. Then

$$\begin{aligned} |\langle C_1, C_2\rangle_I| &= \left|\frac{\kappa_1\kappa_2' + \kappa_2\kappa_1'}{2} - \Re(\xi_1\overline{\xi_2})\right| = \left|\Re(i \cdot ie^{\pm i\phi})\right| \\ &= \left|\Re(e^{\pm i\phi})\right| = |\cos(\phi)|, \end{aligned}$$

finishing the proof. □

Remark 3.5 Henceforth, we shall simply take the convention that the angle ϕ between two intersecting oriented circles is the unique real number $-\pi < \phi \leq \pi$ such that $\langle C_1, C_2\rangle_I = \cos(\phi)$.

Definition 3.8 We say that two oriented circles C_1, C_2 are *externally tangent* if either they intersect at a single point and their interiors do not intersect. We say that two oriented circles are *internally tangent* if either they intersect at a single point and their interiors intersect.

Figure 3.10 shows both internally and externally tangent circles.

Corollary 3.4 *Let C_1, C_2 be two oriented circles.*

1. *$\langle C_1, C_2\rangle_I = 1$ if and only if C_1, C_2 are externally tangent or $C_1 = -C_2$.*
2. *$\langle C_1, C_2\rangle_I = -1$ if and only if C_1, C_2 are internally tangent or $C_1 = C_2$.*
3. *$\langle C_1, C_2\rangle_I = 0$ if and only if C_1, C_2 are orthogonal to each other.*

Proof The proof is left as an exercise to the reader. (See Exercise 3.2.9.) □

▶ **Example** *Determine whether the two circles with centers* $1+i$ *and* $3-4i$ *and radii* 2 *and* 5 *intersect. If they do, determine the angle of intersection.*
The square of the distance between the centers is

$$d^2 = |(3-4i-(1+i)|^2 = |2-5i|^2 = 4+25 = 29.$$

After this, we simply compute the inversive distance.

$$\langle C_1, C_2\rangle_I = \frac{d^2 - r_1^2 - r_2^2}{2r_1r_2} = \frac{29-4-25}{2\cdot 2\cdot 5} = 0.$$

Thus, the circles don't just intersect—they are orthogonal to one another, which is to say that the angle of intersection is $\pi/2$.

3.6 Steiner's Porism Revisited

We are finally ready to look at the two examples of applications of inversive geometry a little more closely. We shall start with Steiner's porism. Previously, we managed to prove that whether or not a Steiner chain is open or closed does not depend on the choice of starting point used in the construction. However, we did not give any simple criterion for determining whether a Steiner chain is open or closed. We shall now rectify this.

Theorem 3.14 *Let* C_1, C_2 *be two non-intersecting generalized circles. The Steiner chain that they define is closed and contains n circles other than the initial two if and only if*

$$|\langle C_1, C_2\rangle_I| = 2\sec(\pi/n)^2 - 1.$$

Proof We already know that any Steiner chain can be reduced via Möbius transformations to the case where both circles are concentric at the origin, as in Figure 3.11. The inversive distance is invariant under Möbius transformations, so if the theorem is true in that case, then it must be true in the general case. But

$$|\langle C_1, C_2\rangle_I| = \frac{|d^2 - r_1^2 - r_2^2|}{2r_1r_2} = \frac{r_1^2 + r_2^2}{2r_1r_2},$$

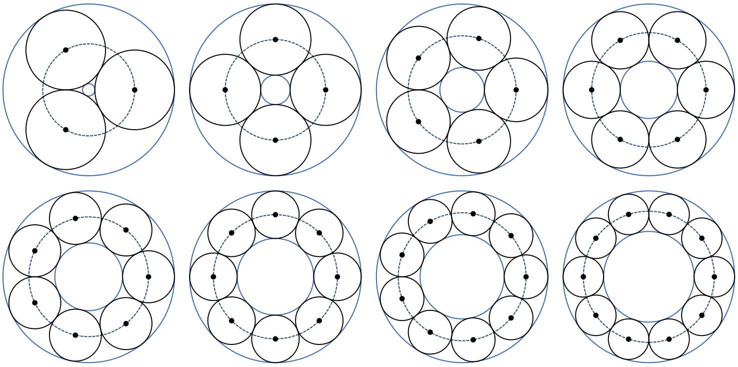


Fig. 3.11 Steiner chains of various lengths, in the standard configuration where the constructed circles all have centers on the unit circle.

and if we rescale so that the inner circle has radius 1, then this further simplifies to

$$\frac{1}{2}\left(r_2 + \frac{1}{r_2}\right).$$

There can be at most one radius r_2 such that the resulting Steiner chain is closed with n circles. On the other hand, the function

$$\begin{aligned} (1, \infty) &\to (1, \infty) \\ x &\mapsto \frac{1}{2}\left(x + \frac{1}{x}\right) \end{aligned}$$

is bijective—its inverse is $x \mapsto x + \sqrt{x^2 - 1}$. Therefore, there is at most one value for the inversive distance such that the resulting Steiner chain is closed with n circles. It remains to compute the inversive distance for the circles in such a chain. Such a chain is easy enough to construct. We take the circles in the chain to have centers $e^{\frac{2\pi ki}{n}}$ for $k = 0, 1, 2, \ldots n - 1$ so that they are equidistant around the unit circle; for them to be tangent, their radii have to half the distance between these centers, or

$$\begin{aligned} \frac{1}{2}\left|1 - e^{\frac{2\pi i}{n}}\right| &= \frac{1}{2}\sqrt{2 - 2\Re\left(e^{\frac{2\pi i}{n}}\right)} = \frac{1}{2}\sqrt{2 - 2\sin\left(\frac{2\pi}{n}\right)} \\ &= \frac{1}{2}\sqrt{4\sin\left(\frac{\pi}{n}\right)^2} = \sin\left(\frac{\pi}{n}\right). \end{aligned}$$

There are two circles with centers at the origin that are tangent to all of these circles—one has radius $1 - \sin(\pi/n)$ and the other has radius $1 + \sin(\pi/n)$. The desired invariant is thus

$$\begin{aligned}|\langle C_1, C_2\rangle_I| &= \frac{r_1^2+r_2^2}{2r_1r_2} = \frac{(1-\sin(\pi/n))^2+(1+\sin(\pi/n))^2}{2(1-\sin(\pi/n))(1+\sin(\pi/n))}\\ &= \frac{2+2\sin(\pi/n)^2}{2(1-\sin(\pi/n)^2)}\\ &= \frac{2-\cos(\pi/n)^2}{\cos(\pi/n)^2}\\ &= 2\sec(\pi/n)^2-1,\end{aligned}$$

precisely as claimed. □

Corollary 3.5 *Let C_1, C_2 be two generalized circles. They define a closed Steiner chain if and only if*

$$\frac{\pi}{\arccos\left(\sqrt{\frac{2}{1+|\langle C_1,C_2\rangle_I|}}\right)}$$

is an integer.

Proof We know that C_1 and C_2 intersect if and only if $|\langle C_1, C_2\rangle| \leq 1$. However, for non-negative x,

$$\arccos\left(\sqrt{\frac{2}{1+x}}\right)^{-1}$$

is defined (or, at least, a real number) if and only if $x > 1$. In this same range, it is easy to check that

$$x \mapsto \frac{\pi}{\arccos\left(\sqrt{\frac{2}{1+x}}\right)}$$

is the inverse function to $x \mapsto 2\sec(\pi/x)^2 - 1$, and so the claim is an immediate consequence of Theorem 3.14. □

One possible way to generalize closed Steiner chains is to allow the constructed circles to intersect, rather than halting once it becomes impossible to add another non-intersecting circle. This gives analogs of Steiner chains that "wrap around" in some sense, as in Figure 3.12. Sometimes, these chains keep wrapping forever. Sometimes, they eventually overlap on top of themselves. I leave determining when both cases occur as an interesting problem for the reader.

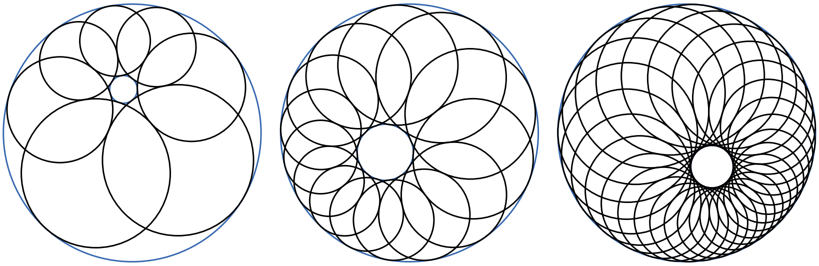


Fig. 3.12 Some examples of generalized Steiner chains.

3.7 Apollonian Gaskets Revisited

When we originally defined the Apollonian gasket, we did not worry about the orientation of the starting four circles. This is inconvenient, but the good news is that it is easy to see that, given any Descartes configuration, there is only one way to choose an orientation on the four generalized circles such that their interiors do not intersect. Furthermore, when we do this, the four oriented circles are then externally tangent and so we get that if C_1, C_2, C_3, C_4 are the four oriented circles, it must be that

$$\langle C_i, C_j \rangle_I = \begin{cases} 1 & \text{if } i \neq j \\ -1 & \text{if } i = j. \end{cases}$$

With this in mind, we make the following definition.

Definition 3.9 An *oriented Descartes configuration* is a quadruple of oriented circles that are all externally tangent to one another.

It is easy to see that we have shown the following.

Lemma 3.4 *Four oriented circles C_1, C_2, C_3, C_4 form an oriented Descartes configuration if and only if*

$$\langle C_i, C_j \rangle_I = \begin{cases} 1 & \textit{if } i \neq j \\ -1 & \textit{if } i = j. \end{cases}$$

We already know that we can use Descartes swaps to get new Descartes configurations from old ones. In fact, it is easy to see that they generated oriented Descartes configurations from existing ones. As a consequence, it is natural to make the following definition.

Definition 3.10 Let C_1, C_2, C_3, C_4 be an oriented Descartes configuration. The *oriented Apollonian gasket with starting configuration* C_1, C_2, C_3, C_4 is the smallest set $\hat{S}$ of oriented circles in $\mathbb{C}P^1$ such that

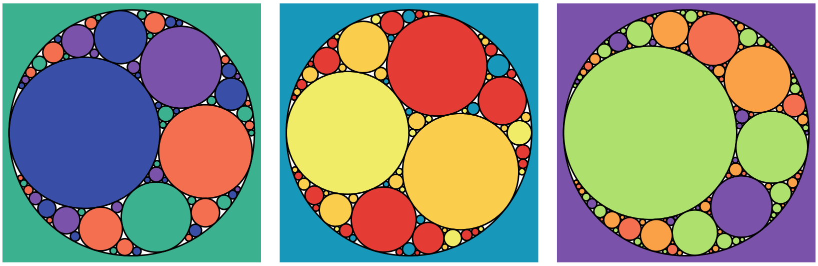


Fig. 3.13 A selection of oriented Apollonian gaskets.

1. $\hat{\mathcal{S}}$ contains C_1, C_2, C_3, C_4 and
2. if C_1', C_2', C_3', C_4' are all oriented circles in $\hat{\mathcal{S}}$ that form an oriented Descartes configuration, then all of the Descartes swaps of these circles are also in $\hat{\mathcal{S}}$.

Some examples are drawn in Figure 3.13. The oriented Apollonian gasket differs from our previous definition exclusively in that all we have done is added an orientation to all of the circles in the collection. However, this will be easier to show once we have found a more convenient way to describe the oriented Apollonian gasket. We begin by defining a nice subset of the gasket.

Lemma 3.5 *Let A be the oriented Apollonian gasket with initial configuration C_1, C_2, C_3, C_4. Let $\mathcal{G}_A$ be the Apollonian group. Define*

$$\mathcal{P}_A = \left\{ \gamma(C) \middle| \gamma \in \mathcal{G}_A,\ C \in \{C_1, C_2, C_3, C_4\} \right\}.$$

Then $\mathcal{P}_A \subset A$.

Proof Choose any $g \in \mathcal{G}_A$. Let γ_i be the reflection through the dual circle that swaps out C_i. We can write $g = \gamma_{i_n} \circ \gamma_{i_{n-1}} \circ \ldots \circ \gamma_{i_1}$ for some $1 \leq i_1, i_2, \ldots i_n \leq 4$. Define $g_{k,l} = \gamma_{i_k} \circ \gamma_{i_{n-1}} \circ \ldots \circ \gamma_{i_l}$. Notice that $(g_{k,1}(C_1), g_{k,1}(C_2), g_{k,1}(C_3), g_{k,1}(C_4))$ is the Descartes swap of $(g_{k,2}(C_1), g_{k,2}(C_2), g_{k,2}(C_3), g_{k,2}(C_4))$—this is because the image under a linear fractional transformation of a Descartes swap is a Descartes swap. Now, we shall prove by induction that for any element $g \in \mathcal{G}_A$ that can be written as a composition of no more than k of the γ_i, $g(\{C_1, C_2, C_3, C_4\}) \in A$. Indeed, if $k = 0$, this is obvious, since $g = id$. Otherwise, assume it is true for $k - 1$—we already saw that $(g_{k,1}(C_1), g_{k,1}(C_2), g_{k,1}(C_3), g_{k,1}(C_4))$ is the Descartes swap of $(g_{k,2}(C_1), g_{k,2}(C_2), g_{k,2}(C_3), g_{k,2}(C_4))$. However, $g_{k,2}$ can be written as a composition of $k-1$ γ_i's, hence $\{g_{k,2}(C_1), g_{k,2}(C_2), g_{k,2}(C_3), g_{k,2}(C_4)\} \subset A$. However, since A is closed under Descartes swaps, it follows that $g(C_1), g(C_2), g(C_3), g(C_4) \in A$ as well. We conclude that $\mathcal{P}_A \subset A$. □

In actuality, $A = \mathcal{P}_A$. However, to prove this, we shall need a few lemmas.

Lemma 3.6 *Let C_1 be the line $y = 0$ traversed from right to left, C_2 be the line $y = 1$ traversed from left to right, C_3 be the circle $x^2 + (y/2)^2 = 1/4$ oriented counter-clockwise, and C_4 be the circle $(x-1)^2 + (y/2)^2 = 1/4$ oriented counter-*

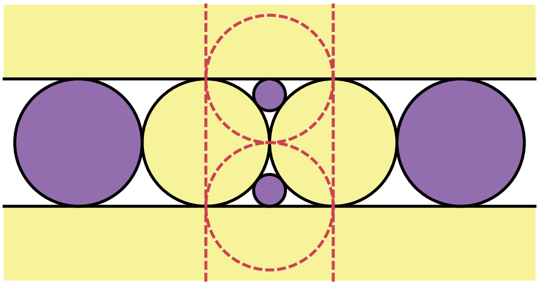


Fig. 3.14 The standard oriented Descartes configuration is depicted in yellow. The Descartes swaps are drawn in purple. The dual circles are drawn in red.

clockwise. Let A be the oriented Apollonian gasket with initial configuration C_1, C_2, C_3, C_4. *Let* $\mathcal{P}_A$ *be defined as in Lemma 3.5. Then*

$$\mathcal{P}_A \subset \left\{ \pm\gamma.C_1 \middle| \gamma \in SL(2, \mathbb{Z}[i]) \right\}$$

where

$$SL(2, \mathbb{Z}[i]) = \left\{ \begin{pmatrix} a_0 + a_1 i & b_0 + b_1 i \\ c_0 + c_1 i & d_0 + d_1 i \end{pmatrix} \in SL(2, \mathbb{C}) \middle| a_0, a_1, b_0, b_1, c_0, c_1, d_0, d_1 \in \mathbb{Z} \right\}.$$

Remark 3.6 The set $\mathbb{Z}[i]$ is known as the *Gaussian integers*—defined as the set of complex numbers $a + bi$ where $a, b \in \mathbb{Z}$—and is of fundamental importance in elementary number theory. We will not pursue this further in this text, but it is a common feature of number theory books such as Rosen's [13].

Proof This configuration is illustrated in Figure 3.14. The reflections through the dual circles are

$$g_1(z) = \frac{(2-i)\overline{z} + 2i}{-2i\overline{z} + 2 + i} \quad g_2(z) = \frac{i}{2i\overline{z} - i}$$

$$g_3(z) = \frac{i\overline{z} - 2i}{-i} \qquad g_4(z) = \frac{i\overline{z}}{-i}.$$

This way of writing them might seem odd—it is chosen to demonstrate that $g_i(z) = \Psi(\gamma_i) \circ \text{conj}$ for

$$\gamma_1 = \begin{pmatrix} 2-i & 2i \\ -2i & 2+i \end{pmatrix} \quad \gamma_2 = \begin{pmatrix} i & 0 \\ 2i & -i \end{pmatrix}$$
$$\gamma_3 = \begin{pmatrix} i & -2i \\ 0 & -i \end{pmatrix} \qquad \gamma_4 = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix},$$

where each γ_i is in $SL(2, \mathbb{C})$. In fact, something stronger is true—each γ_i is an element of the set

$$SL(2,\mathbb{Z}[i]) = \left\{ \begin{pmatrix} a_0 + a_1 i & b_0 + b_1 i \\ c_0 + c_1 i & d_0 + d_1 i \end{pmatrix} \in SL(2,\mathbb{C}) \middle| a_0, a_1, b_0, b_1, c_0, c_1, d_0, d_1 \in \mathbb{Z} \right\}.$$

This set is a group (see Exercise 3.3.2), but really we only need that it is closed under matrix multiplication, which is very easy to check, and that if $\gamma \in SL(2,\mathbb{Z}[i])$ then there exists some $\tilde{\gamma} \in SL(2,\mathbb{Z}[i])$ such that

$$\mathrm{conj} \circ \Psi(\gamma) = \Psi(\tilde{\gamma}) \circ \mathrm{conj}$$

which is also very easy to check. One final thing to note is that for $j = 2, 3, 4$, $\tilde{\gamma}_j.C_1 = -C_j$ if we take

$$\tilde{\gamma}_2 = \begin{pmatrix} 1 & i \\ 0 & 1 \end{pmatrix} \; \tilde{\gamma}_3 = \begin{pmatrix} 0 & i \\ i & 1 \end{pmatrix} \; \tilde{\gamma}_4 = \begin{pmatrix} i & 1+i \\ i & 1 \end{pmatrix}.$$

Note that $\tilde{\gamma}_2, \tilde{\gamma}_3, \tilde{\gamma}_4 \in SL(2,\mathbb{Z}[i])$ as well. By these observations and the definition of $\mathcal{P}_A$, we conclude that every element in it can be written in the form $\pm\gamma.C_1$ for some $\gamma \in SL(2,\mathbb{Z}[i])$. □

Lemma 3.7 *Let A be an oriented Apollonian gasket with initial configuration C_1, C_2, C_3, C_4, let $\mathcal{G}_A$ be the Apollonian group, and let $\mathcal{P}_A$ be defined as in Lemma 3.5. If $D_1, D_2 \in \mathcal{P}_A$ intersect, then they are tangent.*

Proof Since linear fractional transformations preserve oriented circles and tangencies, we can assume that the initial configuration is the standard one defined in Lemma 3.6. Thus, we know that $D_1 = \pm\gamma_1.C_1$ and $D_2 = \pm\gamma_2.C_2$ for some $\gamma_1, \gamma_2 \in SL(2,\mathbb{Z}[i])$. By Corollary 3.3, we know that D_1 and D_2 intersect if and only if $|\langle D_1, D_2\rangle| \leq 1$. But by the invariance of the inversive distance, we see that

$$|\langle D_1, D_2\rangle| = |\langle \gamma_1.C_1, \gamma_2.C_2\rangle| = \left|\langle C_1, \gamma_1^{-1}\gamma_2.C_2\rangle\right|.$$

Let

$$\gamma = \gamma_1^{-1}\gamma_2 = \begin{pmatrix} a_0 + a_1 i & b_0 + b_1 i \\ c_0 + c_1 i & d_0 + d_1 i \end{pmatrix}.$$

Noting that $\mathrm{inv}(C_1) = (0, 0, -i)$, we see that

$$\begin{aligned} &|\langle C_1, \gamma_1^{-1}\gamma_2.C_2\rangle| \\ &\quad = \frac{1}{2}\mathrm{tr}\left(\gamma \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} \overline{\gamma}^T \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}^{-1}\right) \\ &\quad = \frac{1}{2}\mathrm{tr}\left(\begin{pmatrix} a_0 + a_1 i & b_0 + b_1 i \\ c_0 + c_1 i & d_0 + d_1 i \end{pmatrix} \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} \begin{pmatrix} a_0 - a_1 i & c_0 - c_1 i \\ b_0 - b_1 i & d_0 - d_1 i \end{pmatrix} \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}^{-1}\right) \\ &\quad = |a_0 d_0 + a_1 d_1 - b_0 c_0 - b_1 c_1|. \end{aligned}$$

We can actually say a little bit more, because we know that by definition,

$$\begin{aligned} 1 &= \det\left(\begin{pmatrix} a_0 + a_1 i & b_0 + b_1 i \\ c_0 + c_1 i & d_0 + d_1 i \end{pmatrix}\right) \\ &= a_0 d_0 - a_1 d_1 - b_0 c_0 + b_1 c_1 + i\,(a_1 d_0 + a_0 d_1 - b_1 c_0 - b_0 c_1) \end{aligned}$$

whence

$$\begin{aligned} |\langle D_1, D_2\rangle| &= |a_0 d_0 + a_1 d_1 - b_0 c_0 - b_1 c_1| \\ &= |a_0 d_0 + a_1 d_1 - b_0 c_0 - b_1 c_1 + 1 - (a_0 d_0 - a_1 d_1 - b_0 c_0 + b_1 c_1)\,| \\ &= |1 + 2(a_1 d_1 - b_1 c_1)|\,. \end{aligned}$$

However, $a_1 d_1 - b_1 c_1$ is some integer, so this implies that $|\langle D_1, D_2\rangle|$ is a positive odd integer, which means that if $|\langle D_1, D_2\rangle| \leq 1$ then $|\langle D_1, D_2\rangle| = 1$. □

Lemma 3.8 *Let A be an oriented Apollonian gasket with initial configuration C_1, C_2, C_3, C_4, let $\mathcal{G}_A$ be the Apollonian group, and let $\mathcal{P}_A$ be defined as in Lemma 3.5. Let D_1, D_2, D_3, D_4 be a Descartes configuration of circles in $\mathcal{P}_A$. Then there exists an element $\gamma \in \mathcal{G}_A$ such that $\{D_1, D_2, D_3, D_4\} = \{\gamma(C_1), \gamma(C_2), \gamma(C_3), \gamma(C_4)\}$.*

Remark 3.7 Note that the statement of the lemma refers to Descartes configurations, rather than oriented Descartes configurations. This is not a mistake. One consequence of this lemma is that all Descartes configurations inside $\mathcal{P}_A$ are oriented Descartes configurations automatically. An illustration of how to reduce an arbitrary Descartes quadruple to the base one is shown in Figure 3.15.

Proof Let g_i be the reflection through the dual circle defined by the fact that $g_i(C_j) = C_j$ if $i \neq j$—by definition, we know that every element in $\mathcal{G}_A$ can be written as a product of these g_is. Furthermore, by the definition of A, $D_1 = \gamma_1 C_i$ for some $\gamma_1 \in \mathcal{G}_A$ and $i \in \{1, 2, 3, 4\}$—without loss of generality, we shall assume that $D_1 = \gamma_1 C_1$. There will be many different choices for $\gamma \in \mathcal{G}_A$ such that $D_1 = \gamma C_1$; choose γ_1 so that if we write $\gamma_1^{-1} D_2 = \gamma_2 C_j$ from some $\gamma_2 \in \mathcal{G}_A$ and some $j \in \{1, 2, 3, 4\}$, then γ_2 can be written as a product of the g_i's in the shortest possible way. We claim that, in fact, γ_2 is the identity.

Suppose not, and write $\gamma_2 = g_{i_1} g_{i_2} \ldots g_{i_k}$ for some integers $i_1, i_2, \ldots i_k$ in the shortest possible way. Notice that if $i_1 \neq 1$, then if we took $\gamma_1' = \gamma_1 g_{i_1}$, we would have $\gamma_1'^{-1}(D_1) = g_{i_1}^{-1}(\gamma_1^{-1}(D_1)) = g_{i_1}^{-1}(C_1) = C_1$ and $\gamma_1'^{-1}(D_2) = g_{i_1}^{-1}\gamma_1^{-1}(D_2) = g_{i_2} \ldots g_{i_k}(C_j)$. Due to the way that γ_1 was chosen, this is impossible; hence, $i_1 = 1$. What is more, we can see that $i_1 \neq i_2$, $i_2 \neq i_3$, and so on—this is because otherwise we could cancel out the corresponding swaps and get a shorter way of expressing γ_2. Finally, it must be that $i_k = j$—otherwise, we could write $\gamma_2^{-1}(D_2) = g_{i_1} g_{i_2} \cdots g_{i_{k-1}}(C_j)$ since $g_l(C_j) = C_j$ if $l \neq j$. Now, g_{i_k} will move the interior of C_j into the interior of the j-th dual circle; $\gamma_{i_{k-1}}$ will move it into the interior of the $i_{k-1} - th$ dual circle, and so on. The last inversion will be g_1, moving this set into the interior of the first dual circle. Thus, $\gamma_1^{-1} D_2$ must be contained in the interior of the first dual circle. We know that D_1 is tangent to D_2, so it must be that $\gamma_1^{-1} D_1 = C_1$ is tangent to D_2. But C_1 does not intersect the first dual circle so it cannot possibly

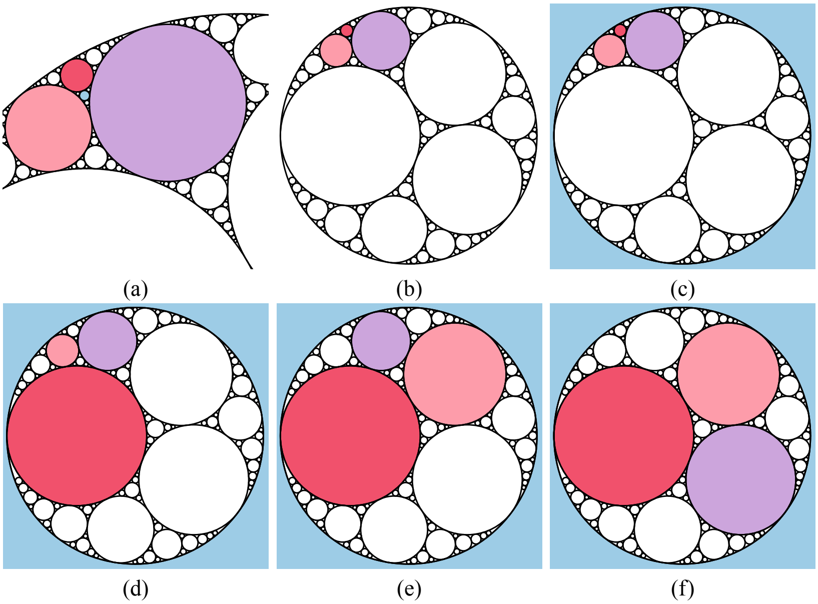


Fig. 3.15 (a) shows a Descartes configuration inside an Apollonian gasket in close up; (b) shows the same, but zoomed out. In (c), the smallest circle in the configuration is moved to the largest circle in the gasket. In (d), this move is chosen such that one of the other circles in the configuration is moved to one in the base quadruple. (e) and (f) show how the remaining two circles are moved into their correct positions.

be tangent to a circle contained in its interior. This is a contradiction, and so we conclude that indeed γ_2 is the identity.

Since γ_2 is the identity, without loss of generality, we can assume that $\gamma_1^{-1}(D_2) = C_2$. Furthermore, for simplicity, we can take C_1 to be the real line traversed from right to left and C_2 to be the line $y = 1$ traversed from left to right. Then $\gamma_1^{-1}(D_3)$ and $\gamma_1^{-1}(D_4)$ must be circles tangent to both of these lines. This implies that they must be two circles with radii $1/2$ and centers $x_0 + i/2$ and $x_0 + 1 + i/2$. However, it is easy to see that $\mathcal{P}_A$ contains all such circles with x_0 an integer, and so $\gamma_1^{-1}(D_3)$ and $\gamma_1^{-1}(D_4)$ will have to intersect them. By Lemma 3.7, they must actually be tangent to them, which can only happen if x_0 is an integer. Thus, by applying g_3 and g_4, we can move $\gamma_1^{-1}(D_3)$ and $\gamma_1^{-1}(D_4)$ onto C_3 and C_4. Without loss of generality, we can assume that $\gamma_1^{-1}(D_3) = \gamma_2(C_3)$ and $\gamma_1^{-1}(D_4) = \gamma_2(C_4)$ for some γ_2 which is a composition of g_3's and $g_4's$. Then if we take $\gamma = \gamma_1\gamma_2$, $D_i = \gamma_i(C_i)$ for $i = 1, 2, 3, 4$, as desired. □

We now fully understand the circumstances under which a Descartes configuration can appear inside the Apollonian gasket. As an immediate consequence, we get the following theorem which describes in detail what the gasket looks like and what properties it has.

Theorem 3.15 (The Structure Theorem for Apollonian Gaskets) *Let A be an oriented Apollonian gasket with initial configuration C_1, C_2, C_3, C_4. Let $\mathcal{G}_A$ be the Apollonian group. Then*

$$A = \mathcal{P}_A = \left\{ \gamma(C) \middle| \gamma \in \mathcal{G}_A,\ C \in \{C_1, C_2, C_3, C_4\} \right\}.$$

Furthermore, it satisfies the following properties:

1. *Every Descartes configuration in A is the image of $\{C_1, C_2, C_3, C_4\}$ under some element $\gamma \in \mathcal{G}_A$.*
2. *If we forget about orientation, then the circles in A are the Apollonian gasket with initial configuration C_1, C_2, C_3, C_4.*
3. *For every pair of circles $D_1, D_2 \in A$, if D_1 and D_2 intersect then either $D_1 = D_2$ or they are externally tangent.*
4. *If D_1, D_2, D_3, D_4 is an oriented Descartes configuration in A, then A is also the oriented Apollonian gasket with initial configuration D_1, D_2, D_3, D_4.*

Proof By Lemma 3.5, A is contained in the given set $\mathcal{P}_A$. By Lemma 3.8, every Descartes quadruple in $\mathcal{P}_A$ can be obtained as $\{\gamma(C_1), \gamma(C_2), \gamma(C_3), \gamma(C_4)\}$ for some $\gamma \in \mathcal{G}_A$. But this means that $\mathcal{P}_A$ also contains $\{\gamma(C_1), \gamma(C_2), \gamma(C_3), \gamma(C_4')\}$ where C_4' is the Descartes swap of C_4—from this, we see that $\mathcal{P}_A$ contains all Descartes swaps of circles in $\mathcal{P}_A$. By the definition of A as the smallest set containing the Descartes swaps, $A = \mathcal{P}_A$. It immediately follows that every Descartes configuration in A is the image of $\{C_1, C_2, C_3, C_4\}$. It is clear that A is contained inside the Apollonian gasket with initial configuration C_1, C_2, C_3, C_4—however, since it is closed under Descartes swaps of all Descartes configurations, these two sets must actually be equal since the Apollonian gasket is defined to be the smallest such set. Finally, choose two circles $D_1, D_2 \in A$ which intersect. By Lemma 3.7, D_1 is tangent to D_2. As in the proof of Lemma 3.8, we can find an element $\gamma \in \mathcal{G}_A$ such that $\gamma(D_1), \gamma(D_2)$ are in $\{C_1, C_2, C_3, C_4\}$. For the circles in the initial configuration, they are either equal or externally tangent, so this must be true of D_1 and D_2 as well. For the last part, by Lemma 3.8, there exists some $g \in \mathcal{G}_A$ such that

$$\{C_1, C_2, C_3, C_4\} = \{g^{-1}(D_1), g^{-1}(D_2), g^{-1}(D_3), g^{-1}(D_4)\}.$$

Let $\gamma_1, \gamma_2, \gamma_3, \gamma_4$ be the Descartes swaps of C_1, C_2, C_3, C_4. Then it isn't hard to see that $g \circ \gamma_i \circ g^{-1}$ will be the Descartes swaps of D_1, D_2, D_3, D_4. As the map

$$\begin{aligned} \mathcal{G}_A &\to \mathcal{G}_A \\ \gamma &\mapsto g \circ \gamma \circ g^{-1} \end{aligned}$$

is a bijection, we see that actually $\mathcal{G}_A$ can also be described as the smallest subgroup Möb(2) containing the Descartes swaps of D_1, D_2, D_3, D_4. This implies that C_1, C_2, C_3, C_4 are contained in the oriented Apollonian gasket with starting configuration D_1, D_2, D_3, D_4 since, as stated above,

$$\{C_1, C_2, C_3, C_4\} = \{g^{-1}(D_1), g^{-1}(D_2), g^{-1}(D_3), g^{-1}(D_4)\}.$$

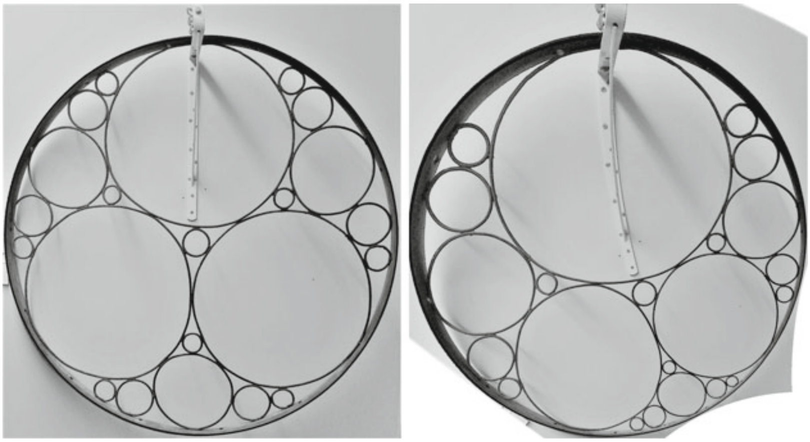


Fig. 3.16 A sculpture that hung in the math department at CUNY's Graduate Center when I was a post-doc there, along with a sculpture that could have hung there, but didn't.

But that means that it is the smallest collection of oriented circles which contains C_1, C_2, C_3, C_4 and which is closed under Descartes swaps, which means that it is just A. □

Thus, at long last, we know that illustrations like in Figure 3.16 really are representative of Apollonian gaskets.

3.8 Descartes' Theorem

In 1643, René Descartes wrote a letter to Princess Elisabeth of the Palatine, with whom he held regular correspondence primarily on questions of philosophy. Included in that letter was the following result, which we state in more modern language.

Theorem 3.16 (Descartes' Theorem) *Let C_1, C_2, C_3, C_4 be an oriented Descartes configuration. Let b_1, b_2, b_3, b_4 be their bends. Then $(b_1+b_2+b_3+b_4)^2 = 2(b_1^2 + b_2^2 + b_3^2 + b_4^2)$.*

Remark 3.8 We don't know exactly what Descartes' proof was, but I can say with confidence that it was not the proof that shall be given here. This is because our proof makes heavy use of matrix multiplication, which was only first described in 1812 by Jacques Philippe Marie Binet—this proof is most likely newer still. There are many, many other known proofs [9].

Proof We know that C_1, C_2, C_3, C_4 are a Descartes configuration if and only if $\langle C_i, C_j \rangle = 1$ if $i \neq j$, and -1 otherwise. If we take the inversive coordinates of C_i to be (b_i, c_i, x_i, y_i), then this can be stated as

$$\begin{pmatrix} b_1 & b_2 & b_3 & b_4 \\ c_1 & c_2 & c_3 & c_4 \\ x_1 & x_2 & x_3 & x_4 \\ y_1 & y_2 & y_3 & y_4 \end{pmatrix}^T \begin{pmatrix} 0 & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix} \begin{pmatrix} b_1 & b_2 & b_3 & b_4 \\ c_1 & c_2 & c_3 & c_4 \\ x_1 & x_2 & x_3 & x_4 \\ y_1 & y_2 & y_3 & y_4 \end{pmatrix} = \begin{pmatrix} -1 & 1 & 1 & 1 \\ 1 & -1 & 1 & 1 \\ 1 & 1 & -1 & 1 \\ 1 & 1 & 1 & -1 \end{pmatrix}.$$

This can be expressed a little more compactly as $D^T MD = R$ if we call

$$D = \begin{pmatrix} b_1 & b_2 & b_3 & b_4 \\ c_1 & c_2 & c_3 & c_4 \\ x_1 & x_2 & x_3 & x_4 \\ y_1 & y_2 & y_3 & y_4 \end{pmatrix} \quad M = \begin{pmatrix} 0 & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix} \quad R = \begin{pmatrix} -1 & 1 & 1 & 1 \\ 1 & -1 & 1 & 1 \\ 1 & 1 & -1 & 1 \\ 1 & 1 & 1 & -1 \end{pmatrix}.$$

It follows that $M = (D^T)^{-1}RD^{-1}$. But if we take the inverse of both sides, we will get the relation $DR^{-1}D^T = M^{-1}$. One checks that $R^{-1} = R/4$ and

$$M^{-1} = \begin{pmatrix} 0 & 2 & 0 & 0 \\ 2 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix},$$

so

$$\begin{pmatrix} b_1 & b_2 & b_3 & b_4 \\ c_1 & c_2 & c_3 & c_4 \\ x_1 & x_2 & x_3 & x_4 \\ y_1 & y_2 & y_3 & y_4 \end{pmatrix} \begin{pmatrix} -1 & 1 & 1 & 1 \\ 1 & -1 & 1 & 1 \\ 1 & 1 & -1 & 1 \\ 1 & 1 & 1 & -1 \end{pmatrix} \begin{pmatrix} b_1 & b_2 & b_3 & b_4 \\ c_1 & c_2 & c_3 & c_4 \\ x_1 & x_2 & x_3 & x_4 \\ y_1 & y_2 & y_3 & y_4 \end{pmatrix}^T = \begin{pmatrix} 0 & 8 & 0 & 0 \\ 8 & 0 & 0 & 0 \\ 0 & 0 & -4 & 0 \\ 0 & 0 & 0 & -4 \end{pmatrix}$$

and in particular

$$\begin{pmatrix} b_1 & b_2 & b_3 & b_4 \end{pmatrix} \begin{pmatrix} -1 & 1 & 1 & 1 \\ 1 & -1 & 1 & 1 \\ 1 & 1 & -1 & 1 \\ 1 & 1 & 1 & -1 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{pmatrix} = 0.$$

Multiplying out the expression above, one gets

$$-b_1^2 - b_2^2 - b_3^2 - b_4^2 + 2b_1b_2 + 2b_1b_3 + 2b_1b_4 + 2b_2b_3 + 2b_2b_4 + 2b_3b_4 = 0,$$

or

$$\begin{aligned} &2(b_1^2 + b_2^2 + b_3^2 + b_4^2) \\ &\quad = b_1^2 + b_2^2 + b_3^2 + b_4^2 + 2b_1b_2 + 2b_1b_3 + 2b_1b_4 + 2b_2b_3 + 2b_2b_4 + 2b_3b_4 \\ &\quad = (b_1 + b_2 + b_3 + b_4)^2, \end{aligned}$$

as was claimed. □

Corollary 3.6 *Let C_1, C_2, C_3, C_4 be an oriented Descartes configuration. Let b_1, b_2, b_3, b_4 be their bends. If C_4' is the Descartes swap of C_4, then its bend is $b_4' = 2b_1 + 2b_2 + 2b_3 - b_4$.*

Proof Since C_1, C_2, C_3, C_4' is also an oriented Descartes configuration, we must have that b_4, b_4' are both solutions to the quadratic polynomial equation

$$(b_1 + b_2 + b_3 + X)^2 = 2(b_1^2 + b_2^2 + b_3^2 + X^2),$$

which is more conveniently rearranged as

$$X^2 - 2(b_1 + b_2 + b_3)X + (b_1^2 + b_2^2 + b_3^2 - 2b_1b_2 - 2b_1b_3 - 2b_2b_3) = 0.$$

The sum of the roots must be $b_4 + b_4' = 2(b_1 + b_2 + b_3)$, whence the result. □

Descartes' theorem and its corollary were rediscovered many times, including by English radiochemist Frederick Soddy in 1936, who then wrote the poem "The Kiss Precise" about it which was published in *Nature* [15]. It would be remiss for me not to include at least an excerpt from it.

Four circles to the kissing come.
The smaller are the benter.
The bend is just the inverse of
The distance from the center.
Though their intrigue left Euclid dumb
There's now no need for rule of thumb.
Since zero bend's a dead straight line
And concave bends have minus sign,
The sum of the squares of all four bends
Is half the square of their sum.

One of Soddy's other contributions to the study of Descartes configurations is the following observation [16].

Corollary 3.7 *Let A be an oriented Apollonian gasket with initial configuration C_1, C_2, C_3, C_4. If the bends of C_1, C_2, C_3, C_4 are integers, then all of the bends in A are integers.*

Proof If the bends are integers b_1, b_2, b_3, b_4, then $b_4' = 2b_1 + 2b_2 + 2b_3 - b_4$ is also an integer. Since every circle in the oriented Apollonian gasket is produced via Descartes swaps, all of them will have integer bends. □

It is common to call an Apollonian gasket *integral* if the bends of all the circles in it are integers. Some examples are shown in Figure 3.17. Ever since Soddy first made this observation, number theorists have been interested in learning more about integral Apollonian gaskets and, in particular, what sort of integers show as bends of such geometric objects. While there has been extensive progress on this question, at

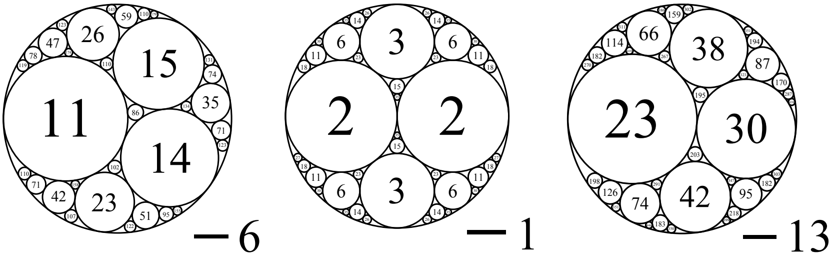


Fig. 3.17 A collection of integral Apollonian gaskets, drawn with their bends.

the time of writing, it is still open, with the best-known result due to Jean Bourgain and Alex Kontorovich [1]. Most of the known partial results make use of heavy algebraic and analytic machinery, so we will not include them here.

Problems

3.1 COMPUTATIONAL EXERCISES

1. Draw a closed Steiner chain of length at least 8.
2. Find a closed Steiner chain of length 6 such that both of the following two conditions hold.

   a. One of the circles in the chain has inversive coordinates $(-1, 1, 0)$.
   b. One of the circles defining the chain has inversive coordinates $(3, 1, 2)$.

   To specify the chain uniquely, it is enough to give the inversive coordinates of the other circle defining the chain; if you want to be a real go-getter, you can find the inversive coordinates of the circles in the chain as well, but be advised that this is significantly harder.
3. a) Suppose that b_1, b_2, b_3, b_4 are the bends of the initial configuration of an Apollonian gasket. Let b_4' be the bend of the Apollonian swap of b_4. Check that

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 2 & 2 & 2 & -1 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ b_4' \end{pmatrix}.$$

   b) Find the matrices such that multiplying them by (b_1, b_2, b_3, b_4) gives the other three Apollonian swaps.

4. Find oriented circles C_1, C_2, C_3, C_4 which form a Descartes configuration, and such that their bends are 0, 1, 1, 4, respectively.
5. Find oriented circles C_1, C_2, C_3, C_4 which form a Descartes configuration, and such that their bends are $-1, 2, 2, 3$, respectively.
6. Calculate the bends of all of the circles that one can get from the standard oriented Descartes configuration in no more than ten Descartes swaps. (You will want a computer for this.) Investigate the set that you find—can you make any conjectures about what it does and does not contain?

3.2 PROOFS

1. Let C_1, C_2 be two perpendicular generalized circles. For any point p on C_1, prove that there exists a unique generalized circle C_3 that is perpendicular to both C_1 and C_2 and which passes through p. *(Hint: use a linear fractional transformation to move one of the points where C_1 and C_2 intersect to ∞. How does this help?)*

2. Let C be a generalized circle that is perpendicular to $y = 0$ and $x = 0$. Prove that the center of C is 0.
3. Define a *Pappus chain* as follows: we start with two generalized circles C_1 and C_2 that intersect in a single point. Choose a point p on C_1 and construct a generalized circle D_0 that is tangent to C_2 and tangent to C_1 at p. Add on the generalized circles that are tangent to D_0, C_1, and C_2—we claim that there are two of them, D_1 and D_{-1}. Add on the generalized circles that are tangent to D_1, C_1, and C_2, and tangent to D_{-1}, C_1, C_2—we claim that there are again two of them, D_2 and D_{-2}. Continue inductively: at each step, add circles D_k and D_{-k} that are tangent to D_{k-1}, C_1, and C_2 and tangent to D_{-k+1}, C_1, and C_2. The Pappus chain is the union of C_1, C_2, and all of the D_k's.

   a) Draw a picture of a Pappus chain.
   b) Prove that given p, C_1, and C_2, there exists a unique generalized circle D_0 that is tangent to C_2 and tangent to C_1 at p.
   c) Prove that there exist two generalized circles D_1, D_{-1} that are tangent to C_1, C_2, and D_0.
   d) Use induction to prove that for every integer k, there exist two generalized circles D_{k+1} and D_{k-1} that are tangent to D_k, C_1, and C_2.
   e) Prove that if two circles in the Pappus chain intersect, then they are tangent.
   f) What effect does changing the initial point p have?

4. Our goal is to prove that the Apollonian gasket with starting configuration C_1, C_2, C_3, C_4 is equal to the set $\mathcal{A}$ defined as

$$\mathcal{A} = \bigcup_{n=1}^{\infty} \mathcal{A}_n,$$

   where $\mathcal{A}_1$ is the starting configuration, $\mathcal{A}_2$ consists of all Descartes swaps of quadruples in $\mathcal{A}_1$, $\mathcal{A}_3$ consists of all Descartes swaps of quadruples in $\mathcal{A}_2$, and so on.

   a) Prove that the Apollonian gasket contains each $\mathcal{A}_n$. *(Hint: you may want to use induction.)*
   b) Prove that the Apollonian gasket contains $\mathcal{A}$.
   c) Prove that $\mathcal{A}$ contains all Descartes swaps of elements in $\mathcal{A}$.
   d) Conclude that $\mathcal{A}$ is the Apollonian gasket.

5. Prove that for any oriented circle C, $\operatorname{inv}(-C) = -\operatorname{inv}(C)$.
6. We confirm some of the basic properties of the conjugate transpose.

a) Prove that $\overline{(AB)} = \overline{A}\,\overline{B}$ for all 2×2 matrices A, B with complex coefficients. Here,

$$\overline{\begin{pmatrix} a & b \\ c & d \end{pmatrix}} = \begin{pmatrix} \overline{a} & \overline{b} \\ \overline{c} & \overline{d} \end{pmatrix}.$$

b) Prove that if $A \in SL(2, \mathbb{C})$ then $\overline{A} \in SL(2, \mathbb{C})$.

c) Prove that $(AB)^T = B^T A^T$. Here,

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^T = \begin{pmatrix} a & c \\ b & d \end{pmatrix}.$$

d) Prove that if $A \in SL(2, \mathbb{C})$, then $A^T \in SL(2, \mathbb{C})$.

e) Prove that $\overline{(AB)}^T = \overline{B}^T \overline{A}^T$ for all 2×2 matrices A, B with complex coefficients.

f) Prove that if $A \in SL(2, \mathbb{C})$, then $\overline{A}^T \in SL(2, \mathbb{C})$.

7. Prove that every matrix $SL(2, \mathbb{C})$ can be written as a product of matrices of the forms

$$\begin{pmatrix} u & 0 \\ 0 & u^{-1} \end{pmatrix}, \begin{pmatrix} 1 & \tau \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

8. a) Let M, N be $n \times n$ matrices with i, j-th entries $a_{i,j}$ and $b_{i,j}$, respectively. By the definition of matrix multiplication, the i, j-th entry of MN is

$$c_{i,j} = \sum_{k=1}^{n} a_{i,k} b_{k,j}.$$

Use this observation to prove that $\mathrm{tr}(MN) = \mathrm{tr}(NM)$.

b) Use your answer to the previous part to show that for any $n \times n$ matrix M and any $G \in GL(n, \mathbb{C})$, $\mathrm{tr}(GMG^{-1}) = \mathrm{tr}(M)$.

9. Prove Corollary 3.4.

10. The following describes an analog of the Apollonian gasket which comes from a paper by Guttler and Mallows [3].

a) Prove that for any circles C_1, C_2, C_3 that are externally tangent to one another, there exist exactly two triples of circles D_1, D_2, D_3 such that D_1, D_2, D_3 are externally tangent to one another, D_1 is externally tangent to C_2 and C_3, D_2 is externally tangent to C_1 and C_3, and D_3 is externally tangent to C_1 and C_2. *(Hint: use a linear fractional transformation to move C_1, C_2, C_3 into a configuration that is easier to think about. There are a number of choices for how to do this.)*

b) We shall call a sextuple C_1, C_2, C_3, D_1, D_2, D_3 satisfying the properties above an *octahedral configuration*. Prove that six oriented circles C_1, C_2, C_3, D_1, D_2, D_3 form an octahedral configuration if and only if

$$\langle C_i, C_j\rangle_I = \begin{cases} 1 & \text{if } i \neq j \\ -1 & \text{if } i = j \end{cases} \quad \langle D_i, D_j\rangle_I = \begin{cases} 1 & \text{if } i \neq j \\ -1 & \text{if } i = j \end{cases}$$

$$\langle C_i, D_j\rangle_I = \begin{cases} 1 & \text{if } i \neq j \\ -3 & \text{if } i = j. \end{cases}$$

(Hint: use the fact that the inversive distance does not change under linear fractional transformations, and force the sextuple into a position where it is comparatively easy to compute what the inversive distance is for an octahedral configuration.)

c) Draw an octahedral configuration. Choose any point in the interior of each oriented circle. For any two points, connect them by a line if their corresponding circles are externally tangent. Looking at the figure you have drawn, can you see why this configuration is called octahedral? *(Hint: what would the vertices and edges of an octahedron look like if you squashed them flat onto a plane?)*

d) Given an octahedral configuration C_1, C_2, C_3, D_1, D_2, D_3, denote by D_1', D_2', D_3' the other three oriented circles such that C_1, C_2, C_3, D_1', D_2', D_3' is an octahedral configuration. Call the map $(C_1, C_2, C_3, D_1, D_2, D_3) \mapsto (C_1, C_2, C_3, D_1', D_2', D_3')$ an *octahedral swap*. Show that any octahedral swap can be understood as an inversion through some circle. *(Hint: prove this for a standard octahedral configuration first.)*

e) Prove that if C_1, C_2, C_3, D_1, D_2, D_3 is an octahedral configuration, then $\mathrm{inv}(C_1) + \mathrm{inv}(D_1) = \mathrm{inv}(C_2) + \mathrm{inv}(D_2) = \mathrm{inv}(C_3) + \mathrm{inv}(D_3)$. *(Hint: first, show that if this holds true for one octahedral configuration, then it is true for all octahedral configurations. Then prove it for a convenient choice of octahedral configuration.)*

f) Given an octahedral configuration C_1, C_2, C_3, D_1, D_2, D_3, define its *condensed coordinates* to be the quadruple $v_1 = \mathrm{inv}(C_1)$, $v_2 = \mathrm{inv}(C_2)$, $v_3 = \mathrm{inv}(C_3)$, $v_4 = \mathrm{inv}(C_1) + \mathrm{inv}(D_1)$. Prove that the condensed coordinates of an octahedral configuration specify it uniquely.

g) Consider an octahedral configuration with condensed coordinates v_1, v_2, v_3, v_4. If V is a matrix with columns v_1, v_2, v_3, v_4 and

$$M = \begin{pmatrix} 0 & -1/2 & 0 & 0 \\ -1/2 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

prove that

$$V^T M V = R := \begin{pmatrix} 1 & -1 & -1 & -2 \\ -1 & 1 & -1 & -2 \\ -1 & -1 & 1 & -2 \\ -2 & -2 & -2 & -4 \end{pmatrix}.$$

(Hint: you can prove this by passing to a standard configuration, but it is probably easier to use the result of part b).)

h) Let C_1, C_2, C_3, D_1, D_2, D_3 be an octahedral configuration. Let the bends of C_1, C_2, C_3, D_1, D_2, D_3 be b_1, b_2, b_3, d_1, d_2, d_3. Prove that $(d_1, d_2, d_3) = (2X - b_1, 2X - b_2, 2X - b_3)$, where X is a root of

$$X^2 - 2X(b_1 + b_2 + b_3) + b_1^2 + b_2^2 + b_3^2 = 0.$$

(Hint: use the previous exercise and emulate the proof of Descartes' theorem.)

i) Let C_1, C_2, C_3, D_1, D_2, D_3 be an octahedral configuration with bends b_1, b_2, b_3, d_1, d_2, d_3. Let C_1, C_2, C_3, D_1',D_2',D_3' be the octahedral swap where the new bends are d_1', d_2', d_3'. Prove that $d_i' = 4(b_1 + b_2 + b_3) - 2b_i - d_i$ for $i = 1, 2, 3$. *(Hint: use the previous exercise and emulate the proof of Corollary 3.6.)*

j) Let C_1, C_2, C_3, D_1, D_2, D_3 be an octahedral configuration. An *octahedral packing* with initial configuration C_1, C_2, C_3, D_1, D_2, D_3 is the smallest set of oriented circles that satisfies the following two properties.

   a. The set contains C_1, C_2, C_3, D_1, D_2, D_3.
   b. The set contains all octahedral swaps of elements in the set.

   We say that an octahedral packing is *integral* if the bends of all the circles contained inside of it are all integers. Prove that an octahedral packing is integral if and only if the bends of the initial configuration are all integers. *(Hint: use the previous exercise and emulate the proof of Corollary 3.7.)*

k) Draw a picture of an octahedral packing.

3.3 PROOFS (Group Theory)

1. Prove that the Apollonian group is a group.

2. Prove that $SL(2, \mathbb{Z}[i])$ is a group.
3. Define $PSL(2, \mathbb{C})$ to be the set of equivalence classes of $SL(2, \mathbb{C})$ where two matrices M_1, M_2 are considered to be equivalent if there exists $\lambda \in \mathbb{C}$ such that $M_1 = \lambda M_2$.

   a) Let $M_1, M_1', M_2, M_2' \in SL(2, \mathbb{C})$ such that M_1, M_1' are equivalent and M_2, M_2' are equivalent. Prove that $M_1 M_2, M_1' M_2'$ are equivalent.
   b) Using the above, prove that $PSL(2, \mathbb{C})$ is a group. *(Hint: The main problem is showing that it has a well-defined multiplication on it. The previous part suggested how to do this.)*
   c) Prove that

$$\begin{aligned} \varphi : PSL(2, \mathbb{C}) &\to PGL(2, \mathbb{C}) \\ M &\mapsto M \end{aligned}$$

   is a well-defined group isomorphism.

4. Given a set S and a group G, an *action* of G on S is a function $A : G \times S \to S$ satisfying the following properties.

   a. If $\imath$ is the identity of G, then $A(\imath, s) = s$ for all $s \in S$.
   b. For all $g, h \in G$ and all $s \in S$, $A(f, A(g, s)) = A(fg, s)$.

   Wherever it is unlikely to cause confusion, it is customary to write $g.s$ instead of $A(g, s)$.

   a) Let X be any set. Define $\text{Sym}(X)$ to be the set of bijective functions $f : X \to X$. Prove that $\text{Sym}(X)$ is a group if we take the operation to be composition of functions.
   b) Let A be an action of G on X. Prove that

$$\begin{aligned} G &\to \text{Sym}(X) \\ g &\mapsto (x \mapsto A(g, x)) \end{aligned}$$

   is a group homomorphism.
   c) Prove that if $\phi : G \to \text{Sym}(X)$ is a group homomorphism, then

$$\begin{aligned} A : G \times X &\mapsto X \\ (g, x) &\mapsto \phi(g)(x) \end{aligned}$$

   is a group action.

5. Prove that

$$\begin{aligned} \text{Isom}(\mathbb{C}) \times \mathbb{C} &\to \mathbb{C} \\ (\phi, z) &\mapsto \phi(z) \end{aligned}$$

   is a group action.

6. Prove that

$$GL(2,\mathbb{C}) \times \mathbb{C}P^1 \to \mathbb{C}P^1$$
$$\left(\begin{pmatrix} a & b \\ c & d \end{pmatrix}, z\right) \mapsto \frac{az+b}{cz+d}$$

is a group action. *(Hint: you can prove this directly, of course, but it might be a little more convenient to use the result of Exercise* 3.3.4 *and Theorem* 2.4.*)*

7. Prove that Definition 3.6 defines a group action of $SL(2,\mathbb{C})$ on the set of oriented circles in $\mathbb{C}P^1$.

8. A group action of G on X is called *transitive* if for every $x, y \in X$, there exists $g \in G$ such that $g.x = y$.

   a) Prove that $SL(2,\mathbb{C})$ acts transitively on the set of oriented circles in $\mathbb{C}P^1$.
   b) Prove that the action of $GL(2,\mathbb{C})$ on $SL(2,\mathbb{C})$ by conjugation is not transitive. *(Hint: find two matrices $M_1, M_2 \in SL(2,\mathbb{C})$ with different traces. Is it possible that $\gamma M_1 \gamma^{-1} = M_2$ for some $\gamma \in GL(2,\mathbb{C})$?)*

9. A group action of G on X is called *free* if for every $g, h \in G$ and $x \in X$, $g.x = g.y$ if and only if $g = h$.

   a) Prove that S^1, the collection of all $z \in \mathbb{C}$ with $|z| = 1$ is a group under multiplication.
   b) Prove that the action of S^1 on $\mathbb{C}^\times$ by left multiplication is free.
   c) Prove that $SL(2,\mathbb{C})$ does not act freely on the set of oriented circles in $\mathbb{C}P^1$.

10. Prove that $PSL(2,\mathbb{C})$ acts freely, transitively on the set of distinct triples of points in $\mathbb{C}P^1$.

11. Prove that the Apollonian group acts freely, transitively on the set of oriented Descartes configurations in the Apollonian gasket.

4 Construction of Hyperbolic Geometry

In which we construct new spaces on which to act.

We have carefully studied the properties of linear fractional transformations on the Euclidean plane; it is now time to look elsewhere. There are many possible candidates for exposition but probably the single most important is hyperbolic space. The hyperbolic plane was the original example of a non-Euclidean space—that is, a geometry that satisfied all of Euclid's axioms for plane geometry save for what is now known as the Fifth Postulate.

The Fifth Postulate

If a line segment intersects two straight lines forming two interior angles on the same side that sum to less than two right angles, then the two lines, if extended indefinitely, meet on that side on which the angles sum to less than two right angles.

This formulation is the way that Euclid originally phrased it, anyway, but I personally prefer a slightly different version known as Playfair's Axiom.

Playfair's Axiom

Given a line l and a point p not on l, there is exactly one line l' that passes through p and does not intersect l.

The idea behind hyperbolic geometry is that rather than there being exactly one non-intersecting line, there are instead many, as shown in Figure 4.1. Hyperbolic geometry was developed in the 19th century when the mathematical soil was finally ready for such a thing to sprout. This is evidenced by the fact that it was discovered

A. Sheydvasser, *Linear Fractional Transformations*, Undergraduate Texts in Mathematics, https://doi.org/10.1007/978-3-031-25002-6_4

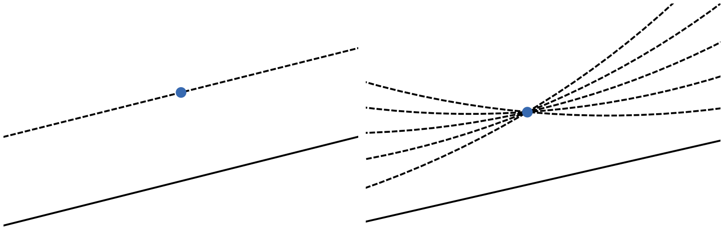


Fig. 4.1 On the left, an illustration of the Playfair axiom in the Euclidean plane. On the right, an example of how this axiom is not satisfied by the hyperbolic plane.

independently by at least four different mathematicians at around the same time: Nikolai Ivanovich Lobachevsky, János Bolyai, Carl Friedrich Gauss, and Franz Taurinus. We shall develop a model of hyperbolic geometry that was originally put forward by Italian mathematician Eugenio Beltrami in 1868, although, in a classic instance of Stigler's law of eponymy, it usually carries Henri Poincaré's name instead. Before we get into this properly, we will first examine what we even mean by a geometry for our purposes.

4.1 Metric Geometry

Historically, geometry was first studied axiomatically (à la Euclid), then analytically using coordinates (à la Descartes), and then using differential forms (à la Gauss and Riemann). Instead of developing any of these potent machines, we will instead leap forward in time to 1906 and consider Maurice Fréchet's notion of a metric space.

Definition 4.1 A *metric space* (X, d) is a set X together with a function $d : X \to [0, \infty)$ called the *metric* which satisfies the following three properties.

1. For all $x, y \in X$, $d(x, y) = 0$ if and only if $x = y$.
2. For all $x, y \in X$, $d(x, y) = d(y, x)$.
3. For all $x, y, z \in X$, $d(x, y) + d(y, z) \geq d(x, z)$.

Fréchet introduced this notion in his thesis, securing its place as a foundational paper in mathematics. Today, metric spaces are used to phrase the basic theorems of real analysis, extend the notions of limits to spaces of functions (as Fréchet himself did), analyze error-correcting codes, and much more. But what *is* a metric space?

I claim that you already familiar with at least one example of a metric space. Specifically, consider the set $\mathbb{R}^n$—that is, n-dimensional Euclidean space. You can take $n = 2$ if you want to feel more comfortable. Then, consider the Euclidean distance function

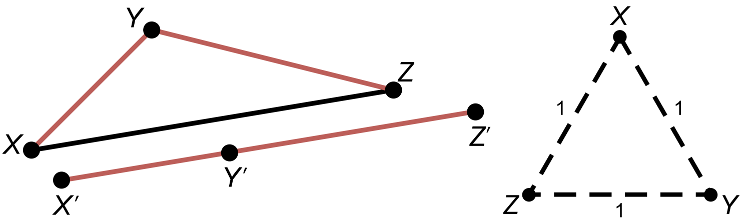


Fig. 4.2 On the left, an illustration of the triangle inequality in the Euclidean plane; on the right, an illustration of the discrete metric space with three points.

$$
\begin{aligned}
d_{\text{Euclid}} &: \mathbb{R}^n \times \mathbb{R}^n \to [0, \infty) \\
((x_1, \ldots, x_n), (y_1, \ldots, y_n)) &\mapsto \sqrt{(x_1 - y_1)^2 + \ldots + (x_n - y_n)^2}.
\end{aligned}
$$

This is a metric space. Indeed, it is true that $d_{\text{Euclid}}(x, y) = 0$ if and only if $x = y$; it is true that $d_{\text{Euclid}}(x, y) = d_{\text{Euclid}}(y, x)$. Neither of these are difficult to prove. What about the last assertion? Well, all it is saying is that if we have three points x, y, z, then the distance from x to y plus the distance from y to z can't be less than the distance from x to z—this is just the *triangle inequality*

$$d_{\text{Euclid}}(x, y) + d_{\text{Euclid}}(y, z) \geq d_{\text{Euclid}}(x, z).$$

Proving this is a little trickier (see Exercise 4.5.4), but it is intuitively clear, as shown in Figure 4.2. In any case, we now understand what a metric space is: it is just a set together with a distance function, where the notion of "distance" is just a straightforward generalization of the usual Euclidean one. For us, this is what we will mean by a "geometry": it is a choice of metric.

Let's give another example of a metric space. Let X be any set whatsoever, and define a function

$$
\begin{aligned}
d_{\text{discrete}} &: X^2 \to [0, \infty) \\
(x, y) &\mapsto \begin{cases} 1 & \text{if } x \neq y \\ 0 & \text{otherwise.} \end{cases}
\end{aligned}
$$

This is known as the discrete metric, and it turns X into a metric space. I leave the proof for the reader—it isn't hard. (See Exercise 4.5.2.) This example highlights that while metric geometry is good enough to talk about distances, it isn't usually good enough to talk about angles. After all, $\{1, 2, 3\}$ together with d_{discrete} is a perfectly good metric space (illustrated in Figure 4.2), but there is no reasonable way to talk about angles between paths in this space, particularly since there is no such thing as a (continuous) path between any two points in this geometry. While this is limiting, we'll see that for the examples that we consider, one can define angles in a reasonable way, so we won't worry about this too much. Other geometric notions that we defined in Chapter 1 generalize much more easily.

Definition 4.2 For any two metric spaces (X, d_X), (Y, d_Y), an *isometry* between them is a function $\Psi : X \to Y$ such that for all $x_1, x_2 \in X$, $d_Y(\Psi(x_1), \Psi(x_2)) = d_X(x_1, x_2)$. Furthermore, Ψ is called an *isometric isomorphism* if it is also bijective. We will write $\text{Isom}(X, d)$ to denote the set of isometries of (X, d) with itself. If the metric d is clear from context, we will instead abbreviate this as $\text{Isom}(X)$.

It is often the case that all of the isometries $X \to X$ are bijective and so the two notions of isometry and isometric isomorphism coincide: this was the case, for example, for isometries of the Euclidean plane. For general metric spaces, though, they can be different. Let's see this by thinking about discrete metric spaces. If X is a discrete metric space, then $\Psi : X \to X$ is an isometry if and only if for all $x, y \in X$,

$$d_{\text{discrete}}(\Psi(x), \Psi(y)) = \begin{cases} 1 & \text{if } \Psi(x) \neq \Psi(y) \\ 0 & \text{otherwise} \end{cases}$$

$$d_{\text{discrete}}(x, y) = \begin{cases} 1 & \text{if } x \neq y \\ 0 & \text{otherwise,} \end{cases}$$

which is to say that $x = y$ if and only if $\Psi(x) = \Psi(y)$. In other words, Ψ is an isometry if and only if it is injective. However, if X is an infinite set, then one can find maps that are injective but not surjective: for instance, if $X = \mathbb{Z}$, the map $n \mapsto 2n$ is just such a thing. Even so, there are restrictions on what isometries can be: any isometry is always injective (see Exercise 4.5.6), any composition of isometries is necessarily an isometry (see Exercise 4.5.7), and if an isometry has an inverse, then that inverse is also an isometry (see Exercise 4.5.8).

▶ **Example** *For $x, y \in \mathbb{R}$, define $|(x, y)|_1 = |x| + |y|$. The taxicab metric on $\mathbb{R}^2$ is defined as $d_1(p_1, p_2) = |p_1 - p_2|_1$. Prove that it is a metric.*
Write $p_1 = (x_1, y_1)$ and $p_2 = (x_2, y_2)$. Then $d_1(p_1, p_2) = |x_1 - x_2| + |y_1 - y_2|\}$. This makes it evident that for all $p_1, p_2 \in \mathbb{R}^2$, $d_1(p_1, p_2) = 0$ if and only if $p_1 = p_2$ and $d_1(p_1, p_2) = d_1(p_2, p_1)$. The only tricky part is the triangle inequality. Adding a third point $p_3 = (x_3, y_3)$, we note that

$$\begin{aligned} d_1(p_1, p_2) + d_1(p_2, p_3) &= |x_1 - x_2| + |y_1 - y_2| + |x_2 - x_3| + |y_2 - y_3| \\ &\geq |x_1 - x_3| + |y_1 - y_3| = d_1(p_1, p_3), \end{aligned}$$

where in the second to last step, we used the fact that $|a - b| + |b - c| \geq |a - c|$, which is just the triangle inequality for $(\mathbb{R}, d_{\text{Euclid}})$.

4.2 The Real Special Linear Group

Up until this point, our general tendency has been to first describe a geometry, and then to work out the transformations that preserve its key properties. Thus, for example, we defined the plane with the Euclidean metric and then determined what the isometries were; just so, we defined the discrete metric in the previous section and then worked out the isometries. Now, we are going to flip the script: we will start with what we want the isometries to be, and we shall try to find a nice metric space that matches. One possible justification for this approach is the following.

Philosophical Principle

> If you want to understand a group G, find a nice space X such that G can be interpreted as the group of transformations $X \to X$ with some convenient properties, such as them being isometries.

Later in the chapter, we will do precisely this for the full group Möb(2). For now, we shall try this for a somewhat smaller, easier to work with, subgroup.

Definition 4.3 The *special linear group on* $\mathbb{R}^2$, denoted by $SL(2, \mathbb{R})$, consists of all 2×2 matrices with real coefficients with determinant 1. The group $PSL(2, \mathbb{R})$ is the image of this group in $\text{Möb}^0(2)$—that is, it consists of all transformations

$$z \mapsto \frac{az + b}{cz + d}$$

where $a, b, c, d \in \mathbb{R}$ and $ad - bc = 1$.

That $SL(2, \mathbb{R})$ is a subgroup of $SL(2, \mathbb{C})$ is easy to check. (See Exercise 4.2.2.) In the next section, we will construct a metric space $X \subset \mathbb{C}P^1$ such that linear fractional transformations in $PSL(2, \mathbb{R})$ are isometries of X. To set the scene, we need a subset X such that every transformation in $PSL(2, \mathbb{R})$ moves it back to itself.

Definition 4.4 The *upper half-plane* $\mathbb{H}^2$ is the set of points $z \in \mathbb{C}$ such that $\Im(z) > 0$. The *boundary* $\partial\mathbb{H}^2$ of the upper half-plane is $\partial\mathbb{H}^2$.

Lemma 4.1 *For any* $\varphi \in PSL(2, \mathbb{R})$, $\varphi(\mathbb{H}^2) = \mathbb{H}^2$.

Proof What is $\mathbb{H}^2$? It is just the interior of the real line, oriented such that i is in the interior. If $\varphi \in \text{Möb}^0(2)$, then $\varphi(\mathbb{H}^2) = \mathbb{H}^2$ if and only if φ sends the real line with this orientation back to itself. If we refer back to the exercise at the end of Section 3.4, we see that, indeed, if $\varphi \in PSL(2, \mathbb{R})$, then this is exactly what happens.

A consequence of this is that it is entirely sensible to ask how elements in $PSL(2, \mathbb{R})$ move elements in $\mathbb{H}^2$—an illustration is shown in Figure 4.3. This is, however, a little complicated. It's a good idea to come up with some simple examples

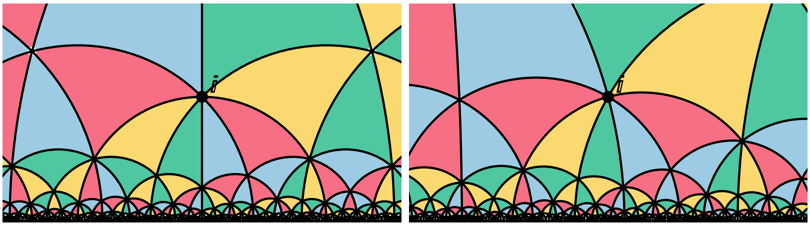


Fig. 4.3 An example of how $SL(2,\mathbb{R})$ moves points in $\mathbb{H}^2$. The illustration on the right is the image of the illustration left under the map $z \mapsto \begin{pmatrix} \cos(\pi/12) & -\sin(\pi/12) \\ \sin(\pi/12) & \cos(\pi/12) \end{pmatrix}.z$

of elements in $PSL(2,\mathbb{R})$. Notice that if $\tau \in \mathbb{R}$, $\lambda > 0$, then the maps $z \mapsto z+\tau$, $z \mapsto \lambda z$, $z \mapsto -z^{-1}$ are all in $PSL(2,\mathbb{R})$ since they correspond to the matrices

$$\begin{pmatrix} 1 & \tau \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} \sqrt{\lambda} & 0 \\ 0 & 1/\sqrt{\lambda} \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix},$$

which are all in $SL(2,\mathbb{R})$. Another useful transformation that preserves the upper half-plane, but is not in $PSL(2,\mathbb{R})$, is $z \mapsto -\overline{z}$. Together, these transformations help us prove nice properties about $PSL(2,\mathbb{R})$.

Theorem 4.1 *The map*

$$SL(2,\mathbb{R}) \to \left\{ \psi \in M\ddot{o}b^0(2) \middle| \psi(\mathbb{H}^2) = \mathbb{H}^2 \right\}$$
$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \mapsto \left(z \mapsto \frac{az+b}{cz+d} \right)$$

is a surjective group homomorphism.

Proof It is easy to check that

$$\left\{ \psi \in \text{Möb}^0(2) \middle| \psi(\mathbb{H}^2) = \mathbb{H}^2 \right\}$$

is a group and so this map is just a restriction of the usual group homomorphism $SL(2,\mathbb{C}) \to \text{Möb}^0(2)$. By Lemma 4.1, we know that the image of $SL(2,\mathbb{R})$ preserves the upper half-plane. Therefore, this is a well-defined group homomorphism; it remains to prove that it is surjective. Choose an arbitrary element $\psi \in \text{Möb}^0(2)$ such that $\psi(\mathbb{H}^2) = \mathbb{H}^2$—we know that there exist $a, b, c, d \in \mathbb{C}$ such that $ad - bc = 1$ and

$$\psi(z) = \frac{az+b}{cz+d}.$$

We shall show that in fact $a, b, c, d \in \mathbb{R}$. We know that $\psi(\partial\mathbb{H}^2) = \partial\mathbb{H}^2$, hence $\psi(\infty) = a/c \in \partial\mathbb{H}^2$ and $\psi^{-1}(0) = -b/a \in \partial\mathbb{H}^2$. By composing with $z \mapsto -1/z$ if necessary, we may assume that $a \neq 0$, hence $c/a, b/a \in \mathbb{R}$. But

$$\begin{pmatrix} 1 & 0 \\ -c/a & 1 \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} 1 & -b/a \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} a & 0 \\ 0 & 1/a \end{pmatrix},$$

and this new matrix is in $SL(2, \mathbb{R})$ if and only if the original one was. Thus, we may assume without loss of generality that $b = c = 0$, so $\psi(z) = a^2 z$. This preserves $\partial \mathbb{H}^2$ if and only if $a^2 \in \mathbb{R}^\times$. Moreover, if $a^2 < 0$, then $\psi(i) \notin \mathbb{H}^2$. This implies that $a^2 > 0$, which in turn means that $a \in \mathbb{R}$, concluding the proof.

Corollary 4.1 *For any $\psi \in M\ddot{o}b(2)$, $\psi(\mathbb{H}^2) = \mathbb{H}^2$ if and only if either $\psi \in PSL(2, \mathbb{R})$ or $\psi \circ \phi \in PSL(2, \mathbb{R})$, where $\phi(z) = -\overline{z}$.*

Proof If $\psi \in \text{Möb}^0(2)$, this follows from Theorem 4.1. If $\psi \notin \text{Möb}^0(2)$, then ψ composed with $z \mapsto -\overline{z}$ is in $\text{Möb}^0(2)$, and so the result follows. □

With this motivation, we define a group $\text{Isom}(\mathbb{H}^2)$.

Definition 4.5 We define

$$\begin{aligned} \text{Isom}(\mathbb{H}^2) &= \left\{ \psi \in \text{Möb}(2) \,\middle|\, \psi(\mathbb{H}^2) = \mathbb{H}^2 \right\} \\ &= PSL(2, \mathbb{R}) \cup \left\{ \psi \circ \phi \,\middle|\, \psi \in PSL(2, \mathbb{R}) \right\}, \end{aligned}$$

where $\phi(z) = -\overline{z}$.

Remark 4.1 It is easy to check that this is a group—see Exercise 4.2.3.

We are jumping the gun here a little—what we are effectively claiming is that this is the isometry of the upper half-plane. However, we haven't defined a metric on $\mathbb{H}^2$ yet, so it is meaningless to talk about isometries! Nevertheless, in the next section, we will define a metric on $\mathbb{H}^2$ and then it really will be the case that $\text{Isom}(\mathbb{H}^2)$ will be the set of isometries $\mathbb{H}^2 \to \mathbb{H}^2$.

Before we are ready to make that construction, though, we will need a few results showing how $\text{Isom}(\mathbb{H}^2)$ and $PSL(2, \mathbb{R})$ act on points in $\mathbb{H}^2$ and its boundary.

Lemma 4.2 *For any three distinct points $z_1, z_2, z_3 \in \partial \mathbb{H}^2$, there exists an element $\psi \in Isom(\mathbb{H}^2)$ such that $\psi(z_1) = 0$, $\psi(z_2) = 1$, and $\psi(z_3) = \infty$.*

Proof We will construct ψ by stages. First, if $z_3 \neq \infty$, define $\psi_1(z) = (z_3 - z)^{-1}$—this is the image of

$$\begin{pmatrix} 0 & 1 \\ -1 & z_3 \end{pmatrix} \in SL(2, \mathbb{R}).$$

If $z_3 = \infty$, define $\psi_1(z) = z$. In either case, $PSL(2, \mathbb{R})$ and $\psi_1(z_3) = \infty$. Since $\psi_1(z_1) \neq \infty$, we may then consider the translation $\psi_2(z) = z - \psi_1(z_1)$. This sends $z_1 \mapsto 0$, $z_3 \mapsto \infty$. Thus $(\psi_2 \circ \psi_1)(z_2)$ is a non-zero real number. If it is less than zero, define $\psi_3(z) = -\overline{z}$; otherwise, $\psi_3(z) = z$. Then $\psi_3 \circ \psi_2 \circ \psi_1 \in \text{Isom}(\mathbb{H}^2)$ and sends $z_1 \mapsto 0$, $z_3 \mapsto \infty$, and z_2 to some positive real number r. Finally, define $\psi_4(z) = z/r$, so $\psi_4 \circ \psi_3 \circ \psi_2 \circ \psi_1$ is the desired element of $\text{Isom}(\mathbb{H}^2)$. □

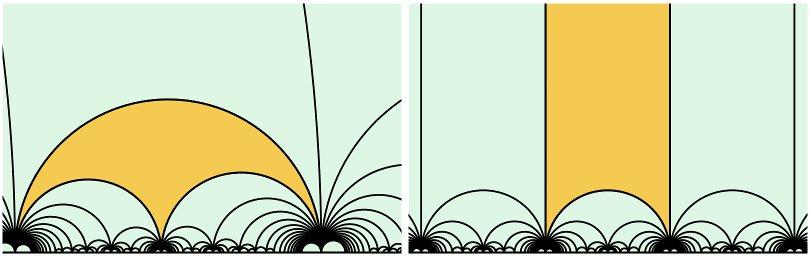


Fig. 4.4 An illustration of how Isom($\mathbb{H}^2$) acts transitively on triples of points on the boundary of $\mathbb{H}^2$; the right-hand side is the image of the left under a transformation that sends the three vertices of the "triangle" to 0, 1, ∞.

Theorem 4.2 *For any pair of distinct triples of points $z_1, z_2, z_3 \in \partial\mathbb{H}^2$, $w_1, w_2, w_3 \in \partial\mathbb{H}^2$, there exists a unique $\psi \in Isom(\mathbb{H}^2)$ such that $\psi(z_1) = w_1$, $\psi(z_2) = w_2$, and $\psi(z_3) = w_3$.*

Proof Showing existence is easy; simply apply Lemma 4.2 to produce $\psi_1, \psi_2 \in$ Isom($\mathbb{H}^2$) such that $\psi_1(z_1) = \psi_2(w_1) = 0$, $\psi_1(z_2) = \psi_2(w_2) = 1$, and $\psi_1(z_3) = \psi_2(w_3) = \infty$. Composing $\psi_2^{-1} \circ \psi_1$ produces the desired element of Isom($\mathbb{H}^2$). On the other hand, if $\phi \in$ Möb(2) satisfies $\phi(\mathbb{H}^2) = \mathbb{H}^2$, $\phi(z_1) = w_1$, $\phi(z_2) = w_2$, and $\phi(z_3) = w_3$, then we claim that $\phi = \psi_2^{-1} \circ \psi_1$. Equivalently, $\psi_3 = \psi_2 \circ \phi \circ \psi_1^{-1}$ is the identity function. We know that $\psi_3(0) = 0$, $\psi_3(1) = 1$, and $\psi_3(\infty) = \infty$—therefore, either $\psi_3(z) = z$ or $\psi_3(z) = \overline{z}$. But conjugation does not preserve the upper half-plane, hence $\psi_3 = id$, establishing uniqueness. □

An example of this transitive action on triples of points on the boundary is shown in Figure 4.4. We also need to know how points off the boundary are moved around.

Theorem 4.3 *For any two distinct points $z_1, z_2 \in \mathbb{H}^2$, there exists exactly two elements $\psi \in Isom(\mathbb{H}^2)$ such that $\psi(z_1) = i$, and $\psi(z_2) = it$ for some real $t > 1$; one element is in $PSL(2, \mathbb{R})$ and the other isn't. Furthermore, the parameter t is the same for both.*

Proof First, we tackle showing that there exist two such elements. We're going to carefully choose two points $a, b \in \partial\mathbb{H}^2$ such that if we choose an element $\psi_1 \in$ Isom($\mathbb{H}^2$) so that $a \mapsto 0$ and $b \mapsto \infty$, then z_1, z_2 will fall on the imaginary axis, with $\Im(\psi_1(z_2)) > \Im(\psi_1(z_1))$. Assuming we can do this, the rest of the construction is easy: first, define $\psi_2(z) = z/\Im(\psi_1(z_1))$, so then $\psi = \psi_2 \circ \psi_1 \in$ Isom($\mathbb{H}^2$) sends $z_1 \mapsto i$ and $z_2 \mapsto it$ for some $t > 1$. The desired second element is given by $\phi \circ \psi$, where $\phi(z) = -\overline{z}$.

That there exists an element in Isom($\mathbb{H}^2$) which can send arbitrary points on the boundary to 0 and ∞ is guaranteed by Theorem 4.2. But how to choose a, b to accomplish what we want? Consider the line l through z_1 and z_2. There are three cases:

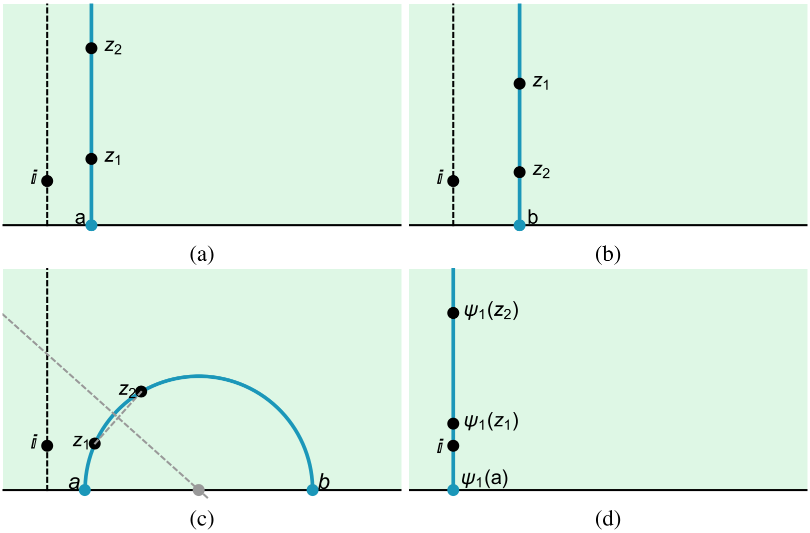


Fig. 4.5 A visual sketch of the existence part of the proof of Theorem 4.3. (a), (b), and (c) show the possible configurations of a pair of points $z_1, z_2 \in \mathbb{H}^2$ and the corresponding choice of a, b. (d) shows where ψ_1 sends z_1, z_2, a, b.

1. l is a vertical line and $\Im(z_1) < \Im(z_2)$.
2. l is a vertical line and $\Im(z_1) > \Im(z_2)$.
3. l is not a vertical line.

In the first case, define $a = \Re(z_1)$ and $b = \infty$. In the second case, define $a = \infty$ and $b = \Re(z_1)$. The third case is the trickiest: consider the perpendicular bisector of the line segment from z_1 to z_2. This perpendicular bisector intersects $\mathbb{R}$ at some point; construct a circle through z_1 and z_2 whose center is that point. This circle is perpendicular to $\mathbb{R}$—define a and b to be the points of intersection, chosen such that there exists a path from a to z_1 to z_2 to b that traverses the circle either clockwise or counterclockwise. All three of these constructions are depicted in Figure 4.5.

Now, suppose that we choose two elements $\psi, \phi \in \text{Isom}(\mathbb{H}^2)$ such that $\psi(z_1) = \phi(z_1) = i$, and $\psi(z_2) = it_1$, $\phi(z_2) = it_2$ for some $t_1, t_2 > 1$. Let $\varphi = \psi \circ \phi^{-1}$. Then $\varphi(i) = i$, and $\varphi(it_2) = it_1$. Consider the line $x = 0$. It is perpendicular to $\mathbb{R}$—since φ preserves $\mathbb{R}$, its image must be a generalized circle orthogonal to $\mathbb{R}$. It also passes through the points i and it_2, so its image must pass through the points i and it_1. However, there is only one generalized circle that passes through those points and is orthogonal to $\mathbb{R}$, and that is the line $x = 0$ itself. Therefore, φ must either fix both 0 and ∞, or it must swap them. Since any element of $\text{Möb}^0(2)$ is fully determined by where it sends three points and we know that $\varphi(i) = i$, $\varphi(0) \in \{0, \infty\}$, $\varphi(\infty) \in \{0, \infty\}$, this leaves us with four possibilities for φ, namely

$$\varphi_1(z) = z \qquad \varphi_2(z) = -\overline{z}$$
$$\varphi_3(z) = -z^{-1} \; \varphi_4(z) = \overline{z}^{-1}.$$

Since we also know that $\varphi(it_2) = it_1$ and $t_1, t_2 > 1$, this rules out the second two possibilities, leaving just two cases. In both, $\varphi(it_2) = it_2$, hence $t_1 = t_2$. □

4.3 The Poincaré Half-Plane

We are finally ready to construct a metric on $\mathbb{H}^2$ such that the corresponding isometry group is $\text{Isom}(\mathbb{H}^2)$. Strictly speaking, this sole requirement is insufficient to uniquely determine the metric even up to scaling—see Exercise 4.5.5. So, we should ask for some additional nice property. There are many possible choices, but here is a simple one: we'll ask that the line $x = 0$ acts like a line in Euclidean space, in the sense that if $d : \mathbb{H}^2 \to [0, \infty)$ is the metric, then $d(it_1, it_2) + d(it_2, it_3) = d(it_1, it_3)$ for any $0 < t_1 < t_2 < t_3$. This additional requirement is just about enough. We start by defining d on the line $x = 0$.

Lemma 4.3 *Let L be the half-line $x = 0$, $y > 0$. For any $r > 0$, there exists a unique function $d_r : L \times L \to [0, \infty)$ such that*

1. $d_r(i, ei) = r$,
2. $d_r(x, y) = d_r(y, x)$ *for any* $x, y \in L$,
3. *for any* $0 \leq t_1 \leq t_2 \leq t_3$, $d_r(it_1, it_2) + d_r(it_2, it_3) = d_r(it_1, it_3)$, *and*
4. *for any* $\gamma \in \text{Isom}(\mathbb{H}^2)$ *such that* $\gamma(L) = L$, *and any* $x, y \in L$, $d_r(x, y) = d_r(\gamma(x), \gamma(y))$.

Concretely,

$$d_r(z_1, z_2) = r \left| \ln\left(\frac{z_2}{z_1}\right) \right| .$$

Remark 4.2 To clear up any confusion: the 'e' in the statement of the lemma is just Euler's number. We could do the argument with e replaced with any other positive real number, but we will see that this is a convenient choice.

Proof If $\gamma \in \text{Isom}(\mathbb{H}^2)$ and $\gamma(L) = L$, then γ fixes the set $\{0, \infty\}$. In $SL(2, \mathbb{R})$, there are only two types of matrices that do this, namely

$$\begin{pmatrix} \lambda & 0 \\ 0 & 1/\lambda \end{pmatrix}, \begin{pmatrix} 0 & \lambda \\ -1/\lambda & 0 \end{pmatrix}$$

for some $\lambda \in \mathbb{R}^\times$. Consequently, we know that $\gamma(z) = \lambda^2 z$, $-\lambda^2/z$, $-\lambda^2\bar{z}$, or $\lambda^2/\bar{z}$. Note that $z \mapsto -\bar{z}$ fixes every point in L, so in fact if $d_r(x, y) = d_r(\gamma(x), \gamma(y))$ for $\gamma(z) = \lambda^2 z$ and $-\lambda^2/z$, then it immediately follows that it is true for the other two. Thus, it shall actually suffice to check that $d_r(x, y) = d_r(\gamma(x), \gamma(y))$ for all $x, y \in L$, $\gamma(z) = \lambda^2 z$ and $\gamma(z) = -1/z$.

We'll start by determining what $d_r(e^{l/n}i, e^{(l+1)/n}i)$ is for $l, n \in \mathbb{Z}$ with $0 \leq l < n$. If we take $\gamma(z) = e^{1/n}z$, then $\gamma(e^{l/n}i) = e^{(l+1)/n}i$ and $\gamma(e^{(l+1)/n}i) = e^{(l+2)/n}i$.

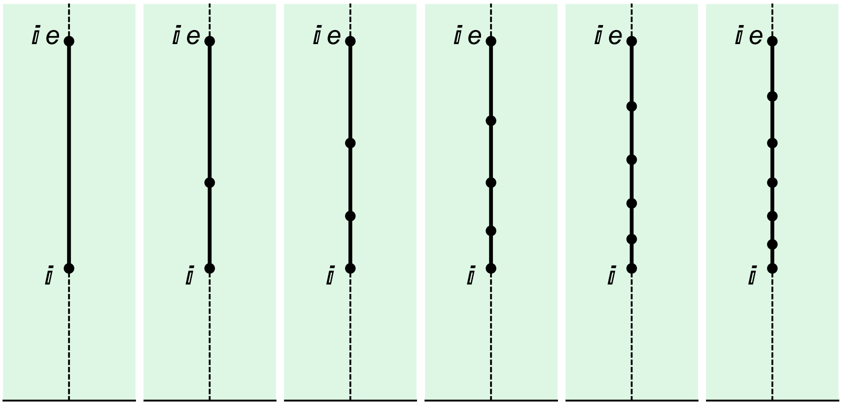


Fig. 4.6 From left to right, the same line segment in $\mathbb{H}^2$ is partitioned into smaller and smaller segments of equal (hyperbolic) length.

From this, we know that $d_r(e^{l/n}i, e^{(l+1)/n}i)$ does not depend on the choice of l. This partitioning is illustrated in Figure 4.6. Using this, we get

$$\begin{aligned} r = d_r(i, ei) &= d_r(i, e^{1/n}i) + d_r(e^{1/n}i, e^{2/n}i) + \ldots + d_r(e^{(n-1)/n}i, ei) \\ &= nd_r(e^{l/n}i, e^{(l+1)/n}i), \end{aligned}$$

so $d_r(e^{l/n}i, e^{(l+1)/n}i) = r/n$. What we have done is partition the line segment from i to ei into pieces of equal length, as in Figure 4.6. We deduce that $d_r(i, e^{l/n}i) = rl/n$ for any $l, n \in \mathbb{Z}$ such that $0 \leq l/n \leq 1$. By applying a transformation $z \mapsto tz$, we can conclude that $d_r(ti, te^q i) = rq$ for any $q \in \mathbb{Q} \cap [0, 1]$. However, for any $q \in \mathbb{Q}$ such that $q > 1$, there exists some $n \in \mathbb{N}$ such that $0 \leq q/n \leq 1$, and therefore

$$\begin{aligned} d_r(ti, te^q i) &= d_r(ti, te^{q/n}i) + d_r(te^{q/n}i, te^{2q/n}i) + \ldots + d_r(te^{q(n-1)/n}i, te^q i) \\ &= d_r(ti, te^{q/n}i) + \ldots + d_r(e^{q(n-1)/n}(ti), e^{q(n-1)/n}(te^{q/n}i)) \\ &= nd_r(ti, te^{q/n}i) = n\,(rq/n) = rq. \end{aligned}$$

So, for any positive rational q and $z \in L$, we know that $d_r(z, e^q z) = d_r(e^q z, z) = rq$. If we replace z with $e^{-q}z$, though, then the above gives us that $d_r(e^{-q}z, z) = d_r(z, e^{-q}z) = rq$. Thus, what we have so far demonstrated is that for any $z \in L$ and $q \in \mathbb{Q}$, $d_r(z, e^q z) = d_r(e^q z, z) = r|q|$. Our next objective must be to extend this so that it applies to all of $\mathbb{R}$ and not merely $\mathbb{Q}$. This can be accomplished as follows. Choose any $t \in \mathbb{R}$ and any $q_1, q_2 \in \mathbb{Q}$ such that $q_1 \leq t \leq q_2$. Since,

$$\begin{aligned} d_r(i, e^{q_1}i) &\leq d_r(i, e^{q_1}i) + d_r(e^{q_1}i, e^t i) = d_r(i, e^t i) \\ d_r(i, e^t i) &\leq d_r(i, e^t i) + d_r(e^t i, e^{q_2}i) = d_r(i, e^{q_2}i), \end{aligned}$$

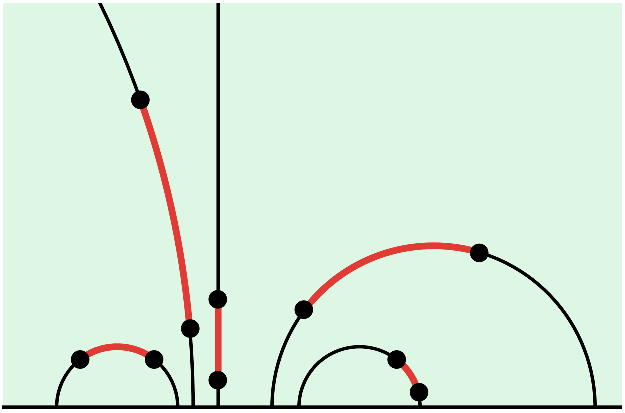


Fig. 4.7 Each of the red segments is the image of each other under some element in Isom($\mathbb{H}^2$). In particular, the (hyperbolic) distance between any two endpoints is exactly the same.

we have that $r|q_1| = d_r(i, e^{q_1}i) \leq d_r(i, e^t i) \leq d_r(i, e^{q_2}i) = r|q_2|$. But we can force q_1 and q_2 to be arbitrarily close to t, so we see that in fact it must be that $d_r(i, e^t i) = r|t|$ for any $t \in \mathbb{R}$. From, this it follows instantly that $d_r(z, e^t z) = r|t|$ for any $z \in L$ and $t \in \mathbb{R}$, which in turn can be restated as $d_r(z, \lambda z) = r|\ln(\lambda)|$ for any $z \in L$, $\lambda \in (0, \infty)$. Of course, for any $z_1, z_2 \in L$, $z_1/z_2 \in (0, \infty)$, so we can state this as

$$d_r(z_1, z_2) = d_r\left(z_1, \frac{z_2}{z_1}z_1\right) = r\left|\ln\left(\frac{z_2}{z_1}\right)\right|$$

for all $z_1, z_2 \in L$. It is easy to check that this is indeed invariant under $z \mapsto \lambda^2 z$ and $z \mapsto -1/z$, and so we are done. □

Having defined this function on the line $x = 0$, we can now uniquely extend it to all of $\mathbb{H}^2$, allowing us to compare the lengths of arbitrarily line segments, as in Figure 4.7.

Lemma 4.4 *For any $r > 0$, there exists a unique function $d_r : \mathbb{H}^2 \times \mathbb{H}^2 \to [0, \infty)$ such that*

1. $d_r(i, ei) = r$,
2. $d_r(x, y) = d_r(y, x)$ *for any* $x, y \in \mathbb{H}^2$,
3. *for any* $0 \leq t_1 \leq t_2 \leq t_3$, $d_r(it_1, it_2) + d_r(it_2, it_3) = d_r(it_1, it_3)$, *and*
4. *for any* $\gamma \in$ *Isom*($\mathbb{H}^2$), $d_r(x, y) = d_r(\gamma(x), \gamma(y))$ *if* $x, y \in \mathbb{H}^2$.

Proof We know that if there exists such a function, then for all $z_1, z_2 \in \mathbb{H}^2$ with $\Re(z_1) = \Re(z_2) = 0$,

$$d_r(z_1, z_2) = d_r\left(z_1, \frac{z_2}{z_1}z_1\right) = r\left|\ln\left(\frac{z_2}{z_1}\right)\right|.$$

On the other hand, by Theorem 4.3, we know that for any two distinct elements $z_1, z_2 \in \mathbb{H}^2$, there exists an element $\gamma \in \text{Isom}(\mathbb{H}^2)$ such that $\gamma(z_1) = i$, and $\gamma(z_2) = it$ for some $t > 1$. Consequently, it must be that

$$d_r(z_1, z_2) = d_r(\gamma(z_1), \gamma(z_2)) = d_r(i, it) = r\,|\ln(t)|\,.$$

This implies that there is at most one function d_r satisfying the desired properties. To prove that there is such a function, we define

$$\begin{aligned} d_r : \mathbb{H}^2 \times \mathbb{H}^2 &\to [0, \infty) \\ (z_1, z_2) &\mapsto \begin{cases} 0 & \text{if } z_1 = z_2 \\ r\ln(t) & \begin{array}{l}\text{if } \gamma(z_1) = i, \gamma(z_2) = it \text{ with } t > 1, \\ \gamma \in \text{Isom}(\mathbb{H}^2).\end{array} \end{cases} \end{aligned}$$

This function is in fact well-defined—while there are multiple elements γ with the desired property, we know by Theorem 4.3 that they will all send z_2 to the same point it. It is easy to see that $d_r(i, ei) = r$ and that $d_r(\gamma(z_1), \gamma(z_2)) = d_r(z_1, z_2)$ for any $z_1, z_2 \in \mathbb{H}^2$ and $\gamma \in \text{Isom}(\mathbb{H}^2)$. It remains to check that $d_r(z_1, z_2) = d_r(z_2, z_1)$. Let γ be such that $\gamma(z_1) = i$ and $\gamma(z_2) = it$. Then

$$d_r(z_2, z_1) = d_r(it, i) = d_r(i, it) = d_r(z_1, z_2).$$

Thus, d_r has all of the required properties. □

It would be good to have a more explicit formula for this function d_r. To get this, we prove a simple classical result.

Theorem 4.4 ($SL(2, \mathbb{R})$ **Imaginary Transformation Law**)
For any $z \in \mathbb{H}^2$ *and*

$$\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL(2, \mathbb{R}),$$

$\Im(\gamma.z) = |cz + d|^{-2}\Im(z)$.

Proof This is a straightforward calculation.

$$\begin{aligned} \Im\left(\frac{az+b}{cz+d}\right) &= |cz+d|^{-2}\Im\left((az+b)\overline{(cz+d)}\right) \\ &= |cz+d|^{-2}\Im\left((az+b)(c\overline{z}+d)\right) \\ &= |cz+d|^{-2}\Im\left(ac|z|^2 + adz + bc\overline{z} + bd\right) \\ &= |cz+d|^{-2}\left(ad\Im(z) - bc\Im(z)\right) = |cz+d|^{-2}\Im(z). \end{aligned}$$

Lemma 4.5 *The unique function* $d_r : \mathbb{H}^2 \times \mathbb{H}^2 \to [0, \infty)$ *defined in Lemma 4.4 is given by*

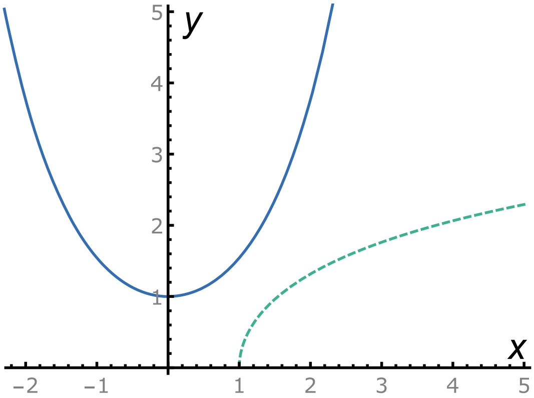


Fig. 4.8 A plot of $y = \cosh(x)$ in blue and a plot of $y = \cosh^{-1}(x)$ in green, dashed.

$$\begin{aligned} d_r : \mathbb{H}^2 \times \mathbb{H}^2 &\to [0, \infty) \\ (z_1, z_2) &\mapsto r \cosh^{-1}\left(1 + \frac{|z_2 - z_1|^2}{2\Im(z_1)\Im(z_2)}\right), \end{aligned}$$

where $\cosh : \mathbb{R} \to [1, \infty)$ *is the hyperbolic cosine, defined by*

$$\cosh(x) = \frac{e^x + e^{-x}}{2}.$$

Remark 4.3 There is something to prove here: namely, that cosh is actually invertible in any sense. Indeed, it is not if we consider it over its full domain—however, if we restrict to $(0, \infty)$, then everything goes through as it should. This is illustrated in Figure 4.8. For this and more about the other hyperbolic functions, see Exercises 4.2.1 and 4.3.1.

Proof By Theorem 4.3, we know that there exists some

$$\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL(2, \mathbb{R})$$

such that $z_1 = \gamma.i$ and $z_2 = \gamma.it$ for some $t > 1$. Furthermore, we know from the preceding two lemmas that $r \ln(t) = d_r(z_1, z_2)$, hence $t = e^{d_r(z_1,z_2)/r}$. This means that

$$\begin{aligned}|z_1 - z_2|^2 = |\gamma.i - \gamma.it|^2 &= \left|\frac{ai+b}{ci+d} - \frac{ait+b}{cit+d}\right|^2 \\ &= |ci+d|^{-2}\,|cit+d|^{-2}\,|(ai+b)(cit+d) - (ait+b)(ci+d)|^2 \\ &= |ci+d|^{-2}\,|cit+d|^{-2}\,|-act+adi+bcti+bd+act-adti-bci-bd|^2 \\ &= |ci+d|^{-2}\,|cit+d|^{-2}\,|ad(1-t)i - bc(1-t)i|^2 \\ &= |ci+d|^{-2}\,|cit+d|^{-2}\,(1-t)^2.\end{aligned}$$

Here the $SL(2,\mathbb{R})$ imaginary transformation law is very useful, since this means that

$$\begin{aligned}\frac{|z_1-z_2|^2}{\Im(z_1)\Im(z_2)} &= \frac{|ci+d|^{-2}\,|cit+d|^{-2}\,(1-t)^2}{\Im(\gamma.i)\Im(\gamma.it)} \\ &= \frac{|ci+d|^{-2}\,|cit+d|^{-2}\,(1-t)^2}{|ci+d|^{-2}\,|cit+d|^{-2}\,\Im(i)\Im(it)} = \frac{(1-t)^2}{t}.\end{aligned}$$

It is now a matter of solving for t and then for $d_r(z_1, z_2)$. We see that

$$\begin{aligned}1 + \frac{|z_1-z_2|^2}{2\Im(z_1)\Im(z_2)} &= 1 + \frac{(1-t)^2}{2t} = \frac{1}{2}\left(t + \frac{1}{t}\right) \\ &= \frac{1}{2}\left(e^{\frac{d_r(z_1,z_2)}{r}} + e^{\frac{-d_r(z_1,z_2)}{r}}\right) = \cosh\left(\frac{d_r(z_1,z_2)}{r}\right),\end{aligned}$$

and so the stated result follows. □

In some sense, the choice of $r > 0$ is arbitrary, but it is customary to set $r = 1$. There are many ways to motivate this—the most obvious at present is that with the explicit description of d_r that we have given, it is the most convenient choice. One can also look at the behavior of d_r close to i or, if one is familiar with Riemannian geometry, at the curvature. In any case, we are finally ready to define hyperbolic distance properly.

Theorem 4.5 *Define*

$$\begin{aligned}d_{hyper} : \mathbb{H}^2 \times \mathbb{H}^2 &\to [0,\infty) \\ (z_1, z_2) &= \cosh^{-1}\left(1 + \frac{|z_1-z_2|^2}{2\Im(z_1)\Im(z_2)}\right).\end{aligned}$$

Then $(\mathbb{H}^2, d_{hyper})$ *is a metric space.*

Remark 4.4 This is commonly referred to as the *Poincaré upper half-plane model of hyperbolic space*.

Proof In terms of proving that $(\mathbb{H}^2, d_{\text{hyper}})$ is a metric space, the only thing that is unclear is whether it satisfies the triangle inequality. The intuitive idea behind the proof is to find a way to "flatten" an arbitrary triangle in $\mathbb{H}^2$ onto the line $x = 0$, as in

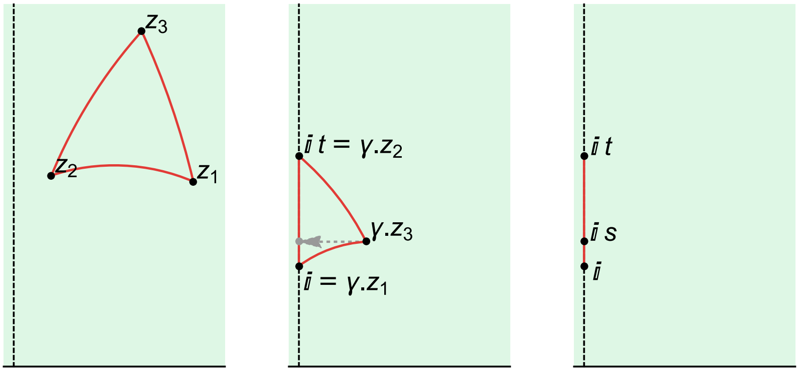


Fig. 4.9 A visualization of the proof of Theorem 4.5. We start with three points z_1, z_2, z_3 in general position. We then apply an element $\varphi \in \mathrm{Isom}(\mathbb{H}^2)$ to move z_1 to i and z_2 to it. Finally, we "flatten" $\varphi(z_3)$ onto the y-axis in a way that can only decrease hyperbolic distance.

Figure 4.9. To prove this formally, start with the following observation: if $z_1, z_2 \in \mathbb{H}^2$, define $w_1 = \mathfrak{I}(z_1)i$ and $w_2 = \mathfrak{I}(z_2)i$; then $d_{\text{hyper}}(z_1, z_2) \geq d_{\text{hyper}}(w_1, w_2)$. Why is this? Well,

$$\begin{aligned} 1 + \frac{|z_1 - z_2|^2}{2\mathfrak{I}(z_1)\mathfrak{I}(z_2)} &\geq 1 + \frac{|w_1 - w_2|^2}{2\mathfrak{I}(z_1)\mathfrak{I}(z_2)} \\ &= 1 + \frac{|w_1 - w_2|^2}{2\mathfrak{I}(w_1)\mathfrak{I}(w_2)} \end{aligned}$$

and $\cosh^{-1}$ is order-preserving in the sense that if $x \leq y$ then $\cosh^{-1}(x) \leq \cosh^{-1}(y)$. (See Exercise 4.3.1.) So, choose any $z_1, z_2, z_3 \in \mathbb{H}^2$. By applying Theorem 4.3, we know that there exists some $\gamma \in SL(2, \mathbb{R})$ such that $\gamma.z_1 = i$, $\gamma.z_3 = it$ for some $t > 1$; we also know that $d_{\text{hyper}}(\gamma.z, \gamma.w) = d_{\text{hyper}}(z, w)$ for any $z, w \in \mathbb{H}^2$. Thus, to prove that

$$d_{\text{hyper}}(z_1, z_3) \leq d_{\text{hyper}}(z_1, z_2) + d_{\text{hyper}}(z_2, z_3),$$

it actually suffices to prove that

$$d_{\text{hyper}}(i, it) \leq d_{\text{hyper}}(i, \gamma.z_2) + d_{\text{hyper}}(\gamma.z_2, it).$$

Here we use the trick that we can "flatten" $\gamma.z_2$ onto the line $x = 0$—we know that $d_{\text{hyper}}(i, \gamma.z_2) \geq d_{\text{hyper}}(i, \mathfrak{I}(\gamma.z_2)i)$ and $d_{\text{hyper}}(\gamma.z_2, it) \geq d_{\text{hyper}}(\mathfrak{I}(\gamma.z_2)i, it)$, and so actually it suffices to prove that for any $t > 1$ and any $s > 0$,

$$d_{\text{hyper}}(i, it) \leq d_{\text{hyper}}(i, is) + d_{\text{hyper}}(is, it).$$

But this is immediate from the additive property of d_{hyper} along the line $x = 0$. Thus, $(\mathbb{H}^2, d_{\text{hyper}})$ is a metric space. □

By construction, $\mathrm{Isom}(\mathbb{H}^2)$ consists of isometries of this metric space. We will show later, after we have a better feel for the geometry of this space, that it consists of all of the isometries. As a step in this direction, let's try to understand this weird metric that we have constructed by thinking about what happens if we choose two points that are very close together. One can show that

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\cosh^{-1}(x)\right) = \frac{1}{\sqrt{x^2-1}}$$

(see Exercise 4.3.1), so

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(\cosh^{-1}\left(1+\frac{x^2}{2}\right)\right) = \frac{x}{\sqrt{\left(1+\frac{x^2}{2}\right)^2-1}} = \frac{2x}{\sqrt{x^2(x^2+4)}} = \mathrm{sgn}(x)\frac{2}{\sqrt{x^2+4}},$$

where $\mathrm{sgn}(x) = 1$ if $x > 0$ and -1 if $x < 0$. This means that we can do a linear approximation for $f(x) = \cosh^{-1}(1+x^2/2)$ in the region $x > 0$ by

$$f(0) + x \lim_{t\to 0^+} f'(t) = x,$$

which will be roughly accurate as long as x is very small. What does this have to do with our metric? Suppose that $z_1, z_2 \in \mathbb{H}^2$ are close together, so $|z_1 - z_2|/y \approx 0$, where $y = \Im(z_1)$. Then

$$\begin{aligned} d_{\text{hyper}}(z_1, z_2) &= \cosh^{-1}\left(1 + \frac{|z_1 - z_2|^2}{2\Im(z_1)\Im(z_2)}\right) \\ &\approx \cosh^{-1}\left(1 + \frac{|z_1 - z_2|^2}{2y^2}\right) \\ &\approx \frac{|z_1 - z_2|}{y} = \frac{1}{y} d_{\text{Euclid}}(z_1, z_2). \end{aligned}$$

Ah-ha! What we have determined is that if we just look at points that are quite close together, the hyperbolic metric is essentially just the Euclidean one, but scaled according to the distance above the x-axis. In particular, we see what is happening is that as we get closer and closer to the line $y = 0$, the more space is compacted—shorter and shorter Euclidean distances correspond to larger and larger hyperbolic distances. For a visualization of this, see Figure 4.10.

▶ **Example** *Let $z_1 \neq z_2 \in \mathbb{H}^2$ be two points such that $d_{\text{hyper}}(z_1, z_2) = \lambda$. Suppose Ψ is an isometry of $\mathbb{H}^2$ such that $\Psi(z_1) = i$ and $\Psi(z_2) = it$ for some $t > 1$.* Find *t as a function of λ.*

Since Ψ is an isometry, $d_{\text{hyper}}(\Psi(z_1), \Psi(z_2)) = \lambda$ by definition. Therefore,

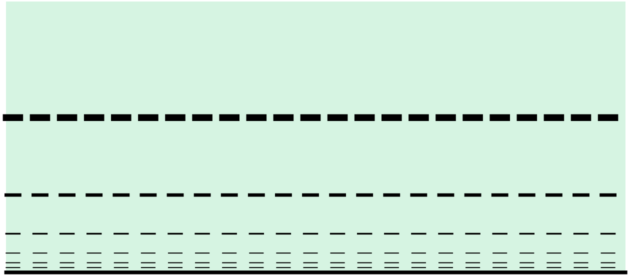


Fig. 4.10 All of the dashed curves are equidistant from one to the next in the hyperbolic metric.

$$\lambda = d_{\text{hyper}}(i, it) = \left| \ln \left(\frac{it}{i} \right) \right| = \ln(t),$$

so $t = e^{\lambda}$.

4.4 Circles and Lines

Our goal in this section is to find hyperbolic analogs of Euclidean circles and lines. The first definition is quite straightforward, as it is exactly the original Euclidean definition, but with d_{Euclid} replaced by d_{hyper}.

Definition 4.6 For any $z \in \mathbb{H}^2$ and $r > 0$, the *hyperbolic circle with center z and radius r* is the locus of points $w \in \mathbb{H}^2$ such that $d_{\text{hyper}}(w, z) = r$.

What does a hyperbolic circle look like? Let's first work this out in the special case where the center is i.

Lemma 4.6 *The hyperbolic circle with center i and radius r is the Euclidean circle with center* $\cosh(r)i$ *and radius* $\sinh(r)$*, where* $\sinh : \mathbb{R} \to \mathbb{R}$ *is the hyperbolic sine, defined by*

$$\sinh(x) = \frac{1}{2}\left(e^{x} - e^{-x}\right).$$

Proof If w is a point on this hyperbolic circle, then

$$d_{\text{hyper}}(w, i) = \cosh^{-1}\left(1 + \frac{|w - i|^2}{2\Im(w)}\right) = r,$$

or equivalently,

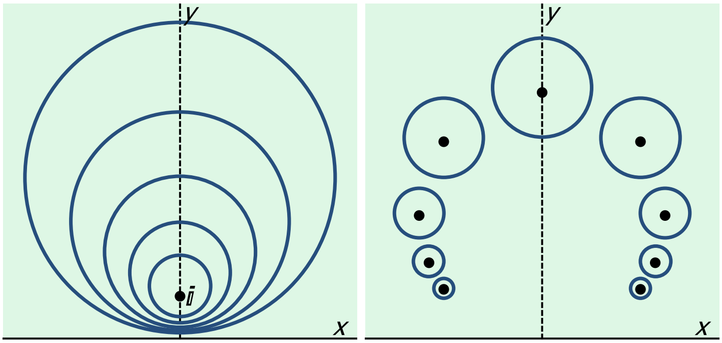


Fig. 4.11 On the left, a family of hyperbolic circles with center i. On the right, a collection of hyperbolic circles all with radius $1/5$. The one furthest from the x-axis has hyperbolic center $i/4$.

$$1 + \frac{|w - i|^2}{2\Im(w)} = \cosh(r)$$

$$|w - i|^2 + 2(1 - \cosh(r))\Im(w) = 0.$$

Let's write $w = x + iy$. Then the above can be rewritten as

$$x^2 - 2y^2 + y^2 + 1 + 2(1 - \cosh(r))y = 0,$$

which we recognize as the equation of a Euclidean circle. Determining the center and radius of this circle is just an exercise in completing the square, which I leave as an exercise. (See Exercise 4.2.4.) □

This is enough to determine what hyperbolic circles are like in general.

Theorem 4.6 *Hyperbolic circles are Euclidean circles (but with different centers and radii).*

Remark 4.5 To better understand the differences, some families of hyperbolic circles are shown in Figure 4.11.

Proof If C is the hyperbolic circle with center z and radius r, then we can choose an element $\varphi \in PSL(2, \mathbb{R})$ such that $\varphi(z) = i$. Then $\varphi(C)$ is the hyperbolic circle with center i and radius r, which we know from the previous lemma is just a Euclidean circle. We know that linear fractional transformations preserve generalized circles, so that means that $C = \varphi^{-1}(\varphi(C))$ is a generalized circle contained inside $\mathbb{H}^2$—which must simply be a Euclidean circle. □

Marvelous! Next, we give the same treatment to hyperbolic lines. It is less obvious how to define these. One potential suggestion would be to say that a line should be the shortest path between two points $z_1, z_2 \in \mathbb{H}^2$, as measured with respect to the hyperbolic distance. This is possible to make sense of but complicated. We will

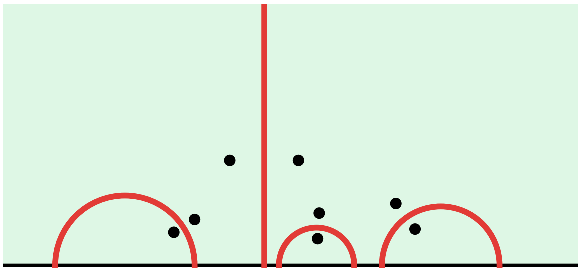


Fig. 4.12 Pairs of points in the hyperbolic plane and the hyperbolic lines that correspond to them.

instead proceed as follows: recall that a line in the Euclidean plane can be thought of as the locus of points equidistant from two fixed points z_1, z_2—that is, all $w \in \mathbb{C}$ such that $d_{\text{Euclid}}(w, z_1) = d_{\text{Euclid}}(w, z_2)$. This fact was used in the proof of Theorem 2.10, the Algebraic Description of Generalized Circles.

Definition 4.7 A *hyperbolic line in* $\mathbb{H}^2$ is a locus of points in $\mathbb{H}^2$ of the form

$$\left\{ w \in \mathbb{H}^2 \middle| d_{\text{hyper}}(w, z_1) = d_{\text{hyper}}(w, z_2) \right\},$$

for some fixed points $z_1 \neq z_2 \in \mathbb{H}^2$.

Some examples of such loci are depicted in Figure 4.12, suggesting that they are really just Euclidean circles of some kind. As before, our method to actually prove this is to first work out a simple case.

Lemma 4.7 *Let $z_1 \neq z_2 \in \mathbb{H}^2$ such that $z_1 = -\overline{z_2}$. The corresponding hyperbolic line is $x = 0$ (restricted to $\mathbb{H}^2$).*

Proof By definition, $w \in \mathbb{H}^2$ is on this line if and only if

$$\begin{aligned} d_{\text{hyper}}(w, z_1) &= \cosh^{-1}\left(1 + \frac{|w - z_1|^2}{2\Im(w)\Im(z_1)}\right) \\ &= \cosh^{-1}\left(1 + \frac{|w - z_2|^2}{2\Im(w)\Im(z_2)}\right) = d_{\text{hyper}}(w, z_2), \end{aligned}$$

or equivalently if

$$\frac{|w - z_1|^2}{\Im(z_1)} = \frac{|w - z_2|^2}{\Im(z_2)}.$$

If $z_1 = -\overline{z_2}$, then $\Im(z_1) = \Im(z_2)$, so this further simplifies to just $|w - z_1| = |w - z_2|$. This is just the equation defining the Euclidean line of points equidistant from z_1 and z_2—that line is just $x = 0$. □

It is immediate from this lemma that any generalized circle orthogonal to the real line must be a hyperbolic line—$x = 0$ is simply the special case where it is orthogonal at 0 and ∞, but since we know that $\mathrm{Isom}(\mathbb{H}^2)$ allows us to freely move points on the boundary but preserves circles and angles, we get the desired conclusion. What is less obvious is that any hyperbolic line has to be of such a form, and we couldn't possibly get any pathological counterexamples. To prove that, we need to show that we can move arbitrary points in $\mathbb{H}^2$ into the special position of Lemma 4.7. We will do it using a sequence of lemmas.

Lemma 4.8 *Let $z_1 \neq z_2 \in \mathbb{H}^2$. There exists a unique generalized circle passing through both z_1 and z_2 which is orthogonal to the real line.*

Proof Choose $\varphi \in PSL(2, \mathbb{R})$ such that $\varphi(z_1) = i$ and $\varphi(z_2) = it$ for some $t > 1$; since γ preserves generalized circles, $\mathbb{R}$, and angles, there exists a unique generalized circle with the desired properties if and only if there exists a unique generalized circle through i and it which is orthogonal to $\mathbb{R}$. If C is a generalized circle with inversive coordinates (κ, κ', ξ), then it is orthogonal to $\mathbb{R}$ if and only if

$$\frac{1}{2}\mathrm{tr}\left(\begin{pmatrix}\kappa' & \xi \\ \overline{\xi} & \kappa\end{pmatrix}\begin{pmatrix}0 & i \\ -i & 0\end{pmatrix}^{-1}\right) = \Im(\xi) = 0,$$

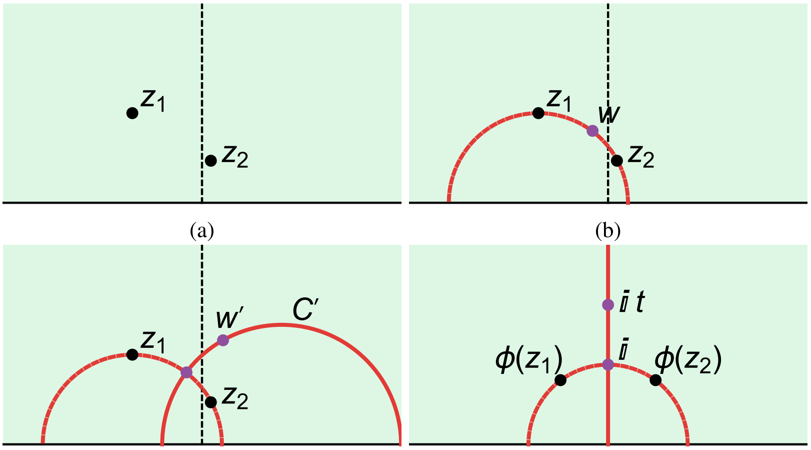
which is to say that its bend-center is real. On the other hand, if C is a circle passing through i and it, then its center is $x + it/2$ for some $x \in \mathbb{R}$, so its bend-center is non-real. Therefore, if C is a generalized circle that passes through i and it and is orthogonal to $\mathbb{R}$, then it is a line—specifically, it has to be the line $x = 0$. □

Lemma 4.9 *Let C be a generalized circle orthogonal to the real line and $z \in \mathbb{H}^2$ be a point on C. There exists a unique generalized circle passing through z which is orthogonal to both the real line and C.*

Proof Choose any other point $z' \in \mathbb{H}^2$ on C and an element $\varphi \in PSL(2, \mathbb{R})$ such that $\varphi(z) = i$ and $\varphi(z') = it$. It is clear that $\varphi(C)$ is the line $x = 0$. If we can show that there exists a unique generalized circle through i which is orthogonal to both $x = 0$ and $y = 0$, we will be done. Under what circumstances is a generalized circle orthogonal to both $x = 0$ and $y = 0$? This happens exactly when it is a circle centered at the origin. (See Exercise 4.2.5.) Such a circle passes through i if and only if it is the unit circle. □

Lemma 4.10 *For any $z_1 \neq z_2 \in \mathbb{H}^2$, there exists $\varphi \in \mathit{Isom}(\mathbb{H}^2)$ such that $\varphi(z_1) = -\overline{\varphi(z_2)}$.*

Proof We proceed geometrically; the essential steps are depicted in Figure 4.13. First, draw a hyperbolic line C through z_1 and z_2—that is, a generalized circle orthogonal to the boundary of $\mathbb{H}^2$. We know there is a unique such circle thanks to Lemma 4.8. Next, choose a point w on this line such that $d_{\text{hyper}}(w, z_1) = d_{\text{hyper}}(w, z_2)$. Why must there be such a point? In principle, we could try to directly solve for it, but it is easier to consider a path $p(t)$ from z_1 to z_2 along C, such that $p(0) = z_1$ and $p(1) = z_2$. Consider the function



(a)

(b)

(c)

(d)

Fig. 4.13 A visualization of the proof of Lemma 4.10. We start with a pair of points $z_1, z_2 \in \mathbb{H}^2$ in (a). In (b), we draw a hyperbolic line through this pair and find the midpoint w between them. In (c), we draw another hyperbolic line C' perpendicular to the first through the midpoint and choose another point w' on this perpendicular line. In (d), we choose an element in $\phi \in PSL(2, \mathbb{R})$ which sends the midpoint to i and the other point on our perpendicular bisector to it. This forces us into the desired configuration.

$$f(t) = d_{\text{hyper}}(p(t), z_1) - d_{\text{hyper}}(p(t), z_2).$$

Clearly, $f(0) = -d_{\text{hyper}}(z_1, z_2) < 0$ and $f(1) = d_{\text{hyper}}(z_1, z_2) > 0$. Therefore, there has to exist some $0 < t < 1$ where $f(t) = 0$—but that is exactly a point such that $w = p(t)$ is equidistant from z_1 and z_2. Using Lemma 4.9, construct a generalized circle C' through this point w orthogonal to both $\mathbb{R}$ and C. Now, choose any other point $w' \in \mathbb{H}^2$ on C' and an element $\varphi \in PSL(2, \mathbb{R})$ such that $\varphi(w) = i$ and $\varphi(w') = it$ for some $t > 1$. This forces $\varphi(C')$ to be the line $x = 0$ and $\varphi(C)$ to be the unit circle. Therefore, $\varphi(z_1) = e^{i\theta_1}$, $\varphi(z_2) = e^{i\theta_2}$ for some $0 < \theta_1, \theta_2 < \pi$. Since w was equidistant from z_1 and z_2, i is equidistant from $\varphi(z_1)$ and $\varphi(z_2)$, which is to say that

$$\cosh^{-1}\left(1 + \frac{\left|e^{i\theta_1} - i\right|^2}{2\sin(\theta_1)}\right) = \cosh^{-1}\left(1 + \frac{\left|e^{i\theta_2} - i\right|^2}{2\sin(\theta_2)}\right),$$

or equivalently,

$$\frac{\left|e^{i\theta_1} - i\right|^2}{2\sin(\theta_1)} = \frac{\left|e^{i\theta_2} - i\right|^2}{2\sin(\theta_2)}.$$

This, too, we can simplify, since $|e^{i\theta} - i|^2 = 2(1 - \sin(\theta))$, and so we see that in fact when all is said and done we must have $\sin(\theta_1) = \sin(\theta_2)$. The only way for this to happen is if $\theta_2 = \pi - \theta_1$, hence $\varphi(z_1) = e^{i\theta_1} = -\overline{(-e^{-i\theta_1})} = -\overline{\varphi(z_2)}$. □

We see that we have proved the following.

Theorem 4.7 *Hyperbolic lines are generalized circles orthogonal to the real line (restricted to $\mathbb{H}^2$). Furthermore, they have the following properties.*

1. *For any $z_1 \neq z_2 \in \mathbb{H}^2$, there exists a unique hyperbolic line passing through them both.*
2. *For any hyperbolic line l and a point $z \in l$, there exist a unique hyperbolic line l' that is orthogonal to l at z.*

Proof If l is a hyperbolic line defined by a pair of points z_1, z_2, we know by Lemma 4.10 that we can choose $\varphi \in \mathrm{Isom}(\mathbb{H}^2)$ such that $\varphi(z_1) = -\overline{\varphi(z_2)}$, hence $\varphi(l)$ is the line $x = 0$ by Lemma 4.7. As we discussed before, this is sufficient to prove that hyperbolic lines are exactly the generalized circles orthogonal to $\mathbb{R}$. The two additional properties are simply the statements of Lemma 4.8 and 4.9 in this new language. □

Before we conclude this section, I want to point out something about our definitions: specifically, it is quite obvious that you could extend both of them to any metric space whatsoever, given that they are entirely phrased in terms of the metric d_{hyper}. Indeed, both of these definitions do show up in the metric geometry literature, but under different names. What we have called a hyperbolic circle is in general known as a sphere; similarly, what we know as a hyperbolic line is in general known as a hyper-plane. The difference in terminology is simply because the hyperbolic plane is "low-dimensional" in a sense; indeed, one sees that in $(\mathbb{R}^3, d_{\text{hyper}})$, what we have termed a circle will be a sphere and what we have termed a line will be a plane. (See Exercise 4.2.6.)

▶ **Example** *Let $C_{\lambda,r}$ be the hyperbolic circle with center λi and radius r, where $\lambda, r > 0$. Find the Euclidean center and radius of C_r as functions of λ* and r.
We know that the hyperbolic circle C_r' with center i and radius r has Euclidean center $\cosh(r)i$ and radius $\sinh(r)$. If

$$\gamma = \begin{pmatrix} \sqrt{\lambda} & 0 \\ 0 & 1/\sqrt{\lambda} \end{pmatrix} \in SL(2, \mathbb{R}),$$

then $\gamma.i = \lambda i$, and therefore $\gamma.C_r'$ is the hyperbolic circle with center λi and radius r—that is, C_r. But, of course, that just means that C_r is the Euclidean circle with center $\lambda \cosh(r)i$ and radius $\lambda \sinh(r)$.

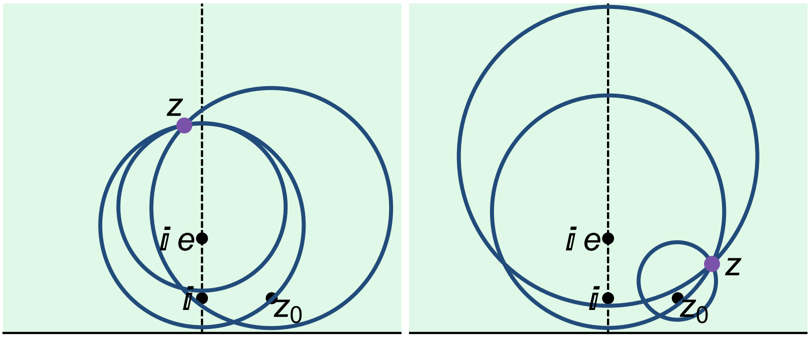


Fig. 4.14 The geometric intuition behind the proof of Theorem 4.8. Given a point $z \in \mathbb{H}^2$ (drawn in purple in both examples), there are circles passing through it with centers at i, ei, and z_0; these circles intersect in a unique point. Thus, if the radii of the circles are known, the point z is specified uniquely. Two examples of this process of triangulation are shown.

4.5 Isometries and Geometric Notions

It is time to make good on our promise of showing that all isometries of $\mathbb{H}^2$ are contained in the group $\mathrm{Isom}(\mathbb{H}^2)$.

Theorem 4.8 *$Isom(\mathbb{H}^2, d_{hyper}) = Isom(\mathbb{H}^2)$.*

Proof We have already shown that $\mathrm{Isom}(\mathbb{H}^2) \subset \mathrm{Isom}(\mathbb{H}^2, d_{\text{hyper}})$. It remains to show that if $\Psi \in \mathrm{Isom}(\mathbb{H}^2, d_{\text{hyper}})$, then $\Psi \in \mathrm{Isom}(\mathbb{H}^2)$. Let $z_1 = \Psi(i)$ and $z_2 = \Psi(ei)$. Then there exists $\varphi \in PSL(2, \mathbb{R})$ such that $\varphi(z_1) = i$ and $\varphi(z_2) = it$ for some $t > 1$. In fact, since

$$d_{\text{hyper}}(i, it) = d_{\text{hyper}}(z_1, z_2) = d_{\text{hyper}}(i, ei),$$

we see that

$$\left|\ln\left(\frac{it}{i}\right)\right| = |\ln(t)| = 1 = \left|\ln\left(\frac{ei}{i}\right)\right|,$$

and since $t > 1$, it must be that $t = e$. Therefore, $\tilde{\Psi} = \varphi \circ \Psi$ has the property that $\tilde{\Psi}(i) = i$ and $\tilde{\Psi}(ei) = ei$. Note that $\tilde{\Psi} \in \mathrm{Isom}(\mathbb{H}^2)$ if and only if Ψ is. Therefore, without loss of generalization, we may assume that Ψ fixes both i and ei. Next, choose any point $z_0 \in \mathbb{H}^2$ to the right of the line $x = 0$. Define $r_1 = d_{\text{hyper}}(z_0, i)$ and $r_2 = d_{\text{hyper}}(z_0, ei)$. Since Ψ is an isometry, it must be that

$$r_1 = d_{\text{hyper}}(z_0, i) = d_{\text{hyper}}\left(\Psi(z_0), \Psi(i)\right) = d_{\text{hyper}}(\Psi(z_0), i);$$

similarly, $d_{\text{hyper}}(\Psi(z_0), ei) = r_2$. Now, draw the hyperbolic circles centered at i and ei with radii r_1 and r_2, respectively. Both z_0 and $\Psi(z_0)$ must lie on both of these circles; however, as hyperbolic circles are Euclidean circles, we see that there are exactly two intersection points. Moreover, by symmetry, these two intersection points are swapped by the map $\phi(z) = -\overline{z}$. Therefore, either $\Psi(z_0) = z_0$ or $(\phi \circ \Psi)(z_0) = z_0$. Since $(\phi \circ \Phi)(i) = i$ and $(\phi \circ \Psi)(ei) = ei$, we see that we may assume without loss of generality that $\Psi(i) = i$, $\Psi(ei) = ei$, and $\Psi(z_0) = z_0$. I claim that actually this forces $\Psi(z) = z$ identically, which of course means that $\Psi \in \mathrm{Isom}(\mathbb{H}^2)$ as claimed.

Why is this? We can just do triangulation, as in Figure 4.14. For any point $z \in \mathbb{H}^2$, we may use the same argument as for z_0 to conclude that either $\Psi(z) = z$ or $\Psi(z) = -\overline{z}$. However, since z_0 is to the right of the $x = 0$ line, either $d_{\text{hyper}}(z, z_0) < d_{\text{hyper}}(-\overline{z}, z_0)$ or $d_{\text{hyper}}(z, z_0) > d_{\text{hyper}}(-\overline{z}, z_0)$—they cannot be equal. Since Ψ is an isometry, it must preserve this inequality. So, for instance, if $d_{\text{hyper}}(z, z_0) < d_{\text{hyper}}(-\overline{z}, z_0)$, then $d_{\text{hyper}}(\Psi(z), z_0) < d_{\text{hyper}}(\Psi(-\overline{z}), z_0)$. But this can only occur if $\Psi(z) = z$. □

We will do a more careful classification of isometries of $\mathbb{H}^2$ in the next chapter; for now, we will gainfully employ the fact that we know what the isometry group is to define various geometrical terms.

Definition 4.8 The *orientation-preserving isometry group of* $\mathbb{H}^2$ is $\mathrm{Isom}^0(\mathbb{H}^2) = PSL(2, \mathbb{R})$. Any isometry of $\mathbb{H}^2$ is either *orientation-preserving* (if it is in $\mathrm{Isom}^0(\mathbb{H}^2)$) or otherwise *orientation-reversing*.

There is no reasonable way to define the orientation for arbitrary metric spaces, but we already know that $\mathrm{Isom}(\mathbb{H}^2)$ splits neatly into these two pieces, so this is entirely sensible.

Definition 4.9 Let $p_1, p_2 : [0, 1] \to \mathbb{H}^2$ be two differentiable paths such that $p_1(1) = p_2(1) = z_0$. The *(hyperbolic) angle of intersection at* z_0 between these two paths is the Euclidean angle of intersection.

Remark 4.6 We know that all of the isometries of $\mathbb{H}^2$ preserve Euclidean angles of intersection, so the definition we have chosen works perfectly well: in particular, we know that hyperbolic isometries preserve hyperbolic angles, exactly as one expects.

Definition 4.10 A *(hyperbolic) polygon* is a connected subset of $\mathbb{H}^2$ bounded by a finite number of (hyperbolic) line segments, referred to as its *sides*. The *vertices* of the polygon are the intersection points of the sides. The *angles* of the polygon are the angles of the intersection at each of the vertices.

While all of these definitions are just like the Euclidean plane, their behavior in the hyperbolic plane is not always analogous. The reader might recall call, for example, that the sum of the angles of a Euclidean polygon with n sides is $(n-2)\pi$. Not so for hyperbolic polygons, as one can see from Figure 4.15!

4.6 The Poincaré Disk Model

One slightly frustrating characteristic of the Poincaré half-plane model is that most of the geometry is "off at infinity"—we can only see a tiny fraction of the whole

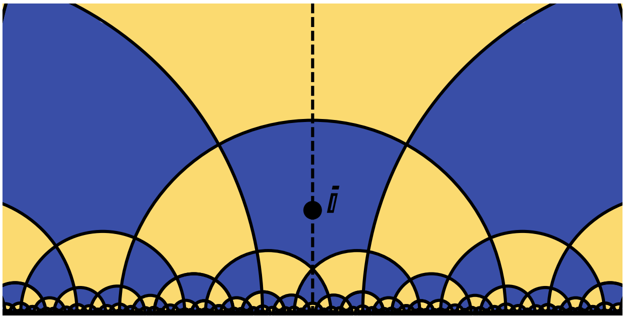


Fig. 4.15 A tiling of the hyperbolic plane by pentagons. Note that the angles of each pentagon have measure $\pi/2$, so their total sum is 2.5π.

hyperbolic plane. What if we could represent it in a different way that didn't have this problem? We know that there exists a linear fractional transformation φ such that $\varphi(\mathbb{H}^2)$ is, say, the unit disk, so there is nothing preventing us from changing the underlying metric space. This would still have the highly desirable property that the isometries of this new space would still be Möbius transformations, with everything that implies. Before we work out the details, there is an alternate characterization of the hyperbolic metric that will be very useful.

Theorem 4.9 *Let $z_1, z_2 \in \mathbb{H}^2$ be distinct points. Let l be the hyperbolic line between them. Let $a, b \in \partial\mathbb{H}^2$ be the intersection points of l with the boundary. Then*

$$d_{hyper}(z_1, z_2) = |\ln([z_1, z_2; a, b])|.$$

Proof We know that elements of $SL(2, \mathbb{R})$ acting as linear fractional transformations preserve both the cross-ratio and the hyperbolic distance; moreover, they can move any two points $z_1, z_2 \in \mathbb{H}^2$ to $i, e^t i$ where $t = d_{\text{hyper}}(z_1, z_2)$. Thus, it shall suffice to prove the theorem for this case, where it is easy to check that a and b are 0 and ∞, although not necessarily in that order—however, it doesn't matter whether we choose $a = 0$ and $b = \infty$ or vice versa. Indeed,

$$\left|\ln\left([i, e^t i; 0, \infty]\right)\right| = \left|\ln\left(e^t\right)\right| = t$$
$$\left|\ln\left([i, e^t i; \infty, 0]\right)\right| = \left|\ln\left(e^{-t}\right)\right| = t,$$

proving the theorem. □

This is very interesting, because the cross-ratio is preserved by all elements in $PSL(2, \mathbb{C})$, and not just $PSL(2, \mathbb{R})$. This motivates the following definition.

Definition 4.11 The *Poincaré disk model* of hyperbolic space consists of the unit disk

$$\mathbb{D}^2 = \{z \in \mathbb{C} | |z| < 1\}$$

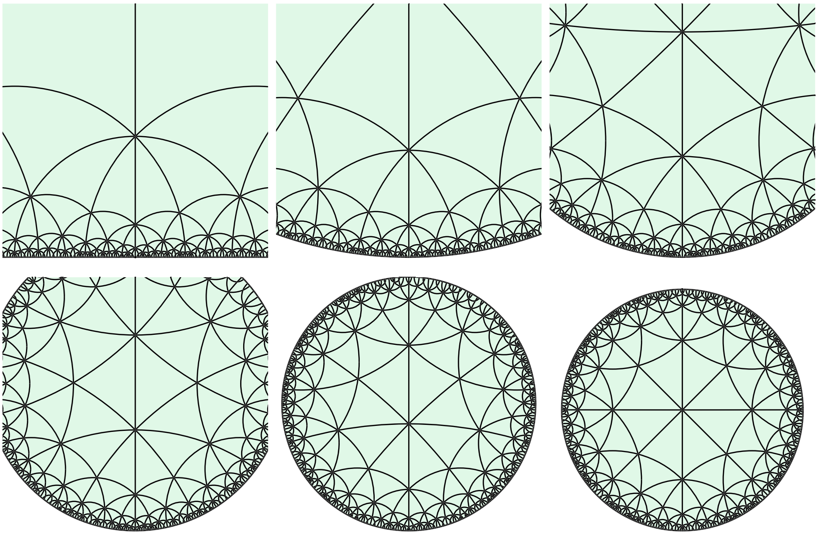


Fig. 4.16 An illustration of the transformation from the Poincaré half-plane to the Poincaré disk.

equipped with a distance function which we shall call, by abuse of notation, $d_{\text{hyper}} : \mathbb{D}^2 \times \mathbb{D}^2 \to [0, \infty)$ and defined as follows. If $z_1 = z_2$, then $d_{\text{hyper}}(z_1, z_2) = 0$. Otherwise, choose a generalized circle C passing through z_1 and z_2 and orthogonal to the boundary—let its intersection points be a and b. Then

$$d_{\text{hyper}}(z_1, z_2) = |\ln([z_1, z_2; a, b])| .$$

The *boundary* $\partial\mathbb{D}^2$ is the unit circle.

Remark 4.7 The existence of a unique generalized circle through z_1 and z_2 orthogonal to the boundary is proved just as it was for $\mathbb{H}^2$.

Theorem 4.10 (**Isometric Isomorphism Between the Poincaré Half-Plane and the Poincaré Disk**) *The Poincaré disk* $(\mathbb{D}^2, d_{hyper})$ *is a metric space. Moreover,*

$$\begin{aligned} \Psi : \mathbb{H}^2 &\to \mathbb{D}^2 \\ z &\mapsto \frac{iz+1}{z+i} \end{aligned}$$

is an isometric isomorphism.

Remark 4.8 See Figure 4.16 for an idea of how to gradually morph one space into the other while preserving the hyperbolic metric.

Proof Consider the linear fractional transformation $\psi(z) = (iz + 1)/(z + i)$. Since $\psi(-1) = -1$, $\psi(0) = -i$, and $\psi(i) = i$, ψ sends $\partial\mathbb{H}^2$, oriented from left to right, to the unit circle, oriented counterclockwise. In particular, Ψ, the restriction of ψ to $\mathbb{H}^2$, is a bijection from $\mathbb{H}^2$ to $\mathbb{D}^2$. For any two points $z_1, z_2 \in \mathbb{H}^2$, since Ψ preserves generalized circles and angles, it will send the generalized circle passing through them and orthogonal to the real line to the generalized circle passing through $\Psi(z_1)$, $\Psi(z_2)$ and orthogonal to the unit circle. It will also send the intersection points a, b with the old boundary to the intersection points $\Psi(a)$, $\Psi(b)$ with the new boundary. Finally, Ψ preserves the cross-ratio and consequently,

$$\begin{aligned} d_{\text{hyper}}(z_1, z_2) &= |\ln([z_1, z_2; a, b])| = |\ln([\Psi(z_1), \Psi(z_2); \Psi(a), \Psi(b)])| \\ &= d_{\text{hyper}}(\Psi(z_1), \Psi(z_2)). \end{aligned}$$

From this, we can conclude that $(\mathbb{D}^2, d_{\text{hyper}})$ is a metric space—after all, if the triangle inequality held in $\mathbb{H}^2$, then it must now also hold in $\mathbb{D}^2$—and Ψ is an isometric isomorphism. □

Corollary 4.2 (Explicit Formula for Distance in the Poincaré Disk) *Let $z_1, z_2 \in \mathbb{D}^2$. Then*

$$d_{hyper}(z_1, z_2) = \cosh^{-1}\left(1 + \frac{2|z_1 - z_2|^2}{\left(1 - |z_1|^2\right)\left(1 - |z_2|^2\right)}\right).$$

Proof We know that $\Phi(z) = (z + i)/(iz + 1)$ is an isometric isomorphism between $\mathbb{D}^2$ and $\mathbb{H}^2$—this is just the inverse of Ψ. Therefore,

$$d_{\text{hyper}}(z_1, z_2) = d_{\text{hyper}}\left(\Phi(z_1), \Phi(z_2)\right) = \cosh^{-1}\left(1 + \frac{|\Phi(z_1) - \Phi(z_2)|^2}{2\Im\left(\Phi(z_1)\right)\Im\left(\Phi(z_2)\right)}\right).$$

From there, it is an easy calculation. (See Exercise 4.2.7.) □

Since we can take the isometric isomorphism between $\mathbb{H}^2$ and $\mathbb{D}^2$ to be a linear fractional transformation, all of the geometric notions that we had defined for $\mathbb{H}^2$ apply equally well in $\mathbb{D}^2$. In particular:

1. Hyperbolic circles in $\mathbb{D}^2$ are loci of points centered around a fixed point at a fixed distance—these are always Euclidean circles, but with shifted centers and radii.
2. Hyperbolic lines in $\mathbb{D}^2$ are loci of points equidistant from two fixed points—these are always generalized circles orthogonal to the boundary $\partial\mathbb{D}^2$.
3. The isometry group $\text{Isom}(\mathbb{D}^2)$ splits into two pieces: the orientation-preserving subgroup $\text{Isom}^0(\mathbb{D}^2)$ and the orientation-reversing transformations.
4. The hyperbolic angle between any two intersecting paths in $\mathbb{D}^2$ is simply the Euclidean angle between them—this is always preserved by the isometries.

Since everything in $\mathbb{H}^2$ corresponds neatly with its counterparts in $\mathbb{D}^2$, we think of both of them as describing the same geometric object—concretely, the hyperbolic plane—with $\mathbb{H}^2$ and $\mathbb{D}^2$ simply being different models of it. Thus, we may view geometric arrangements as in Figure 4.17 as having counterparts in either model. There is an overarching paradigm behind this.

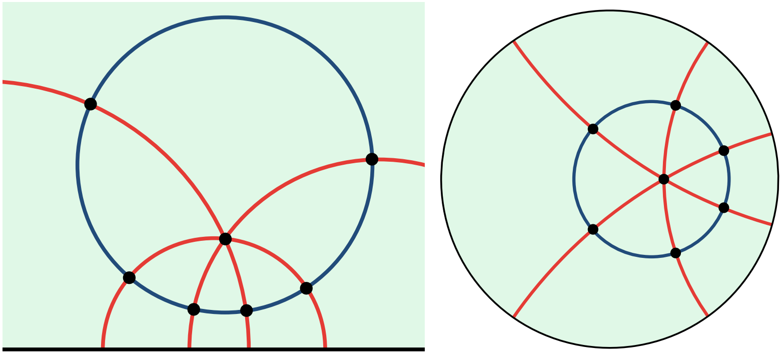


Fig. 4.17 An arrangement of lines, circles, and points in the hyperbolic plane. On the left, this is viewed in $\mathbb{H}^2$; on the right, it is viewed in $\mathbb{D}^2$.

Philosophical Principle

Consider isomorphic geometries as simply different descriptors of the same underlying mathematical object. Use whichever viewpoint (i.e., isomorphic geometry) is most convenient for solving the problem you are currently facing.

It will be easier to work with the Poincaré disk model if we can better understand its isometry group. We know that $\text{Isom}^0(\mathbb{D}^2)$ must consist of linear fractional transformations. Just as we did for the Poincaré upper half-plane, we would like a description of these transformations in terms of some nice matrix group.

Definition 4.12 For any two natural numbers m, n, let $\text{diag}(m, n)$ be the $(m+n) \times (m+n)$ diagonal matrix (i.e., all off-diagonal entries are zero) such that the first m diagonal entries are 1, and the other n are -1. The *indefinite unitary group* $U(m, n)$ consists of all $(m+n) \times (m+n)$ complex matrices M such that $\overline{M}^T \text{diag}(m, n) M = \text{diag}(m, n)$. The *special indefinite unitary group* $SU(m, n)$ consists of all matrices in $U(m, n)$ with determinant 1 or, equivalently, it is the intersection of $U(m, n)$ with $SL(m+n, \mathbb{C})$.

For any m, n, $SU(m, n)$ is a group under matrix multiplication (see Exercise 4.2.8); it is deeply tied to the theory of Hermitian forms, but we won't pursue this point of view. In our case, we are just interested in the group $SU(1, 1)$.

Lemma 4.11

$$SU(1,1) = \left\{ \begin{pmatrix} \alpha & \beta \\ \overline{\beta} & \overline{\alpha} \end{pmatrix} \in SL(2, \mathbb{C}) \right\}.$$

Proof We know that by definition,

$$M = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} \in SL(2, \mathbb{C})$$

is in $SU(1,1)$ if and only if $M\mathrm{diag}(1,-1)\overline{M}^T = \mathrm{diag}(1,-1)$. After this, it is a simple computation that

$$\begin{pmatrix}\overline{\alpha} & \overline{\gamma}\\ \overline{\beta} & \overline{\delta}\end{pmatrix}\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}\begin{pmatrix}\alpha & \beta\\ \gamma & \delta\end{pmatrix} = \begin{pmatrix}\overline{\alpha} & -\overline{\gamma}\\ \overline{\beta} & -\overline{\delta}\end{pmatrix}\begin{pmatrix}\alpha & \beta\\ \gamma & \delta\end{pmatrix}$$

$$= \begin{pmatrix}|\alpha|^2 - |\gamma|^2 & \overline{\alpha}\beta - \overline{\gamma}\delta\\ \overline{\beta}\alpha - \overline{\delta}\gamma & |\beta|^2 - |\delta|^2\end{pmatrix} = \begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}.$$

If $\delta = \overline{\alpha}$ and $\gamma = \overline{\beta}$, it is easy to see that this equality holds, so we simply need to show the converse. Since $\alpha\delta - \beta\gamma = 1$, $|\alpha|^2\delta - \overline{\alpha}\beta\gamma = \overline{\alpha}$. But $|\alpha|^2 = 1 + |\gamma|^2$ and $\overline{\alpha}\beta = \overline{\gamma}\delta$, whence $\overline{\alpha} = (1 + |\gamma|^2)\delta - \overline{\gamma}\delta\gamma = \delta$. We know that $|\delta|^2 \neq 0$ since $|\delta| - |\beta|^2 = 1$. Therefore, since $\overline{\alpha}\beta - \overline{\gamma}\delta = (\beta - \overline{\gamma})\delta = 0$, we can safely conclude that $\beta = \overline{\gamma}$. □

Theorem 4.11 (**Accidental Isomorphism Between** $SL(2,\mathbb{R})$ **and** $SU(1,1)$) *The map*

$$\Psi : SL(2,\mathbb{R}) \to SU(1,1)$$

$$M \mapsto \begin{pmatrix}\frac{1}{\sqrt{2}} & -\frac{i}{\sqrt{2}}\\ -\frac{i}{\sqrt{2}} & \frac{1}{\sqrt{2}}\end{pmatrix} M \begin{pmatrix}\frac{1}{\sqrt{2}} & -\frac{i}{\sqrt{2}}\\ -\frac{i}{\sqrt{2}} & \frac{1}{\sqrt{2}}\end{pmatrix}^{-1}$$

is a group isomorphism.

Proof We first need to show that this map really does send elements of $SL(2,\mathbb{R})$ to elements of $SU(1,1)$. Indeed, if $M \in SL(2,\mathbb{R})$, then

$$\Psi(M)\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}\overline{\Psi(M)}^T$$

$$= \begin{pmatrix}\frac{1}{\sqrt{2}} & -\frac{i}{\sqrt{2}}\\ -\frac{i}{\sqrt{2}} & \frac{1}{\sqrt{2}}\end{pmatrix} M \begin{pmatrix}\frac{1}{\sqrt{2}} & -\frac{i}{\sqrt{2}}\\ -\frac{i}{\sqrt{2}} & \frac{1}{\sqrt{2}}\end{pmatrix}^{-1}\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}\begin{pmatrix}\frac{1}{\sqrt{2}} & \frac{i}{\sqrt{2}}\\ \frac{i}{\sqrt{2}} & \frac{1}{\sqrt{2}}\end{pmatrix}^{-1} M^T \begin{pmatrix}\frac{1}{\sqrt{2}} & \frac{i}{\sqrt{2}}\\ \frac{i}{\sqrt{2}} & \frac{1}{\sqrt{2}}\end{pmatrix}$$

$$= \begin{pmatrix}\frac{1}{\sqrt{2}} & -\frac{i}{\sqrt{2}}\\ -\frac{i}{\sqrt{2}} & \frac{1}{\sqrt{2}}\end{pmatrix} M \begin{pmatrix}0 & -i\\ i & 0\end{pmatrix} M^T \begin{pmatrix}\frac{1}{\sqrt{2}} & \frac{i}{\sqrt{2}}\\ \frac{i}{\sqrt{2}} & \frac{1}{\sqrt{2}}\end{pmatrix}.$$

If we write

$$M = \begin{pmatrix}a & b\\ c & d\end{pmatrix},$$

then one checks that

$$\begin{aligned}
&\begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{i}{\sqrt{2}} \\ -\frac{i}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} M \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} M^T \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{i}{\sqrt{2}} \\ \frac{i}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} \\
&\qquad = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{i}{\sqrt{2}} \\ -\frac{i}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} 0 & -i(ad-bc) \\ i(ad-bc) & 0 \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{i}{\sqrt{2}} \\ \frac{i}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} \\
&\qquad = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{i}{\sqrt{2}} \\ -\frac{i}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{i}{\sqrt{2}} \\ \frac{i}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.
\end{aligned}$$

We claim that

$$\begin{aligned}
SU(1,1) &\to SL(2,\mathbb{R}) \\
M &\mapsto \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{i}{\sqrt{2}} \\ -\frac{i}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}^{-1} M \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{i}{\sqrt{2}} \\ -\frac{i}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}
\end{aligned}$$

is the inverse map. To be precise, this is the inverse map of Ψ if it is well-defined, but it is far from obvious that this map really does send elements of $SU(1,1)$ to elements of $SL(2,\mathbb{R})$. However, this really is so, since

$$\begin{aligned}
&\begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{i}{\sqrt{2}} \\ -\frac{i}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}^{-1} \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{i}{\sqrt{2}} \\ -\frac{i}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} \\
&\qquad = \begin{pmatrix} \frac{\alpha+\overline{\alpha}}{2} - i\frac{\beta-\overline{\beta}}{2} & -i\frac{\alpha-\overline{\alpha}}{2} + \frac{\beta+\overline{\beta}}{2} \\ i\frac{\alpha-\overline{\alpha}}{2} + \frac{\beta+\overline{\beta}}{2} & \frac{\alpha+\overline{\alpha}}{2} + i\frac{\beta-\overline{\beta}}{2} \end{pmatrix} \\
&\qquad = \begin{pmatrix} \Re(\alpha) + \Im(\beta) & \Im(\alpha) + \Re(\beta) \\ -\Im(\alpha) + \Re(\beta) & \Re(\alpha) - \Im(\beta) \end{pmatrix} \in SL(2,\mathbb{R}).
\end{aligned}$$

This shows that Ψ is a well-defined bijection. That it is a group homomorphism is easily confirmed; hence, it is a group isomorphism. □

Theorem 4.12 *The map*

$$\begin{aligned}
SU(1,1) &\to Isom^0(\mathbb{D}^2, d_{hyper}) \\
\begin{pmatrix} \alpha & \beta \\ \overline{\beta} & \overline{\alpha} \end{pmatrix} &\mapsto \left(z \mapsto \frac{\alpha z + \beta}{\overline{\beta} z + \overline{\alpha}} \right)
\end{aligned}$$

is a surjective group homomorphism; two matrices M, N map to the same isometry if and only if $N = \pm M$. Moreover, for any $\phi \in Isom(\mathbb{D}^2)$, either $\phi \in Isom^0(\mathbb{D}^2)$ or $\phi \circ conj \in Isom^0(\mathbb{D}^2)$, where $conj(z) = \overline{z}$.

Proof I leave this one to the reader. (See Exercise 4.2.9.) □

▶ **Example** *Find the orientation-preserving isometries of $\mathbb{H}^2$ that fix the point i.*

We could solve this problem directly by solving $i = (ai + b)/(ci + d)$. There is, however, an easier way. The image of i under the standard isometric isomorphism $\mathbb{H}^2 \to \mathbb{D}^2$ is $(i^2 + 1)/(i + i) = 0$. So, we can instead look for orientation-preserving isometries of $\mathbb{D}^2$ that fix 0. Consider $y = 0$ and $x = 0$; these are both hyperbolic lines passing through 0. They intersect the boundary at ± 1 and $\pm i$, respectively. If Ψ is an isometry of $\mathbb{D}^2$ fixing 0, then the image of $y = 0$ and $x = 0$ must be generalized circles passing through 0 and orthogonal to the unit circle, which is to say that they must be lines through the origin. Furthermore, we know that the angle between them has to be preserved, which means that they are just rotated by some fixed angle θ—in particular, $\Psi(1) = e^{i\theta}$ and $\Psi(i) = e^{i\theta} i$. There is only one linear fractional transformation sending $0 \mapsto 0$, $1 \mapsto e^{i\theta}$, and $i \mapsto e^{i\theta} i$; it must be that $\Psi(z) = e^{i\theta} z$. That is, the orientation-preserving isometries of $\mathbb{D}^2$ that fix 0 are precisely the Euclidean rotations around the origin! Now, we have to translate this observation back to $\mathbb{H}^2$. We have that $\Psi(z) = U.z$, where

$$U = \begin{pmatrix} e^{i\theta/2} & 0 \\ 0 & e^{-i\theta/2} \end{pmatrix} \in SU(1, 1),$$

so we can use the accidental isomorphism between $SU(1, 1)$ and $SL(2, \mathbb{R})$ to get the corresponding transformation

$$\begin{aligned} \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{i}{\sqrt{2}} \\ -\frac{i}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}^{-1} \begin{pmatrix} e^{i\theta/2} & 0 \\ 0 & e^{-i\theta/2} \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{i}{\sqrt{2}} \\ -\frac{i}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} &= \begin{pmatrix} \frac{1}{2} e^{-\frac{i\theta}{2}} + \frac{1}{2} e^{\frac{i\theta}{2}} & \frac{1}{2} i e^{-\frac{i\theta}{2}} - \frac{1}{2} i e^{\frac{i\theta}{2}} \\ \frac{1}{2} i e^{\frac{i\theta}{2}} - \frac{1}{2} i e^{-\frac{i\theta}{2}} & \frac{1}{2} e^{-\frac{i\theta}{2}} + \frac{1}{2} e^{\frac{i\theta}{2}} \end{pmatrix} \\ &= \begin{pmatrix} \cos(\theta/2) & \sin(\theta/2) \\ -\sin(\theta/2) & \cos(\theta/2) \end{pmatrix}. \end{aligned}$$

Therefore, we conclude that the orientation-preserving isometries of $\mathbb{H}^2$ that fix i are those of the form

$$z \mapsto \begin{pmatrix} \cos(\theta/2) & \sin(\theta/2) \\ -\sin(\theta/2) & \cos(\theta/2) \end{pmatrix} .z = \frac{\cos(\theta/2) z + \sin(\theta/2)}{-\sin(\theta/2) z + \cos(\theta/2)},$$

for some $\theta \in \mathbb{R}$.

▶ **Example** *Let C be the hyperbolic circle in $\mathbb{D}^2$ centered at 0 and with radius r. Find the Euclidean center and radius of C as functions of r.*

We already proved that hyperbolic circles are Euclidean circles previously, but now we can give a nicer proof of that fact. If

$$d_{\text{hyper}}(z, 0) = \cosh^{-1}\left(1 + \frac{2|z|^2}{1 - |z|^2}\right) = \cosh^{-1}\left(\frac{1 + |z|^2}{1 - |z|^2}\right) = r,$$

then

$$\frac{1+|z|^2}{1-|z|^2} = \cosh(r),$$

so

$$|z|^2 = \frac{\cosh(r)-1}{\cosh(r)+1},$$

which is a Euclidean circle centered at 0. Since

$$\begin{aligned}\tanh(r/2)^2 &= \left(\frac{e^{r/2}-e^{-r/2}}{e^{r/2}+e^{-r/2}}\right)^2\\ &= \frac{e^r+e^{-r}-2}{e^r+e^{-r}+2}\\ &= \frac{\cosh(r)-1}{\cosh(r)+1},\end{aligned}$$

the radius of this circle is $\tanh(r/2)$, where $\tanh(z) = \sinh(z)/\cosh(z)$, which is the hyperbolic tangent.

4.7 Quaternions

I made a promise earlier in this chapter that we would find a nice metric space such that Möb(2) will be its group of isometries. We shall do this in the next section, when we shall introduce three-dimensional hyperbolic space. Before that, we require an interlude to talk about something which might seem completely unconnected from what has come before in this chapter. The rough reason for this is the following: complex numbers are very useful for describing two-dimensional hyperbolic space, as we have seen. To try to expand two-dimensional hyperbolic space to three-dimensional hyperbolic space, it thus makes sense to try to expand the complex numbers and look at some sort of larger algebraic structure.

The question of how to do this became of great interest to mathematicians in the 19th century and was ultimately resolved by the Irish mathematician William Rowan Hamilton defining the quaternions in 1843[1] . Figure 4.18 shows his portrait, taken when he was in his mid-50s. Hamilton was seeking some extension of the complex numbers with a well-defined notion of addition and multiplication which would have nice properties. Originally, his efforts concentrated on finding some such

[1] It's worth noting that Carl Friedrich Gauss independently defined the quaternions in 1819, but Gauss had a bad habit of only publishing results that he felt were very important and well-presented. As a result, this as well as many other of his findings were only published decades after his death in 1855, with the credit going to mathematicians who did publish.

Fig. 4.18 William Rowan Hamilton, circa 1860.

operations on triples of real numbers (x, y, z), analogous to how complex numbers can be represented by pairs of real numbers (x, y). What he did not know was that—assuming some very reasonable restrictions on what this operation had to be—this was impossible. This was the Frobenius theorem for real division algebras, which wouldn't be proved until 1877, a decade after Hamilton's death.

The profound realization that pushed Hamilton to a more fruitful avenue came in 1843, when he was walking with his wife along the Royal Canal in Dublin. Hamilton was so energized by this revelation that he immediately took out his knife and carved the equation $i^2 = j^2 = k^2 = ijk = -1$ on a stone of the Broom Bridge. Sadly, the original inscription has not survived, but there is to this day a plaque there commemorating this moment in mathematical history.

What did Hamilton's equation mean? His idea was that rather than looking at triples, one should instead look at *quadruples* of real numbers $(x, y, z, t) \in \mathbb{R}^4$. Furthermore, for convenience, write these as $x + yi + zj + tk$. The rules for addition are exactly what you would expect:

$$\begin{aligned}(x_1 + y_1 i + z_1 j + t_1 k) &+ (x_2 + y_2 i + z_2 j + t_2 k)\\ &= (x_1 + x_2) + (y_1 + y_2)i + (z_1 + z_2)j + (t_1 + t_2)k.\end{aligned}$$

The rules for multiplication are a little more complicated but are mostly captured in this equation $i^2 = j^2 = k^2 = ijk = -1$, which tells you what to do with the symbols i, j, and k. To get the full rule-set, so to speak, you need the following additional information.

1. For any real number r and quaternion $x + yi + zj + tk$, $r(x + yi + zj + tk) = (x + yi + zj + tk)r = (rx) + (ry)i + (rz)j + (rt)k$. (This is expected—this is how scaling vectors in $\mathbb{R}^4$ normally works.)
2. For all quaternions q_1, q_2, q_3, $q_1(q_2 q_3) = (q_1 q_2)q_3$. (This is the associative property that we previously saw in the definition of groups.)

3. For all quaternions q_1, q_2, q_3, $q_1(q_2 + q_3) = q_1q_2 + q_1q_3$ and $(q_1 + q_2)q_3 = q_1q_3 + q_2q_3$. (This is the *distributive property* of multiplication, which we know holds true for real and complex numbers.)

Once you know all of this, the general multiplication of quaternions can be deduced. Let's try a few simple cases first. Suppose we want to know what ij is. We know that $ijk = -1$, so $ijk^2 = -k$; since $k^2 = -1$, $ij = k$. What about ji? One might expect this to be k as well, but this is not so:

$$ji = -ji(ijk) = -j(i^2)jk = j^2k = -k.$$

This was the second part of Hamilton's insight; to get the desired extension of the complex numbers, one needs to turn to a multiplication that is not commutative, which is to say that order in which things are multiplied matters. This shouldn't faze us too much—we have mostly been dealing with non-commutative multiplication thus far already. But in the mid-1800s, this was rather unexpected. In any case, using similar reasoning, one can work out that

$$\begin{array}{lll} ij = k & jk = i & ki = j \\ ji = -k & kj = -i & ki = -j. \end{array}$$

(See Exercise 4.1.3.) This gives enough information to do quaternion multiplication in general. For example,

$$\begin{aligned}(1 - i + j)(2i + k) &= 1(2i + k) - i(2i + k) + j(2i + k)\\ &= 2i + k - 2i^2 - ik + 2ji + jk\\ &= 2i + k + 2 + j - 2k + i\\ &= 2 + 3i + j - k,\end{aligned}$$

or even

$$\begin{aligned}(x_1 + y_1i + z_1j + t_1k)&(x_2 + y_2i + z_2j + t_2k)\\ &= x_1(x_2 + y_2i + z_2j + t_2k)\\ &+ y_1i(x_2 + y_2i + z_2j + t_2k)\\ &+ z_1j(x_2 + y_2i + z_2j + t_2k)\\ &+ t_1k(x_2 + y_2i + z_2j + t_2k)\\ &= (x_1x_2 - y_1y_2 - z_1z_2 - t_1t_2)\\ &+ (x_1y_2 + y_1x_2 + z_1t_2 - t_1z_2)i\\ &+ (x_1z_2 - y_1t_2 + z_1x_2 + t_1y_2)j\\ &+ (x_1t_2 + y_1z_2 - z_1y_2 + t_1x_2)k.\end{aligned}$$

This allows us to give an unambiguous definition of the quaternions.

Definition 4.13 The *quaternions* $\mathcal{H}$ are the set of all formal symbols $x + yi + zj + tk$ with $x, y, z, t \in \mathbb{R}$, along with two binary operations dubbed addition $+$ and multiplication $\cdot$, defined by

$$\begin{aligned}(x_1 + y_1 i + z_1 j + t_1 k) &+ (x_2 + y_2 i + z_2 j + t_2 k)\\ &= (x_1 + x_2) + (y_1 + y_2)i + (z_1 + z_2)j + (t_1 + t_2)k\end{aligned}$$

and

$$\begin{aligned}(x_1 + y_1 i + z_1 j + t_1 k) \cdot (x_2 + y_2 i + z_2 j + t_2 k)&\\ = (x_1x_2 - y_1y_2 - z_1z_2 - t_1t_2)&\\ + (x_1y_2 + y_1x_2 + z_1t_2 - t_1z_2)i&\\ + (x_1z_2 - y_1t_2 + z_1x_2 + t_1y_2)j&\\ + (x_1t_2 + y_1z_2 - z_1y_2 + t_1x_2)k.&\end{aligned}$$

Remark 4.9 It would make sense to denote the quaternions by Q, but sadly this conflicts with the convention that $\mathbb{Q}$ denotes the rationals. Instead, they are usually denoted by an ‘H’ in honor of Hamilton. Some authors use $\mathbb{H}$ as the symbol, but as we use this for hyperbolic space, I have compromised to use $\mathcal{H}$ instead.

Remark 4.10 This is by no means the only possible way to express the quaternions. An alternative construction is explored in Exercise 4.2.13.

Quaternions were initially quite popular after Hamilton introduced them, and there was a push to incorporate them in physics; indeed, Maxwell wrote down his equations for electromagnetism in terms of quaternions. After a time, this approach was abandoned in favor of matrices and vectors, which were championed by Heaviside, Gibbs, and others. Quaternions fell into partial ignominy, although they continued to have interest in pure mathematics. Their resurgence as objects of inquiry for applied mathematics came through computer science: quaternions, like matrices, can be used to describe three-dimensional rotations (see Figure 4.19 and Exercise 4.4.5) but they do not suffer from a phenomenon known as gimbal lock.

Much like the complex numbers have complex conjugation, the quaternions have quaternion conjugation, defined as

$$\overline{x + yi + zj + tk} = x - yi - zj - tk.$$

In other words, you leave the real component of the quaternion alone and change the signs of the other, “imaginary”, components. Quaternion conjugation has many of the same properties as complex conjugation. It can be used to define the norm and trace of a quaternion, for example.

Definition 4.14 For any quaternion q, its *norm* is $|q| = \sqrt{q\overline{q}}$, and its *trace* is $\mathrm{tr}(q) = q + \overline{q}$.

It might not be obvious how to interpret the square root here. As it happens, quaternions can possess infinitely many different quaternion square roots—for instance, it isn’t difficult to see that $(\cos(\theta)i + \sin(\theta)j)^2 = -1$ for all $\theta \in \mathbb{R}$. However, everything is completely aboveboard, because

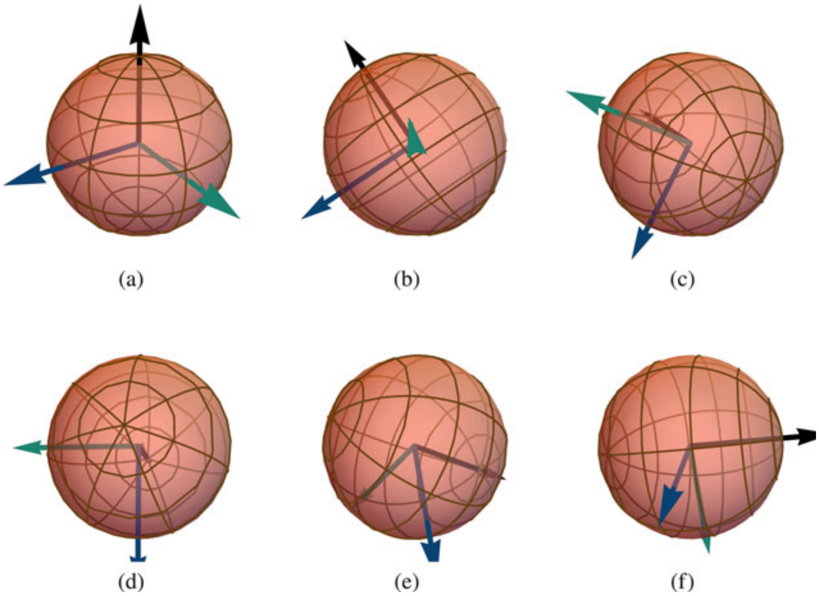


Fig. 4.19 Quaternions are convenient for describing smooth rotations. The last image (f) is a rotation of the first image (a) using the transformation $\mathbf{v} \mapsto q\mathbf{v}q^{-1}$, where $q = (-5 + 4i + 2j - 2k)/7$ and we identify three-dimensional vectors $\mathbf{v} = (x, y, z)$ with quaternions $xi + yj + zk$. The intermediate images (b)-(e) are also rotations using quaternions, using an interpolation between 1 and q.

$$\begin{aligned}(x + yi + zj + tk)\overline{(x + yi + zj + tk)} &= (x + yi + zj + tk)(x - yi - zj - tk) \\ &= x^2 + y^2 + z^2 + t^2 > 0.\end{aligned}$$

Thus, $q\overline{q}$ is a non-negative real number and so it has a unique non-negative square root. This square root, the norm, is nothing more than the Euclidean norm of the vector (x, y, z, t). Similarly, the trace has a straightforward interpretation as well, since

$$x+yi+zj+tk+\overline{x + yi + zj + tk} = x+yi+zj + tk - x - yi - zj - tk = 2x,$$

so it is twice the real part of the quaternion. The term “trace” might seem odd here because we previously used this term in reference to matrices. This is not a coincidence. (See Exercise 4.2.13.) There are a number of other important properties enjoyed by quaternion conjugation, summarized below.

Theorem 4.13 **(Basic Properties of Quaternion Conjugation)** *Let q, q' be quaternions. Then*

1. $q = \overline{q}$ *if and only if* $q \in \mathbb{R}$,
2. $|q| = |\overline{q}|$,
3. $q = 0$ *if and only if* $|q| = 0$,

4. $\overline{qq'} = \overline{q'}\,\overline{q}$*,and*
5. $|qq'| = |q|\,|q'|$.

Proof I leave the first three as exercises for the reader (specifically, Exercise 4.2.17). We will prove the fourth property by pure computation. Let

$$q = x_1 + y_1 i + z_1 j + t_1 k \qquad\qquad q' = x_2 + y_2 i + z_2 j + t_2 k,$$

so

$$\begin{aligned} qq' = &(x_1x_2 - y_1y_2 - z_1z_2 - t_1t_2)\\ &+ (x_1y_2 + y_1x_2 + z_1t_2 - t_1z_2)i\\ &+ (x_1z_2 - y_1t_2 + z_1x_2 + t_1y_2)j\\ &+ (x_1t_2 + y_1z_2 - z_1y_2 + t_1x_2)k. \end{aligned}$$

Replacing q by q' switches the index 1 with the index 2 in the above product. Replacing q with $\overline{q}$ and q' with $\overline{q'}$ switches the sign of each term with exactly one factor of x_1 or x_2. Taken together, these changes mean that

$$\begin{aligned} \overline{q'}\,\overline{q} &= (x_1x_2 - y_1y_2 - z_1z_2 - t_1t_2)\\ &\quad + (-x_2y_1 - y_2x_1 + z_2t_1 - t_2z_1)i\\ &\quad + (-x_2z_1 - y_2t_1 - z_2x_1 + t_2y_1)j\\ &\quad + (-x_2t_1 + y_2z_1 - z_2y_1 - t_2x_1)k\\ &= (x_1x_2 - y_1y_2 - z_1z_2 - t_1t_2)\\ &\quad - (x_1y_2 + y_1x_2 + z_1t_2 - t_1z_2)i\\ &\quad - (x_1z_2 - y_1t_2 + z_1x_2 + t_1y_2)j\\ &\quad - (x_1t_2 + y_1z_2 - z_1y_2 + t_1x_2)k, \end{aligned}$$

which is nothing more than $\overline{qq'}$. Finally, knowing that $\overline{qq'} = \overline{q'}\,\overline{q}$, it immediately follows that

$$\begin{aligned} |qq'| &= \left(qq'\overline{qq'}\right)^{\frac{1}{2}} = \left(qq'\overline{q'}\,\overline{q}\right)^{\frac{1}{2}}\\ &= \left(q|q'|^2\overline{q}\right)^{\frac{1}{2}} = |q'|\,(q\overline{q})^{\frac{1}{2}}\\ &= |q'||q| = |q||q'|, \end{aligned}$$

since at the end we are simply dealing with multiplication of real numbers, which is commutative. □

One quick consequence of this is that for any real number r, $\overline{rq} = \overline{q}\,\overline{r} = \overline{q}r = r\overline{q}$. A second is that if $|q| \neq 0$, then we may consider the quaternion $\overline{q}/|q|^2$, and since

$$q\overline{q}/|q|^2 = |q|^2/|q|^2 = 1 = \overline{q}/|q|^2 q,$$

we see that this is q^{-1}—that is, any non-zero quaternion is invertible. This means that one can cancel non-zero quaternions like one can with complex numbers—for example, if $qq' = qq''$ and $q \neq 0$, then $q' = q''$. However, the reader should be left with two very important warnings.

1. Cancellation can either be done on the left or on the right, but one cannot mix one and the other. For instance, it does not follow from $ij = -ji$ that $j = -j$.
2. For complex numbers, $(zw)^{-1} = z^{-1}w^{-1}$. This is not true for quaternions—one must instead use $(qq')^{-1} = q'^{-1}q^{-1}$. Indeed, $(ij)^{-1} = -k = (-j)(-i) = j^{-1}i^{-1}$.

▶ **Example** *Any vector* $\mathbf{v} \in \mathbb{R}^3$ *can be identified with a traceless quaternion as follows: if* $\mathbf{v} = (x, y, z)$*, then the corresponding quaternion is* $xi + yj + zk$*. Therefore, we can write any quaternion in the form* $t + \mathbf{v}$ *for some* $t \in \mathbb{R}$ *and* $\mathbf{v} \in \mathbb{R}^3$*. Given* $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{R}^3$*, express* $\mathbf{v}_1\mathbf{v}_2$*— that is, their product as quaternions—in terms of standard vector operations.*

Write $\mathbf{v}_1 = (x_1, y_1, z_1)$ and $\mathbf{v}_2 = (x_2, y_2, z_2)$. Then

$$\begin{aligned}\mathbf{v}_1\mathbf{v}_2 &= (x_1 i + y_1 j + z_1 k) \cdot (x_2 i + y_2 j + z_2 k)\\ &= -(x_1x_2 + y_1y_2 + z_1z_2) + (y_1z_2 - y_2z_1)\, i\\ &+ (x_2z_1 - x_1z_2)\, j + (x_1y_2 - x_2y_1)\, k.\end{aligned}$$

The real component is easy to recognize—this is $\mathbf{v}_1 \cdot \mathbf{v}_2$, the dot product. The rest is a little more obscure: it is the cross product $\mathbf{v}_1 \times \mathbf{v}_2$. Therefore, $\mathbf{v}_1\mathbf{v}_2 = -\mathbf{v}_1 \cdot \mathbf{v}_2 + \mathbf{v}_1 \times \mathbf{v}_2$.

4.8 Hyperbolic 3-Space

Hyperbolic 3-space is to the hyperbolic plane what 3-dimensional Euclidean space is to the Euclidean plane. There are many different models of hyperbolic space. For example, one can picture it by taking an open ball and defining a distance function that measures points farther from the origin as being farther apart—this is the Poincaré ball model, which is an analog to the Poincaré disk model. Another option is to build an analog of the Poincaré half-plane model, as follows. Take the set of all points $(x, y, z) \in \mathbb{R}^3$ with $z > 0$—this is the upper half of three-dimensional Euclidean space. Then, define a distance function on this set that measures points with smaller z-component as being farther apart—this is the Poincaré half-space model. Other models also exist, but for our purposes, we will stick with the Poincaré half-space model as it will be the easiest to define and picture. So, with this introduction, we define

$$\mathbb{H}^3 = \left\{ x + yi + zj \,\middle|\, z > 0 \right\},$$

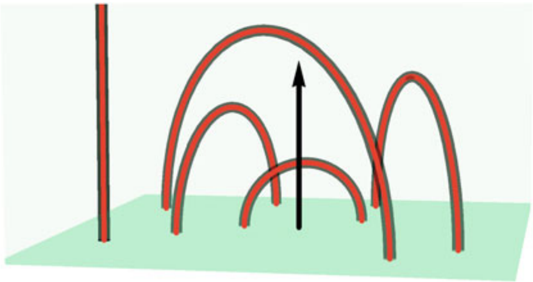


Fig. 4.20 The Poincaré half-space, together with a collection of hyperbolic lines (to be defined later) rendered in red.

where we choose to identify $\mathbb{R}^3$ with a subset of the quaternions in much the same way that we identified $\mathbb{R}^2$ with $\mathbb{C}$. The important thing to figure out here is what the right distance function for this space should be. Recall that for points $z, w \in \mathbb{H}^2$, we had

$$d_{\text{hyper}}(z_1, z_2) = \cosh^{-1}\left(1 + \frac{|z_1 - z_2|^2}{2\Im(z_1)\Im(z_2)}\right).$$

It, therefore, seems reasonable to guess that the following definition is likely sensible.

Definition 4.15 The *Poincaré half-space model* of three-dimensional hyperbolic space consists of the set $\mathbb{H}^3$ together with the distance function

$$\begin{aligned} d_{\text{hyper}} : \mathbb{H}^3 \times \mathbb{H}^3 &\to [0, \infty) \\ (q_1, q_2) &\mapsto \cosh^{-1}\left(1 + \frac{|q_1 - q_2|^2}{2\pi_j(q_1)\pi_j(q_2)}\right), \end{aligned}$$

where $\pi_j(q)$ is the j-th component of q—i.e., if $q = x + yi + zj + tk$, then $\pi_j(q) = z$. The *boundary* $\partial\mathbb{H}^3$ of $\mathbb{H}^3$ is $\mathbb{C}P^1$.

Figure 4.20 illustrates what this space looks like. The definition we have chosen is exactly right—in principle, we could justify this by some argument similar to that from Section 4.3. It is perhaps not immediately clear that this is a metric space, but we shall prove that later. One of the advantages of defining $\mathbb{H}^3$ this way is that there is a simple way of describing how matrices in $SL(2, \mathbb{C})$ transform it.

Definition 4.16 For any

$$M = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL(2, \mathbb{C}),$$

and any quaternion $q \in \mathcal{H}$, define

$$M.q = \begin{cases} (aq+b)(cq+d)^{-1} & \text{if } q \neq -d/c \\ \infty & \text{otherwise.} \end{cases}$$

Notice that since $-d/c \in \mathbb{C}$, which does not belong to $\mathbb{H}^3$, we are always guaranteed that if $q \in \mathbb{H}^3$, then $M.q$ is a well-defined quaternion. What is not obvious is that this quaternion is again in $\mathbb{H}^3$, but this is nevertheless true.

Lemma 4.12 *For any $q \in \mathbb{H}^3$ and any $M \in SL(2,\mathbb{C})$, $M.q \in \mathbb{H}^3$. Furthermore, if $\pi_j(q)$ is the j-th component of q and*

$$M = \begin{pmatrix} a & b \\ c & d \end{pmatrix},$$

then $\pi_j(M.q) = |cq+d|^{-2}\pi_j(q)$.

Proof We can write $q = \alpha + rj$ for some complex number α and some positive real number r. Note that for any complex number β, $\beta j = j\overline{\beta}$. After that, we first compute

$$|cq+d|^2 = |c\alpha + d + crj|^2 = |c\alpha + d|^2 + r^2|c|^2,$$

where we have used the basic fact that $|(x, y, z)|^2 = |(x, y)|^2 + z^2$. Then,

$$\begin{aligned} |cq+d|^2(M.q) &= (aq+b)\overline{(cq+d)} \\ &= (a\alpha + b + raj)(\overline{c\alpha} + \overline{d} - rj\overline{c}) \\ &= \left(a\overline{c}|\alpha|^2 + b\overline{c\alpha} + b\overline{d} + r^2 a\overline{c}\right) + raj\overline{c\alpha} + raj\overline{d} - ra\alpha j\overline{c} - rbj\overline{c} \\ &= \left(a\overline{c}|q|^2 + b\overline{c\alpha} + b\overline{d}\right) + rac\alpha j + radj - ra\alpha cj - rbcj \\ &= \left(a\overline{c}|q|^2 + b\overline{c\alpha} + b\overline{d}\right) + r(ad - bc)j \\ &= \left(a\overline{c}|q|^2 + b\overline{c\alpha} + b\overline{d}\right) + rj. \end{aligned}$$

Note that this has no k-component and that the j-component is unchanged. □

One question we might have is whether $(MN).q = M.(N.q)$—we saw that this is how it worked when we defined how $SL(2,\mathbb{C})$ acted on $\mathbb{C}P^1$. Indeed, this still holds for $\mathbb{H}^3$.

Lemma 4.13 *For any $\gamma_1, \gamma_2 \in SL(2,\mathbb{C})$ and $q \in \mathbb{H}^3$, $(\gamma_1\gamma_2).q = \gamma_1.(\gamma_2.q)$.*

Proof Let

$$\gamma_1 = \begin{pmatrix} a_1 & b_1 \\ c_1 & d_1 \end{pmatrix} \quad \gamma_2 = \begin{pmatrix} a_2 & b_2 \\ c_2 & d_2 \end{pmatrix}.$$

Then

$$(\gamma_1\gamma_2).q = ((a_1a_2{+}b_1c_2)q{+}(a_1b_2 + b_1d_2))\,((a_2c_1{+}c_2d_1)q + (b_2c_1 + d_1d_2))^{-1}$$

and

$$\begin{aligned}\gamma_1.(\gamma_2.q) &= \gamma_1.\left((a_2q + b_2)(c_2q + d_2)^{-1}\right)\\ &= \left(a_1\left((a_2q + b_2)(c_2q + d_2)^{-1}\right) + b_1\right)\\ &\cdot\left(c_1\left((a_2q + b_2)(c_2q + d_2)^{-1}\right) + d_1\right)^{-1}\\ &= (a_1(a_2q + b_2) + b_1(c_2q + d_2))\,(c_2q + d_2)^{-1}\\ &\left((c_1(a_2q + b_2) + d_1(c_2q + d_2))\,(c_2q + d_2)^{-1}\right)^{-1}\\ &= (a_1(a_2q + b_2) + b_1(c_2q + d_2))\,(c_1(a_2q + b_2) + d_1(c_2q + d_2))^{-1}\\ &= ((a_1a_2 + b_1c_2)q + (a_1b_2 + b_1d_2))\\ &\cdot((a_2c_1 + c_2d_1)q + (b_2c_1 + d_1d_2))^{-1},\end{aligned}$$

which concludes the lemma. □

The most important thing for us about the transformations $q \mapsto M.q$ is that they don't just give transformations of $\mathbb{H}^3$; they also preserve d_{hyper}.

Lemma 4.14 *For any $q_1, q_2 \in \mathbb{H}^3$ and $M \in SL(2,\mathbb{C})$, $d_{hyper}(M.q_1, M.q_2) = d_{hyper}(q_1, q_2)$.*

Proof Recall that any element in $SL(2,\mathbb{C})$ can be written as a product of matrices of the form

$$N_b = \begin{pmatrix}1 & b\\ 0 & 1\end{pmatrix},\quad K = \begin{pmatrix}0 & 1\\ -1 & 0\end{pmatrix},\quad A_u = \begin{pmatrix}u & 0\\ 0 & u^{-1}\end{pmatrix}$$

for some $b \in \mathbb{C}$, $u \in \mathbb{C}^\times$. If we can prove that d_{hyper} is preserved by these basic matrices, then we automatically know that it will be preserved by all matrices in $SL(2,\mathbb{C})$. In fact, we can be even more conservative in our choices of matrices, because

$$\begin{aligned}A_u = \begin{pmatrix}u & 0\\ 0 & u^{-1}\end{pmatrix} &= \begin{pmatrix}0 & 1\\ -1 & 0\end{pmatrix}\begin{pmatrix}1 & u^{-1}\\ 0 & 1\end{pmatrix}\begin{pmatrix}0 & 1\\ -1 & 0\end{pmatrix}\begin{pmatrix}1 & u\\ 0 & 1\end{pmatrix}\begin{pmatrix}0 & 1\\ -1 & 0\end{pmatrix}\begin{pmatrix}1 & u^{-1}\\ 0 & 1\end{pmatrix}\\ &= KN_{u^{-1}}KN_uKN_{u^{-1}}.\end{aligned}$$

This leaves us with a straightforward computation.

$$\frac{|N_b.q_1 - N_b.q_2|^2}{2\pi_j(N_b.q_1)\pi_j(N_b.q_2)} = \frac{|(q_1 + b) - (q_2 + b)|^2}{2\pi_j(q_1)\pi_j(q_2)} = \frac{|q_1 - q_2|^2}{2\pi_j(q_1)\pi_j(q_2)}$$

and

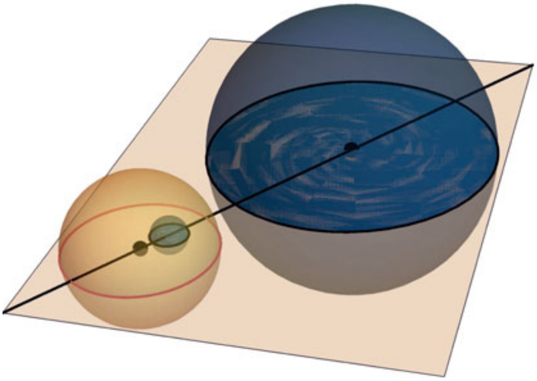


Fig. 4.21 An inversion of a sphere (drawn in blue) through the unit sphere (drawn in yellow). A cross-section of this inversion reveals a circle inversion.

$$\begin{aligned}\frac{|K.q_1 - K.q_2|^2}{2\pi_j(K.q_1)\pi_j(K.q_2)} &= \frac{\left|-q_1^{-1} + q_2^{-1}\right|^2}{2|q_1|^{-2}|q_2|^{-2}\pi_j(q_1)\pi_j(q_2)} \\ &= \frac{|q_1|^2\left|-q_1^{-1} + q_2^{-1}\right|^2|q_2|^2}{2\pi_j(q_1)\pi_j(q_2)} = \frac{|q_1 - q_2|^2}{2\pi_j(q_1)\pi_j(q_2)},\end{aligned}$$

where we make use of the previous lemma. □

For simplicity of exposition, I have chosen to employ the trick shown to allow us to eliminate diagonal matrices A_u as basic generators that we have to consider. However, their action is not so complicated: they are simply compositions of rotations and dilations, just as they were on the plane. We will investigate this more closely in Chapter 5.

In any case, this action by $SL(2, \mathbb{C})$ preserves a number of other important geometric notions. To start with, define a *generalized sphere* to be either a sphere in $\mathbb{R}^3$ or a plane union the point at infinity.

Theorem 4.14 *Let S be a generalized sphere and $\gamma \in SL(2, \mathbb{C})$. Then $\gamma.S$ is a generalized sphere.*

Proof It suffices to prove this for the basic kinds of matrices

$$N_b = \begin{pmatrix} 1 & b \\ 0 & 1 \end{pmatrix}, \quad K = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix},$$

where we are again making use of the trick used in Lemma 4.14 to be able to ignore diagonal matrices. Since $N_b.q = q + b$ is a translation, it preserves generalized

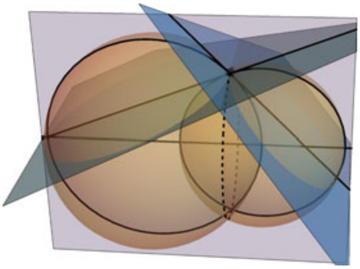


Fig. 4.22 The angle of intersection is shown between two intersecting spheres via their tangent planes. A cross section reveals that this is the angle of intersection between two circles.

spheres. It is less clear what is happening with $K.q = -q^{-1}$. However, note that $q \mapsto \overline{q}$ and $q \mapsto -q$ are reflections and so certainly preserve generalized spheres. On the other hand, the map $q \mapsto \overline{q}^{-1}$ is nothing other than an inversion through the unit sphere—that is, it takes a point q distance r away from the origin to $|q|^{-2}q$, a point on the same ray, but now distance $1/r$ away from the origin. This is illustrated in Figure 4.21. Sphere inversions preserve generalized spheres; we could prove this in a manner similar to the way we did it in Chapter 2, but it is substantially easier to argue by cross-sections and symmetry. To wit, let S' be the sphere of radius r we are trying to invert through the unit sphere centered at the origin. Take any plane that passes through the centers of these two spheres. On this plane, we get a circle of radius 1 and a circle of radius r, and it is easy to see that $q \mapsto \overline{q}^{-1}$ restricted to this plane inverts the circle of radius r through the circle of radius 1. We already know that the result of this will be a generalized circle. Since every cross-section of the image of S is such a generalized circle, and since the bends of all of these circles are the same, we conclude that the image of S must be a generalized sphere. □

It is also true that the transformations given by elements of $SL(2, \mathbb{C})$ preserve angles. We will describe this in a manner analogous to the methodology from Chapter 2: specifically, the angle between two generalized spheres is the angle between their two tangent planes at any of the points of intersection, as in Figure 4.22. That this angle is the same regardless of the choice of intersection point follows from rotational symmetry.

Theorem 4.15 *If S_1, S_2 are generalized spheres that intersect at an angle θ and $\gamma \in SL(2, \mathbb{C})$, then $\gamma.S_1, \gamma.S_2$ also intersect at an angle θ.*

Proof Once again, we only need to prove this for the matrices

$$N_b = \begin{pmatrix} 1 & b \\ 0 & 1 \end{pmatrix}, \quad K = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

Translations preserve angles, as do all reflections, so we only need to consider whether the sphere inversion $q \mapsto \overline{q}^{-1}$ preserves angles or not. Indeed it does, by a cross-sectional argument: take the plane P through the origin and the centers of

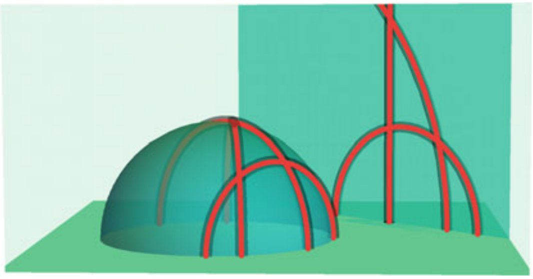


Fig. 4.23 Copies of the hyperbolic plane inside of three-dimensional hyperbolic space.

S_1 and S_2. The intersection of P with S_1 and S_2 gives two circles that intersect at an angle θ. A sphere inversion through the origin preserves this plane P and behaves like a regular circle inversion when restricted to P—consequently, the image under $q \mapsto \overline{q}^{-1}$ will give two generalized spheres intersecting at an angle θ. □

We could also prove that $SL(2, \mathbb{C})$ preserves orientation, but I leave this as a task for the reader to ponder on their own, as we will not need it here. We are actually ready to show that $(\mathbb{H}^3, d_{\text{hyper}})$ is a metric space.

Theorem 4.16 *Let G be the intersection of any generalized sphere orthogonal to $\mathbb{C}P^1$ and $\mathbb{H}^3$. Then $(G, d_{hyper}|_G)$ is isometrically isomorphic to $(\mathbb{H}^2, d_{hyper})$. As a consequence, $(\mathbb{H}^3, d_{hyper})$ is a metric space.*

Remark 4.11 This result gives some insight into why we call $\mathbb{H}^3$ three-dimensional hyperbolic space—as we will see shortly, the generalized spheres G play much the same role as planes in Euclidean space. We know that if we look at any plane in Euclidean three-dimensional space, we get Euclidean two-dimensional space. Just so, if we look at any of these generalized spheres G, we get a copy of hyperbolic two-dimensional space on it, as in Figure 4.23.

Proof The boundary of G is a generalized circle in $\mathbb{C}P^1$. We know that any such circle can be moved to the line $x = 0$ by a transformation $z \mapsto \gamma.z$ for some $\gamma \in SL(2, \mathbb{C})$. The extension of this to $\mathbb{H}^3$ via $q \mapsto \gamma.q$ will move G to the upper half of the xz-plane, which we shall refer to as Y—this follows as we know that such transformations preserve both generalized spheres and angles. Furthermore, we know that this transformation preserves d_{hyper}. Therefore, if we know that $(Y, d_{\text{hyper}}|_Y)$ is isometrically isomorphic to $(\mathbb{H}^2, d_{\text{hyper}})$, then it will automatically follow that $(G, d_{\text{hyper}}|_G)$ is as well, because $(Y, d_{\text{hyper}}|_Y)$ and $(G, d_{\text{hyper}}|_G)$ are related by the isometric isomorphism $q \mapsto \gamma.q$. Why is the choice of Y particularly nice? Well, if we write down the hyperbolic distance restricted in this plane,

$$d_{\text{hyper}}|_Y : Y \times Y \to [0, \infty)$$

$$(x_1 + z_1 j, x_2 + z_2 j) \mapsto \cosh^{-1}\left(1 + \frac{|(x_1 + z_1 j) - (x_2 + z_2 j)|^2}{2(x_1 + z_1 j)(x_2 + z_2 j)}\right),$$

it isn't difficult to notice that this is *exactly* the hyperbolic distance in $\mathbb{H}^2$—all that we have done is renamed i to j! Therefore, the map $x + yi \mapsto x + yj$ is an isometric isomorphism between $(\mathbb{H}^2, d_{\text{hyper}})$ and $(Y, d_{\text{hyper}}|_Y)$.

Why does it follow immediately that $(\mathbb{H}^3, d_{\text{hyper}})$ is a metric space? Well, for any three points q_1, q_2, q_3 in $\mathbb{H}^3$, they lie on some generalized sphere orthogonal to the boundary. Call the intersection of that sphere with $\mathbb{H}^3$ to be G—we know that $(G, d_{\text{hyper}}|_G)$ is a metric space and therefore in particular

1. $d_{\text{hyper}}(q_1, q_2) = d_{\text{hyper}}(q_2, q_1)$,
2. $d_{\text{hyper}}(q_1, q_2) = 0$ if and only if $q_1 = q_2$, and
3. $d_{\text{hyper}}(q_1, q_3) \le d_{\text{hyper}}(q_1, q_2) + d_{\text{hyper}}(q_2, q_3)$.

But since q_1, q_2, q_3 were completely arbitrary, we see that $\mathbb{H}^3$ satisfies all of the properties of a metric space. □

We now know that $(\mathbb{H}^3, d_{\text{hyper}})$ is a perfectly kosher metric space; moreover, we know that $PSL(2, \mathbb{C})$ is a subset of Isom$(\mathbb{H}^3)$. Furthermore, it's not hard to see that *all* of Möb(2) is in Isom$(\mathbb{H}^3)$: any element in it can be written as either $z \mapsto \phi(z)$ or $z \mapsto \phi(-\overline{z})$ for some $\phi \in PSL(2, \mathbb{C})$. But

$$d_{\text{hyper}}\,(-\overline{q_1}, -\overline{q_2})) = \cosh^{-1}\left(1 + \frac{\left|-\overline{q_1} + \overline{q}_2\right|^2}{2\pi_j(-\overline{q_1})\pi_j(-\overline{q_2})}\right) = d_{\text{hyper}}(q_1, q_2),$$

so $q \mapsto -\overline{q}$ is an isometry of $\mathbb{H}^3$. We conclude that every element in Möb(2) can be understood as an isometry of hyperbolic 3-space. That there are no other isometries is less clear, but we will show this once we have defined some basic geometric notions in $\mathbb{H}^3$.

4.9 Hyperbolic Spheres, Planes, and Isometries

We extend various definitions that we had for the hyperbolic plane to $\mathbb{H}^3$. To start, let's consider spheres.

Definition 4.17 The *hyperbolic sphere with center ρ and radius r* in $\mathbb{H}^3$ is the locus of points $q \in \mathbb{H}^3$ such that $d_{\text{hyper}}(q, \rho) = r$.

Theorem 4.17 *Hyperbolic spheres are Euclidean spheres (but with different centers and radii).*

Proof Let S be a hyperbolic sphere centered at a point $q \in \mathbb{H}^3$. Consider the planes that pass through q and are orthogonal to $\mathbb{C}P^1$. Each of them is essentially a copy of $\mathbb{H}^2$ by Theorem 4.16, and so the set of points q' in that plane that are a fixed distance from q is a Euclidean circle. Furthermore, all of these planes are related by a rotation around the vertical line through q', and so the full set in $\mathbb{H}^3$ is what you get by taking such a Euclidean circle and rotating it around that line as the axis—in other words, it is a Euclidean sphere. □

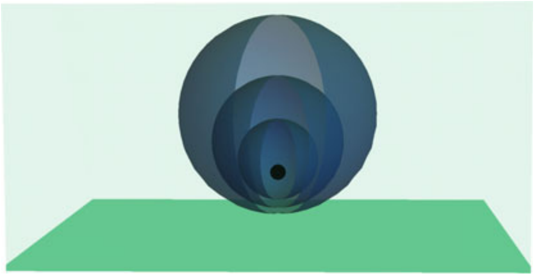


Fig. 4.24 Some concentric hyperbolic spheres which have been cut away slightly for easier viewing.

Figure 4.24 shows a family of concentric hyperbolic spheres to better illustrate this description. Next, we consider hyperbolic planes.

Definition 4.18 A *hyperbolic plane in* $\mathbb{H}^3$ is a locus of points in $\mathbb{H}^3$ of the form

$$\left\{ q \in \mathbb{H}^3 \middle| d_{\text{hyper}}(q, \rho_1) = d_{\text{hyper}}(q, \rho_2) \right\},$$

for some fixed points $\rho_1 \neq \rho_2 \in \mathbb{H}^3$.

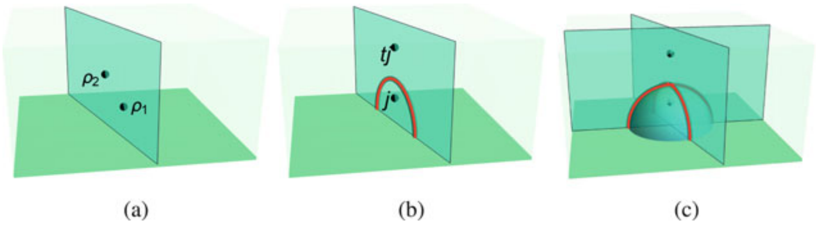
We can characterize what hyperbolic planes look like with the aid of a lemma.

Lemma 4.15 *For any $\rho_1 \neq \rho_2 \in \mathbb{H}^3$, there exists an isometry $\phi \in PSL(2, \mathbb{C})$ such that $\phi(\rho_1) = j$ and $\phi(\rho_2) = tj$ for some $t > 1$.*

Proof There exists a Euclidean plane E that passes through ρ_1 and ρ_2, and which is orthogonal to the boundary. By applying a translation $q \mapsto q + \rho$, we may assume that this plane passes through the origin. Since $E \cap \mathbb{H}^3$ is an isometric copy of $\mathbb{H}^2$, we know that there exists $\psi \in \text{Isom}(\mathbb{H}^2)$ such that $\psi(\rho_1) = j$ and $\psi(\rho_2) = tj$ for some $t > 1$. □

Theorem 4.18 *Hyperbolic planes are generalized spheres orthogonal to the boundary (restricted to $\mathbb{H}^3$). Conversely, any generalized sphere orthogonal to the boundary is a hyperbolic plane.*

Proof Choose any two points $\rho_1 \neq \rho_2 \in \mathbb{H}^3$. By Lemma 4.15, we know that there exists $\phi \in PSL(2, \mathbb{C})$ such that $\phi(\rho_1) = j$ and $\phi(\rho_2) = tj$ for some $t > 1$. Choose any Euclidean plane E' that passes through 0, j, and tj—it will automatically be orthogonal to the boundary, so we know that it is isometrically isomorphic to $\mathbb{H}^2$. Therefore, we know that the set of points $q \in E'$ such that $d_{\text{hyper}}(q, j) = d_{\text{hyper}}(q, tj)$ is a hyperbolic line—specifically, it is a circle centered at 0, and its radius is uniquely determined by t. Note that this does not depend on E', which means that the hyperbolic plane P defined by j and tj must be a Euclidean sphere centered at 0. (This is easiest to see from a picture: see Figure 4.25.) However, $\psi^{-1}(P)$ is the hyperbolic plane defined by ρ_1 and ρ_2; since $\text{Isom}(\mathbb{H}^3)$ preserves both angles and



(a) (b) (c)

Fig. 4.25 A visualization of the proof of Theorem 4.18. We start with two points ρ_1, ρ_2, through which we draw a plane orthogonal to the boundary in (a). In (b), we translate the plane so that it passes through the origin and move these two points to j and tj; at this stage, it is clear that the restriction of the hyperbolic plane to this plane is a circle centered at 0. Finally, in (c), we observe that we have rotational symmetry, so the hyperbolic plane is a sphere.

generalized spheres, we see that $\psi^{-1}(P)$ is also a generalized sphere orthogonal to the boundary.

As for the converse, note that any generalized sphere orthogonal to the boundary is uniquely determined by its intersection with the boundary, which is a circle. For any two circles C_1, C_2, there exists an element $\phi \in PSL(2, \mathbb{C})$ such that $\phi(C_1) = C_2$. Therefore, we can get any generalized sphere orthogonal to the boundary as the image under an element of $PSL(2, \mathbb{C})$ of a hyperbolic plane. But elements in $PSL(2, \mathbb{C})$ are isometries, so any such image is also a hyperbolic plane. □

The proof of this theorem suggests another natural object that we should consider.

Definition 4.19 A set $l \subset \mathbb{H}^3$ is a *hyperbolic line* if there exists a hyperbolic plane P such that $l \subset P$ and l is a hyperbolic line inside P (where P is considered as a copy of $\mathbb{H}^2$ in the usual way).

The awkward part of this definition is that it seems like it might depend on which particular plane P we choose. It does not.

Theorem 4.19 *For any $l \subset \mathbb{H}^3$, the following are equivalent.*

1. *l is a hyperbolic line.*
2. *l is the non-trivial intersection of two hyperbolic planes. (By non-trivial, we mean that the intersection is neither empty nor the union of the two planes.)*
3. *l is either a line or a half-circular arc orthogonal to the boundary $\mathbb{C}P^1$.*

Proof Suppose l is a hyperbolic line contained inside some plane P. It must intersect the boundary at two points $a, b \in \mathbb{C}P^1$. Choose any $\phi \in PSL(2, \mathbb{C})$ such that $\phi(a) = 0$ and $\phi(b) = \infty$. Then $\phi(P)$ must be a Euclidean plane through the origin orthogonal to the boundary, in which case $\phi(l)$ is the vertical line through the origin. Choose any other Euclidean plane E that passes through the origin and is orthogonal to the boundary. Evidently, $\phi(l)$ is the intersection of E and $\phi(P)$, which means that l is the intersection of $\phi^{-1}(E)$ and P. Therefore, if l is a hyperbolic line, then it is the non-trivial intersection of two hyperbolic planes. Conversely, if l is the intersection

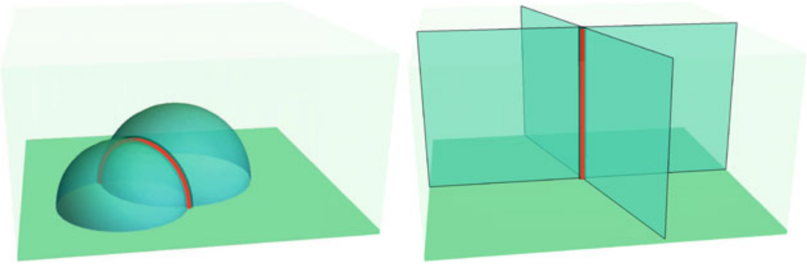


Fig. 4.26 A visualization of the idea behind Theorem 4.19: the general case of the intersection of any two hyperbolic planes (shown on the left) can be transformed to the specific case where they are both planes through the origin (shown on the right).

of two hyperbolic planes, we can use an isometry $\phi \in PSL(2, \mathbb{C})$ to move one of these planes to a Euclidean plane E through the origin orthogonal to the boundary, in which case $\phi(l)$ has to be either a line or a half-circular arc orthogonal to the boundary—that is, a hyperbolic line. However, if $\phi(l)$ is a hyperbolic line in E, then l is a hyperbolic line in $\phi^{-1}(E)$. The construction of this line can be seen in Figure 4.26.

On the other hand, since hyperbolic planes are either spheres or planes orthogonal to the boundary, l is an intersection of two such planes if and only if it is either a line or a half-circular arc orthogonal to the boundary. □

Note that it is immediate from this theorem that if l is a hyperbolic line inside of one plane, then it is a hyperbolic line in any plane that contains it. This is the same as it is in Euclidean space, as is the following characterization.

Theorem 4.20 *For any two distinct points $\rho_1, \rho_2 \in \mathbb{H}^3$, there is a unique hyperbolic line that passes through them. For any three distinct points $\rho_1, \rho_2, \rho_3 \in \mathbb{H}^3$, there exists a unique hyperbolic plane that passes through them.*

Proof For any $\rho_1 \neq \rho_2 \in \mathbb{H}^3$, by Lemma 4.15, there exists an isometry $\phi \in PSL(2, \mathbb{C})$ such that $\phi(\rho_1) = j$ and $\phi(\rho_2) = tj$ for some $t > 1$. There is a unique hyperbolic line that passes through these two points—this is the vertical line through the origin. If we also have a third point ρ_3, then $\phi(\rho_3)$ will be some point off this line, and there is exactly one hyperbolic plane that passes through all three of these points: this will be the Euclidean plane through the origin, orthogonal to the boundary, and passing through $\phi(\rho_3)$. □

This give us enough information to cleanly prove that $\text{Isom}(\mathbb{H}^3) = \text{Möb}(2)$.

Theorem 4.21 **(Characterization of Isometries of $\mathbb{H}^3$)**
If $\Psi \in Isom(\mathbb{H}^3, d_{hyper})$, then either $\Psi(q) = \psi(q)$ for some $\psi \in PSL(2, \mathbb{C})$, or $\Psi(q) = -\overline{\psi(q)}$ for some $\psi \in PSL(2, \mathbb{C})$.

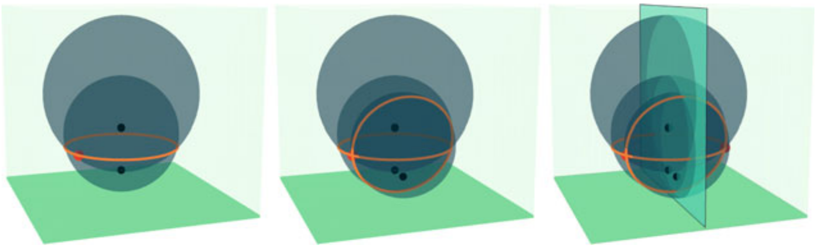


Fig. 4.27 A visualization of the proof of Theorem 4.21. In the first image, we see that measuring distance to a point (drawn in red) from j and ej narrows down its location to a circle which is the intersection of the hyperbolic spheres centered at j and ej. In the second image, we see the effect of adding a third point $i + j$ from which to measure distance: there are now only two possible points. These two points are reflections across the plane through i and j, which is illustrated in the last figure. Adding a fourth point not on this plane allows one to determine the point uniquely.

Proof We already know that the given transformations are isometries; the difficulty is in proving that they are the only ones. Choose any $\Psi \in \text{Isom}(\mathbb{H}^3)$ and consider the points $\rho_1 = \Psi(j)$, $\rho_2 = \Psi(ej)$. By Lemma 4.15, we know that there exists $\phi \in PSL(2, \mathbb{C})$ such that $\phi(\rho_1) = j$ and $\phi(\rho_2) = tj$ for some $t > 1$—indeed, $t = e$. Therefore, without loss of generality, we may assume that $\Psi(j) = j$ and $\Psi(ej) = ej$. For any point $\rho \in \mathbb{H}^3$, define $r_{\rho,1} = d_{\text{hyper}}(j, \rho)$ and $r_{\rho,2} = d_{\text{hyper}}(ej, \rho)$. Furthermore, let $S_{\rho,1}$ be the hyperbolic sphere with center j and radius $r_{\rho,1}$; let $S_{\rho,2}$ be the hyperbolic sphere with center ej and radius $r_{\rho,2}$. Clearly,

1. the intersection of $S_{\rho,1}$ and $S_{\rho,2}$ is a (Euclidean) circle centered on a point on z-axis and parallel to the xy-plane,
2. ρ lies on the aforementioned circle, and
3. $\Psi(\rho)$ also lies on this circle.

We want to show that any rotation around the z-axis can be obtained by an isometry in $PSL(2, \mathbb{C})$. Indeed,

$$\begin{pmatrix} e^{i\theta/2} & 0 \\ 0 & e^{-i\theta/2} \end{pmatrix} \in SL(2, \mathbb{C}),$$

and for any $\rho = z + tj$,

$$\begin{pmatrix} e^{i\theta/2} & 0 \\ 0 & e^{-i\theta/2} \end{pmatrix} .\rho = e^{i\theta/2}(z + tj)e^{i\theta/2} = e^{i\theta}z + j$$

is the desired rotation. So, we can find some rotation $\phi_\theta \in PSL(2, \mathbb{C})$ as above so that $\phi(\Psi(i + j)) = i + j$, but $\phi(j) = j$ and $\phi(ej) = ej$. So, without loss of generality, we may assume that $\Psi(j) = j$, $\Psi(ej) = ej$, and $\Psi(i + j) = i + j$. Now, for any $\rho \in \mathbb{H}^3$, define $r_{\rho,3} = d_{\text{hyper}}(q_1, \rho)$ and $S_{\rho,3}$ to be the hyperbolic sphere with center q_1 and radius $r_{\rho,3}$. We see that

1. the intersection of $S_{\rho,1}$, $S_{\rho,2}$, and $S_{\rho,3}$ is one of two points,

2. these two points are reflections of each other across the plane through j, ej, and $i + j$,
3. the reflection across the aforementioned plane is given by $\rho \mapsto -\overline{\rho}$, and
4. $\rho, \Psi(\rho)$ must be one of these two points.

Now, choose any point q not in the plane through $i, ej, i + j$. We have shown that either $\Psi(q) = q$ or $\Psi(q) = -\overline{q}$. By composing with $\rho \mapsto -\overline{\rho}$ if need be, we can assume that $\Psi(j) = j$, $\Psi(ej) = ej$, $\Psi(i + j) = i + j$, and $\Psi(q) = q$. Now, we can use the same triangulation technique that we had for $\mathbb{H}^2$, which is demonstrated in Figure 4.27. It must be that $\Psi(\rho) = \rho$ for all $\rho \in \mathbb{H}^3$ because the only other possibility would be that $\Psi(\rho) = -\overline{\rho}$, but that would change the distance to q. □

Corollary 4.3 *The set of isometries $Isom(\mathbb{H}^3, d_{hyper})$ is a group if we take function composition to be the operation. Furthermore, the map*

$$\begin{aligned} M\ddot{o}b(2) &\to Isom(\mathbb{H}^3, d_{hyper}) \\ \phi &\mapsto \begin{cases} \rho \mapsto \gamma.\rho & \textit{if } \phi(z) = \gamma.z \textit{ with } \gamma \in SL(2,\mathbb{C}) \\ \rho \mapsto -\overline{\gamma.\rho} & \textit{if } \phi(z) = -\overline{\gamma.z} \textit{ with } \gamma \in SL(2,\mathbb{C}) \end{cases} \end{aligned}$$

is a group isomorphism.

Proof By the previous theorem, we know that this map is surjective. Indeed, it must be injective, since every element in Möb(2) has a distinct effect on the boundary of $\mathbb{H}^3$. Function composition on the left maps to function composition on the right—therefore, Isom($\mathbb{H}^3$, d_{hyper}) is a group and this is an isomorphism. □

Problems

4.1 COMPUTATIONAL EXERCISES

1. a) Prove that

$$T_t := \begin{pmatrix} \cosh(t/2) & \sinh(t/2) \\ \sinh(t/2) & \cosh(t/2) \end{pmatrix} \in SU(1,1)$$

   for all $t \in \mathbb{R}$.

   b) Find $T_t.0$.
2. Let C be a hyperbolic circle in $\mathbb{D}^2$ with center $re^{i\theta}$ and radius R. Find its Euclidean center and radius as functions in r, θ, R. *(Hint: the previous exercise may be helpful.)*
3. Show that

$$\begin{array}{lll} ij = k & jk = i & ki = j \\ ji = -k & kj = -i & ki = -j, \end{array}$$

   where i, j, k are quaternions.
4. a) Find integers a, b, c, d such that if $q = a + bi + cj + dk$, then

   a. $|q|^2 = 2$
   b. $|q|^2 = 3$
   c. $|q|^2 = 5$

   b) Find integers a, b, c, d such that if $q = a + bi + cj + dk$, then $|q|^2 = 30$. *(Hint: the results of the previous part are very helpful here.)*

   c) Make a conjecture regarding for which integers k the equation $a^2 + b^2 + c^2 + d^2 = k$ is solvable in integers.

4.2 PROOFS

1. The three standard hyperbolic functions are *hyperbolic sine*, *hyperbolic cosine*, and *hyperbolic tangent*, defined as

$$\begin{aligned} \sinh(x) &= \frac{e^x - e^{-x}}{2} \\ \cosh(x) &= \frac{e^x + e^{-x}}{2} \\ \tanh(x) &= \frac{\sinh(x)}{\cosh(x)}. \end{aligned}$$

The domains of hyperbolic sine and cosine can be taken to be either $\mathbb{R}$ or $\mathbb{C}$. The hyperbolic tangent can be defined wherever $\cosh(x) \neq 0$. 1

a) Prove that $\cosh(z) = 0$ if and only if $z = \pi i(k + 1/2)$ for some integer k.
b) Sketch the graphs of sinh, cosh, tanh as functions on $\mathbb{R}$.
c) Prove that $\cosh(x)^2 - \sinh(x)^2 = 1$.
d) Prove that

$$\sinh(2x) = 2\sinh(x)\cosh(x)$$
$$\cosh(2x) = \cosh(x)^2 + \sinh(x)^2.$$

e) Prove that $\sinh(iz) = i\sin(z)$ and $\cosh(iz) = \cos(z)$.

2. Prove that $SL(2, \mathbb{R})$ is a subgroup of $SL(2, \mathbb{C})$.
3. Prove that $\text{Isom}(\mathbb{H}^2)$ is a group.
4. Complete the proof of Lemma 4.6.
5. Prove that a generalized circle C is orthogonal to $x = 0$ and $y = 0$ if and only if it is a circle centered at the origin.
6. We show that the metric definition of a plane coincides with the usual definition of a plane in Euclidean space.

a) Show that if $p_1 \neq p_2 \in \mathbb{R}^3$, then the set of points p_3 such that $d_{\text{Euclid}}(p_1, p_3) = d_{\text{Euclid}}(p_2, p_3)$ is a plane. *(Hint: use Euclidean isometries to reduce to a simple case.)*
b) Prove that if P is a plane in $\mathbb{R}^3$, then there exist $p_1 \neq p_2 \in \mathbb{R}^3$ such that $p_3 \in P$ if and only if $d_{\text{Euclid}}(p_1, p_3) = d_{\text{Euclid}}(p_2, p_3)$.

7. Finish the proof of Corollary 4.2.
8. Prove that $SU(m, n)$ is a group under matrix multiplication.
9. Prove Theorem 4.12. *(Hint: use the fact that you already know that the isometries of $\mathbb{H}^2$ come from elements of $SL(2, \mathbb{R})$.)*
10. Consider the line $x = 0$ in $\mathbb{H}^2$. Fix some $r > 0$. For any point it on the line, the hyperbolic line orthogonal to $x = 0$ at that point is the half-circle consisting of all points $te^{i\theta}$ with $0 < \theta < \pi$.

a) Show that for any $t > 0$, there exists a unique point $te^{i\theta_t}$ to the right of the line $x = 0$ such that $d_{\text{hyper}}(it, te^{i\theta_t}) = r$. What is θ_t? Does it depend on t?
b) As one varies the parameter t, what is the curve $te^{i\theta_t}$? Is it a hyperbolic line? How does the situation here differ from the Euclidean plane?

11. Let $\Phi : \mathbb{D}^2 \to \mathbb{D}^2$ be a map such that for any $z_1, z_2, z_3 \in \mathbb{D}^3$,

$$\frac{d_{\text{hyper}}(z_1, z_2)}{d_{\text{hyper}}(z_1, z_3)} = \frac{d_{\text{hyper}}(\Phi(z_1), \Phi(z_2))}{d_{\text{hyper}}(\Phi(z_1), \Phi(z_3))};$$

that is, Φ is a similarity.

a) Prove that there exists some constant $c > 0$ such that

$$d_{\text{hyper}}(\Phi(z_1), \Phi(z_2)) = c d_{\text{hyper}}(z_1, z_2)$$

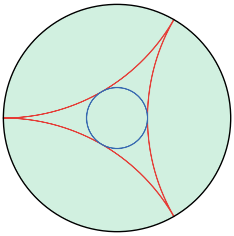
for all $z_1, z_2 \in \mathbb{D}^2$. This is known as the constant of proportionality.

b) Prove that if $c \leq 1$ then the constant of proportionality of Φ^{-1} is greater than or equal to 1.

c) Prove that the image of a hyperbolic line under a similarity is a hyperbolic line.

d) Prove that the image of a hyperbolic circle with hyperbolic radius R under a similarity with a constant of proportionality c is a hyperbolic circle with hyperbolic radius cR.

e) Consider the following configuration of lines and circles in the hyperbolic plane.



That is, in $\mathbb{D}^2$, we consider an idealized hyperbolic triangle with all of its vertices on the boundary, and a circle inscribed inside of it. Prove that any similarity of $\mathbb{D}^2$ must send this configuration to an idealized hyperbolic triangle with all of its vertices on the boundary, and a circle inscribed inside of it. Furthermore, show that there exists some isometry that will send it to the same idealized hyperbolic triangle and circle inscribed inside of it.

f) Prove that for any similarity, its constant of proportionality is 1; that is to say, it is an isometry. *(Hint: consider the configuration from before. How can the hyperbolic radius of the inscribed circle change under a similarity?)*

12. Consider the map

$$\Phi : \mathbb{C} \to \text{Mat}(2, \mathbb{R})$$
$$x + iy \mapsto \begin{pmatrix} x & y \\ -y & x \end{pmatrix}$$

a) Prove that for any $z_1, z_2 \in \mathbb{C}$, $\Phi(z_1 + z_2) = \Phi(z_1) + \Phi(z_2)$ and $\Phi(z_1 z_2) = \Phi(z_1)\Phi(z_2)$. That is, in some sense, every complex number can be faithfully represented by a matrix with real coefficients.
b) Prove that for any $z \in \mathbb{C}$, $|z|^2 = \det(\Phi(z))$ and $2\Re(z) = \mathrm{tr}(\Phi(z))$.

13. Consider the map

$$\Phi : \mathcal{H} \to \mathrm{Mat}(2, \mathbb{C})$$
$$x + yi + zj + tk \mapsto \begin{pmatrix} x + yi & z + ti \\ -z + ti & x - yi \end{pmatrix}$$

a) Prove that for any $q_1, q_2 \in \mathcal{H}$, $\Phi(q_1 + q_2) = \Phi(q_1) + \Phi(q_2)$ and $\Phi(q_1 q_2) = \Phi(q_1)\Phi(q_2)$. That is, in some sense, every quaternion can be faithfully represented by a matrix with complex coefficients.
b) Prove that for any $q \in \mathbb{C}$, $|q|^2 = \det(\Phi(q))$ and $\mathrm{tr}(q) = \mathrm{tr}(\Phi(q))$.
c) Prove that $q^{-1} = \overline{q}/|q|^2$ using the result of the previous part.

14. Find an injective map $\Phi : \mathcal{H} \to \mathrm{Mat}(4, \mathbb{R})$ with the property that for all $q_1, q_2 \in \mathcal{H}$, $\Phi(q_1 + q_2) = \Phi(q_1) + \Phi(q_2)$ and $\Phi(q_1 q_2) = \Phi(q_1)\Phi(q_2)$.
15. Prove that any quaternion q is a solution to the equation $X^2 - \mathrm{tr}(q)X + |q|^2$.
16. Prove that for any $r > 0$, there exist infinitely many quaternions q such that $q^2 = -r$, and there are only two such that $q^2 = r$.
17. We fill in the gaps in the proof of Theorem 4.13. Let q be any quaternion.

a) Prove that $q = \overline{q}$ if and only if $q \in \mathbb{R}$.
b) Prove that $|q| = |\overline{q}|$.
c) Prove that $q = 0$ if and only if $|q| = 0$.

18. Let q be a quaternion.

a) Prove that if $qi = iq$, $qj = jq$, and $qk = kq$, then $q \in \mathbb{R}$.
b) Prove that the following three conditions are equivalent.

a. $q \in \mathbb{R}$.
b. $qq' = q'q$ for all quaternions q' with no real component.
c. $qq' = q'q$ for all quaternions q'.

4.3 PROOFS (Calculus)

1. Let sinh, cosh, tanh be as defined in Exercise 4.2.1.

a) Prove that

$$\lim_{x\to\infty}\sinh(x)=\infty \quad \lim_{x\to-\infty}\sinh(x)=-\infty \quad \lim_{x\to\infty}\cosh(x)=\infty$$
$$\lim_{x\to-\infty}\cosh(x)=\infty \quad \lim_{x\to\infty}\tanh(x)=1 \quad \lim_{x\to-\infty}\tanh(x)=-1.$$

b) Prove that

$$\frac{d}{dx}(\sinh(x))=\cosh(x)$$
$$\frac{d}{dx}(\cosh(x))=\sinh(x)$$
$$\frac{d}{dx}(\tanh(x))=\cosh(x)^{-2}.$$

c) Prove that $\cosh(x) > 0$ for all real x. *(Hint: it is continuous and never zero.)*

d) Prove that sinh and tanh are strictly increasing on $\mathbb{R}$—that is, if $x < y$, then $\sinh(x) < \sinh(y)$ and $\tanh(x) < \tanh(y)$.

e) Prove that $\tanh(x) > 0$ if $x > 0$, $\tanh(x) < 0$ if $x < 0$, and $\tanh(x) = 0$ if $x = 0$.

f) Prove that cosh is strictly increasing on $[0, \infty)$.

g) Prove that, considered as functions on $[0, \infty)$, sinh and cosh attain a minimum at $x = 0$, and that minimum is 0 and 1, respectively.

h) Prove that $\sinh([0, \infty)) = [0, \infty)$ and $\cosh([0, \infty)) = [1, \infty)$.

i) Prove that the functions

$$\sinh : [0,\infty) \to [0,\infty)$$
$$\cosh : [0,\infty) \to [1,\infty)$$
$$\tanh : (-\infty,\infty) \to (-1,1)$$

are all bijective.

j) Since sinh, cosh, tanh defined on the domains above are bijective, they have well-defined inverses. Prove that $\cosh^{-1}(x) = \ln(x + \sqrt{x^2 - 1})$.

k) Prove that

$$\frac{d}{dx}(\cosh^{-1}(x)) = (x^2-1)^{-1/2}.$$

Conclude that $\cosh^{-1}$ is a strictly increasing function.

l) Prove that the function

$$\begin{aligned} p : (-\infty, \infty) &\mapsto \mathbb{R}^2 \\ t &\mapsto (\cosh(t), \sinh(t)) \end{aligned}$$

is injective and that its image is the right half of the hyperbola $x^2 - y^2 = 1$. This is the origin of the term "hyperbolic function."

2. Let $p_1, p_2 : \mathbb{R} \to \mathbb{H}^2$ be paths along hyperbolic lines. Prove that if $\lim_{t\to\infty} p_1(t) \neq \lim_{t\to\infty} p_2(t)$, then

$$\lim_{t\to\infty} d_{\text{hyper}}\left(p_1(t), p_2(t)\right) = \infty.$$

What happens if p_1 and p_2 approach the same point at infinity? Give a geometric interpretation. *(Hint: use an isometry to simplify to some easy to consider case.)*

4.4 PROOFS (Linear Algebra)

1. Define $O(3)$ to be the set of 3×3 real matrices M with the property that $M^T = M^{-1}$, where M^T is the transpose of M. Define $SO(3)$ to be the subset of $O(3)$ consisting of matrices with determinant 1.

   a) Let $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ be the column vectors of a real matrix M. Prove that $M \in O(3)$ if and only if $\mathbf{e}_i \cdot \mathbf{e}_j = 1$ if $i = j$ and is 0 otherwise.
   b) Prove that a 3×3 real matrix M is in $O(3)$ if and only if M preserves the dot product, in the sense that $\mathbf{v}.\mathbf{w} = (M\mathbf{v}).(M\mathbf{w})$ for all $\mathbf{v}, \mathbf{w} \in \mathbb{R}^3$. *(Hint: write the dot product* $\mathbf{v}.\mathbf{w} = \mathbf{v}^T\mathbf{w}$.)
   c) Prove that $O(3)$ is a group under matrix multiplication.
   d) Prove that $SO(3)$ is a subgroup of $O(3)$.
   e) Prove that if $M \in SO(3)$, then $\mathbf{v} \times \mathbf{w} = (M\mathbf{v}) \times (M\mathbf{w})$ for all $\mathbf{v}, \mathbf{w} \in \mathbb{R}^3$.

2. Prove that $\mathcal{H}^1$—the set of quaternions of norm 1—is a group with multiplication as the operation.
3. Prove that

$$\begin{aligned} \mathcal{H}^1 &\to SU(1,1) \\ x + yi + zj + tk &\mapsto \begin{pmatrix} x + yi & z + ti \\ -z + ti & x - yi \end{pmatrix} \end{aligned}$$

   is a group isomorphism.
4. Define $\mathcal{H}^0$ to be the set of quaternions with no real component, and consider the bijection

$$\begin{aligned} Q : \mathbb{R}^3 &\to \mathcal{H}^0 \\ (x, y, z) &\mapsto xi + yj + zk. \end{aligned}$$

a) Prove that if $q' \in \mathcal{H}^0$ and $q \in \mathcal{H}^1$, then $qq'q^{-1} \in \mathcal{H}^0$.

b) Prove that

$$\begin{aligned} A : \mathcal{H}^1 \times \mathbb{R}^3 &\mapsto \mathbb{R}^3 \\ (q, \mathbf{v}) &\mapsto Q^{-1}\left(qQ(\mathbf{v})q^{-1}\right) \end{aligned}$$

is a well-defined action of $\mathcal{H}^1$ on $\mathbb{R}^3$—for simplicity, we shall simply write $q.\mathbf{v}$ for $A(q, \mathbf{v})$.

c) Let $q, q' \in \mathcal{H}^1$. Prove that if $q.\mathbf{v} = q'.\mathbf{v}$ for all $\mathbf{v} \in \mathbb{R}^3$, then either $q = q'$ or $q = -q'$. *(Hint: you might want to use the result from Exercise 4.2.18.)*

5. Consider the map

$$\begin{aligned} \mathcal{H}^1 &\to \mathrm{Mat}(3, \mathbb{R}) \\ q &\mapsto (q.\mathbf{i}, q.\mathbf{j}, q.\mathbf{k}) \end{aligned}$$

where $\mathbf{i} = (1, 0, 0)$, $\mathbf{j} = (0, 1, 0)$, and $\mathbf{k} = (0, 0, 1)$.

a) Prove that $(q.\mathbf{i}, q.\mathbf{j}, q.\mathbf{k}) \in O(3)$. *(Hint: use the fact that* $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3) \in O(3)$ *if and only if* $\mathbf{e}_i \cdot \mathbf{e}_j = 0$ if $i \neq j$, *and* 1 *otherwise.)*

b) Prove that $(q.\mathbf{i}, q.\mathbf{j}, q.\mathbf{k}) \in SO(3)$. *(Hint: use the fact that the determinant of* $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$ *is* $\mathbf{e}_1 \cdot (\mathbf{e}_2 \times \mathbf{e}_3)$.)

c) Prove that if $M = (q.\mathbf{i}, q.\mathbf{j}, q.\mathbf{k})$ then $q.\mathbf{v} = M\mathbf{v}$ for all $\mathbf{v} \in \mathbb{R}^3$.

d) Prove that

$$\begin{aligned} \mathcal{H}^1 &\to SO(3) \\ q &\mapsto (q.i, q.j, q.k) \end{aligned}$$

is a group homomorphism.

6. $SO(3)$ is the group of rotations around the origin in $\mathbb{R}^3$, so the previous exercise shows that quaternions of norm 1 correspond to rotations. Try to prove that all rotations arise in this fashion. *(Hint: this is a hard problem. Don't be ashamed to ask a teacher for help.)*

4.5 PROOFS (Metric Geometry)

1. Let (M, d) be any metric space. Let X be a subset of M. Prove that (X, d) is a metric space.

2. Let X be any set. Define a function

$$\begin{aligned} d_{\text{discrete}} &: X \times X \to \mathbb{R} \\ (x, y) &\mapsto \begin{cases} 0 & \text{if } x = y \\ 1 & \text{otherwise.} \end{cases} \end{aligned}$$

   Prove that (X, d_{discrete}) is a metric space. (It is usually called the *discrete metric space*.)

3. Let X be some set of symbols—it could be letters of the English alphabet, or $\{0, 1\}$, or something else entirely. We shall call X an *alphabet*. A *string* is an element of X^n for some natural number n. For some fixed n, and any two strings $s, t \in X^n$, define $d_{\text{Hamming}}(s, t)$ to be the number of entries in which s and t differ. Prove that $(X^n, d_{\text{Hamming}})$ is a metric space. (It is usually called the *Hamming space* in honor of Richard Hamming, who described it in 1950 as part of his work on error-correcting codes, which have been an important topic in computer science ever since.)

4. We prove that for any positive integer n, the Euclidean distance

$$\begin{aligned} d_{\text{Euclid}} &: \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R} \\ (\mathbf{x}, \mathbf{y}) &\mapsto |\mathbf{x} - \mathbf{y}|, \end{aligned}$$

   where we may define $|\mathbf{v}| = \sqrt{\mathbf{v}^T \mathbf{v}}$, turns $\mathbb{R}^n$ into a metric space.

   a) Prove that for any $\mathbf{x} = (x_1, x_2, \ldots x_n) \in \mathbb{R}^n$ and any i, $|\mathbf{x}| \geq |x_i|$.
   b) Prove that $|\mathbf{x}| = 0$ if and only if $\mathbf{x} = 0$.
   c) Given $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, solve for $c \in \mathbb{R}$ such that $|\mathbf{x} - c\mathbf{y}|^2 = 0$. (You will get some expression in terms of $\mathbf{x}^T \mathbf{y}$, $|\mathbf{x}|$, and $|\mathbf{y}|$.)
   d) Prove the Cauchy-Schwarz inequality: for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, $|\mathbf{x}^T \mathbf{y}| \leq |\mathbf{x}||\mathbf{y}|$. *(Hint: since* $|\mathbf{x} - c\mathbf{y}|^2 \geq 0$, *there can be at most one* $c \in \mathbb{R}$ *such that* $|\mathbf{x} - c\mathbf{y}|^2 = 0$. *Use this and the result of the previous part.)*
   e) Prove that for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, $|\mathbf{x} + \mathbf{y}| \leq |\mathbf{x}| + |\mathbf{y}|$. *(Hint: expand* $|\mathbf{x} + \mathbf{y}|^2$ *and use Cauchy-Schwarz.)*
   f) Prove that $(\mathbb{R}^n, d_{\text{Euclid}})$ is a metric space.

5. Define a function

$$\begin{aligned} f &: \mathbb{H}^2 \times \mathbb{H}^2 \to \mathbb{R} \\ (x, y) &\mapsto \begin{cases} 0 & \text{if } x = y \\ t & \begin{array}{l} \text{if } \gamma(x) = i, \gamma(y) = it \text{ with } t > 1 \\ \text{for some } \gamma \in \text{Möb}(2) \text{ such that } \gamma(\mathbb{H}^2) = \mathbb{H}^2. \end{array} \end{cases} \end{aligned}$$

   Define a function

$$d : \mathbb{H}^2 \times \mathbb{H}^2 \to$$
$$(x, y) = \begin{cases} 0 & \text{if } x = y \\ 2 - f(x, y)^{-1} & \text{otherwise.} \end{cases}$$

a) Prove that $(\mathbb{H}^2, d)$ is a metric space.
b) Prove that $\text{Isom}(\mathbb{H}^2, d) = \text{Möb}(2)$.

6. Let (M, d) be a metric space. For any $f \in \text{Isom}(M, d)$, prove that f is injective.
7. Let (M, d) be a metric space. Consider the set of isometries $\text{Isom}(M)$ with function composition as the operation.

   a) Prove that if $f, g \in \text{Isom}(M)$, then $f \circ g \in \text{Isom}(M)$.
   b) Prove that $f \circ (g \circ h) = (f \circ g) \circ h$ for any $f, g, h \in \text{Isom}(M)$.
   c) Prove that the identity function $id : M \to M$ satisfies $id \circ f = f \circ id = f$ for all $f \in \text{Isom}(M)$. Thus, $(M, \circ)$ satisfies all of the properties of a group other than that it might not have inverses. An algebraic structure with these properties is known as a *monoid*.

8. Let (M, d) be a metric space. Let $\Phi \in \text{Isom}(M)$. Prove that if Φ is surjective, then $\Phi^{-1} \in \text{Isom}(M)$.

Properties of Hyperbolic Geometry

5

In which parallel lines have an odd habit of getting farther and farther apart.

The previous chapter dealt almost exclusively with the problem of constructing various models of the hyperbolic plane and hyperbolic 3-space and showing that we could get the isometry groups via linear fractional transformations. This chapter will reap the rewards of our hard work: we will start by using the techniques we have developed to understand the geometry of these spaces; we will end by using these spaces to prove results about the groups $PSL(2, \mathbb{R})$ and $PSL(2, \mathbb{C})$. Thus, we will see that the group theory and the metric geometry feed each other, enriching both.

5.1 Peculiarities of Hyperbolic Geometry

In the introduction to Chapter 4, we said that hyperbolic space was the original example of a non-Euclidean space—that is, a geometry that satisfied all of Euclid's axioms save for the last one, the Parallel Postulate. Let's write down these axioms.

1. A straight line may be drawn between any two points.
2. Any terminated straight line may be extended indefinitely.
3. A circle may be drawn with any given point as center and any given radius.
4. All right angles are equal.
5. If two straight lines in a plane are met by another line, and if the sum of the internal angles on one side is less than two right angles, then the straight lines will meet if extended sufficiently on the side on which the sum of the angles is less than two right angles.

A. Sheydvasser, *Linear Fractional Transformations*, Undergraduate Texts in Mathematics, https://doi.org/10.1007/978-3-031-25002-6_5

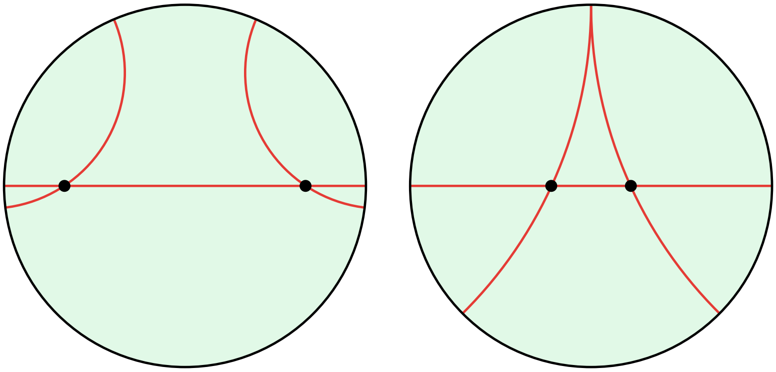


Fig. 5.1 Hyperbolic counterexamples to Euclid's Parallel Postulate.

Why does the hyperbolic plane satisfy all of these save for the last one? Let's go through them one by one.

A straight line... This is certainly true of $\mathbb{H}^2$—indeed, we showed the stronger result that there is a unique line through any two points in $\mathbb{H}^3$. (This is stronger since we know that we can always embed $\mathbb{H}^2$ inside $\mathbb{H}^3$ as a hyperbolic plane.)

Any terminated straight line... If this is true of Euclidean geometry, it certainly has to be true of hyperbolic geometry, given that after applying an isometry, we can assume that the hyperbolic line is just a Euclidean line.

A circle may be drawn... This is certainly true of hyperbolic geometry: we showed exactly how to define hyperbolic circles with arbitrary centers and radii.

All right angles... There is some question here about what exactly is meant by a right angle, but it turns out to be irrelevant since angles in $\mathbb{H}^2$ and $\mathbb{D}^2$ are simply angles in Euclidean geometry.

The last axiom, however, is false in hyperbolic geometry. This is because it is possible for two hyperbolic lines to initially move toward each other, but ultimately to diverge away! Furthermore, it is possible for two hyperbolic lines to keep getting closer and closer but never quite reaching each other. Examples of how this can happen are depicted in Figure 5.1.

It is common in axiomatic geometry to replace the Parallel Postulate with Playfair's axiom "Given a line l and a point p not on l, there is exactly one line l' that passes through p and does not intersect l." Not surprisingly (since it is equivalent to Euclid's Parallel Postulate), this statement is also entirely false for hyperbolic geometry, as demonstrated in Figure 5.2. In fact, for any line l and a point p not on l, there exist infinitely many lines l' that pass through p and do not intersect l. Evidently, hyperbolic lines are shy and do not often want to meet. As an interesting side-note, one historical way of attempting to prove the Parallel Postulate ran as follows: starting with a line l and point p not on the line, consider fixing a ruler with one end on p and the other end orthogonal to l. As you then translate this ruler along

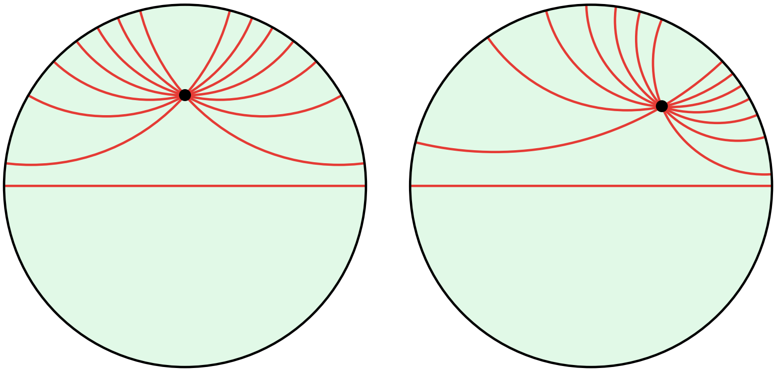


Fig. 5.2 Hyperbolic counterexamples to Playfair's axiom.

the line, what is the curve swept out by the other end? Naively, it might seem like it should be exactly the desired parallel line through p; certainly, that is exactly what you get in Euclidean space. But in the hyperbolic plane, this curve is not a line at all! (See Exercise 4.2.10.)

On the other hand, since the other axioms of Euclidean geometry do hold for hyperbolic space, the ASA, SAS, and SSS triangle congruence theorems all hold for it, exactly as they did for spherical geometry. What is significantly more surprising is that there is a triangle congruence theorem in hyperbolic geometry that has no Euclidean analog: the AAA theorem. That is, any two hyperbolic triangles with the same angle measures must be congruent, in the sense that there is an isometry moving one to the other. We relegate proofs of these triangle theorems to the exercises. (Specifically, Exercises 5.2.8-13.)

We will, however, prove a related theorem: the sum of the angles of a hyperbolic triangle is always strictly less than π. In fact, we want to show something stronger: the amount by which it is smaller is *exactly* the area of the triangle. Now, of course, there is a bit of a snag; we never defined what we mean by hyperbolic area. The rough idea is this: you can set up an integral for the area via a Riemann sum by assuming that the area of a very small rectangle in the hyperbolic plane should be approximately its hyperbolic length times its hyperbolic width. The exact details are worked out in Exercises 5.3.2 and 5.3.3, but the main facts that we are going to use are the following—below, A will always denote a subset of the hyperbolic plane with defined hyperbolic area $\mu(A)$.

1. For any isometry Φ, $\mu(A) = \mu(\Phi(A))$.
2. If $A_1, A_2, \ldots$ are non-intersecting subsets, then $\mu\left(\bigcup_i A_i\right) = \sum_i \mu(A_i)$.
3. Let A be the interior of an idealized hyperbolic triangle with all vertices on the boundary (that is, its sides are all hyperbolic lines which don't intersect inside the plane but do intersect on the boundary). Then $\mu(A) = \pi$.

This is enough for us to get the desired result, with the aid of a lemma.

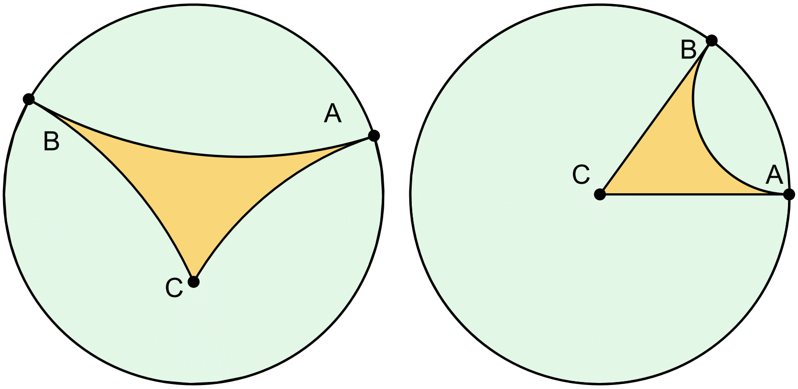


Fig. 5.3 An illustration of the proof of Lemma 5.1: on the left, we start with an idealized triangle ΔABC with one vertex C off the boundary; on the right, we force this triangle into a standard position where the claim is clear.

Lemma 5.1 *Let ΔABC, $\Delta A'B'C'$ be idealized hyperbolic triangles such that A, A', B, B' are on the boundary, but C, C' are not. If $\angle C = \angle C'$, then there exists an isometry Φ such that $\Phi(A) = A'$, $\Phi(B) = B'$, and $\Phi(C) = C'$.*

Proof There exist hyperbolic isometries Ψ, Ψ' such that $\Psi(A) = \Psi'(A') = 1$, $\Psi(C) = \Psi'(C') = 0$, and $\Im(\Psi(B)), \Im(\Psi'(B')) > 0$ in $\mathbb{D}^2$. We can show this as follows: choose any point p on the line segment between C and A. We know that there exists an isometry sending C to 0 and p to some $1 > t > 0$; strictly speaking, we proved this for $\mathbb{H}^2$, but $\mathbb{H}^2$ and $\mathbb{D}^2$ are isometrically isomorphic. That isometry, of course, has to send A to 1. Now, either the image of B is above the real line or it is below it (it cannot be on the real line, as then ΔABC will not be a triangle). The transformation $z \mapsto \bar{z}$ is an isometry of $\mathbb{D}^2$ that will switch the position if required.

So, to summarize, without loss of generality, we may assume that $A = A' = 1$, $C = C' = 0$, and $\Im(B), \Im(B') > 0$. But now we see that ΔABC looks something like a wedge and, in particular, it is obvious that $\angle C$ completely determines where on the boundary B is. (See Figure 5.3.) Therefore, $B = B'$ and we are done. □

Theorem 5.1 **(Lambert's theorem)** *Let ΔABC be a hyperbolic triangle. Then the (hyperbolic) area of ΔABC is $\pi - (a + b + c)$, where a, b, c are the angle measures of the vertices of ΔABC.*

Proof If we take an idealized hyperbolic polygon with n vertices all on the boundary, it can be obtained by pasting together $n - 2$ such idealized triangles, hence its area must be $(n - 2)\pi$. However, if we take such a polygon and put all of its vertices equidistant along the circle, then we can also split this idealized polygon into n isometric pieces. Each of these pieces is an idealized triangle with two vertices on the boundary and the third of angle measure $2\pi/n$. By Lemma 5.1, all of these triangles are related by isometries, hence they must have the same area:

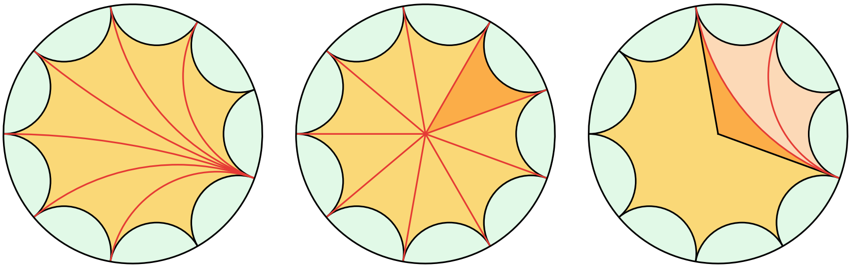


Fig. 5.4 An illustration of the first part of the proof of Lambert's theorem: we decompose an idealized n-gon in two different ways. The first allows us to compute its area; the second allows us to compute the area of idealized triangles with one off-boundary angle of measure $2\pi/n$. The last image shows that we can subtract off idealized triangles to compute the area of idealized triangles with one off-boundary angle of measure $2\pi m/n$.

$(n-2)\pi/n = \pi - 2\pi/n$, to be precise. This proves the theorem for all idealized hyperbolic triangles with $a = 0$, $b = 0$, and $c = 2\pi/n$ for some integer $n \geq 4$.

We can then bootstrap to get the case $a = 0$, $b = 0$, and $c = 2\pi m/n$ for some integer $1 \leq m \leq n$. We do this as follows: first paste together m triangles with $a = 0$, $b = 0$, and $c = 2\pi/n$ at the vertex not on the boundary. This gives an idealized polygon with one vertex of angle measure $2\pi m/n$ and $m+1$ vertices on the boundary. This can be fixed easily enough by subtracting off $m-1$ triangles with all their vertices on the boundary—how this is done is illustrated in Figure 5.4. This gives us the desired triangle with $a = 0$, $b = 0$, and $c = 2\pi m/n$, and its area must be $m(\pi - 2\pi/n) - (m-1)\pi = \pi - 2\pi m/n$, as expected. This allows us to approximate any idealized hyperbolic triangle with two vertices on the boundary arbitrarily closely and so we must conclude that, in general, the area of such a triangle is $\pi - c$ if c is the angle measure of the non-idealized vertex.

Next, consider an idealized hyperbolic triangle with just one vertex on the boundary. Use an isometry to send this vertex to ∞ in the Poincaré half-plane, and then use isometries $z \mapsto z + x_0$ and $z \mapsto \lambda z^2$ to move one of the other vertices to i. Call the angle measure of that vertex b, and the angle measure of the other c. It is easy to see that if you glue to this triangle another idealized triangle with two vertices on the boundary and the last vertex of angle measure $\pi - c$, then together they form an idealized triangle with two vertices on the boundary and the last vertex of angle measure b. Consequently, the area of our desired triangle is $(\pi - b) - (\pi - (\pi - c)) = \pi - b - c$.

Finally, consider a non-idealized triangle ΔABC with angle measures a, b, c. Use an isometry to move A to i in the Poincaré plane, B to some point it with $t > 1$, and C to some point $x + iy$ with $x, y > 0$. Draw a vertical line l through C, and let α be the angle measure between this line and BC. Then one can paste two idealized hyperbolic triangles to ΔABC: one along BC, with angle measures $\pi - b$, α, and 0, and the other along l, with angle measures $\pi - \alpha - c$, 0, and 0. Choosing these idealized hyperbolic triangles so that their sides lie on the hyperbolic line through AC, $x = 0$, and l, we get a new idealized hyperbolic triangle with two vertices on

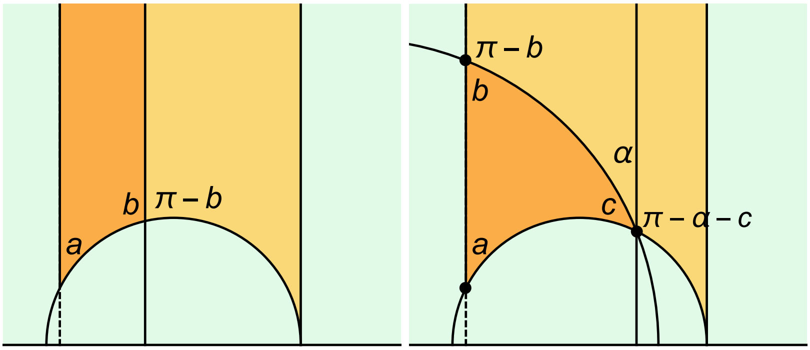


Fig. 5.5 An illustration of the second part of the proof of Lambert's theorem. The first image shows how to subtract off idealized triangles with one off-boundary angle to compute the area of idealized triangles with one vertex on the boundary. The second image shows how to carve up idealized triangles to compute the area of a non-idealized triangle.

the boundary and the remaining one with an angle measure of a—this is shown in Figure 5.5. Therefore, the area of ΔABC is

$$\pi - a - (\pi - (\pi - b + \alpha)) - (\pi - (\pi - \alpha - c)) = \pi - a - b - c,$$

as we initially claimed. □

One surprising consequence of Lambert's theorem is that no hyperbolic triangle can have an area more than π! This would not be strange for a space whose total area was finite, like a sphere, but this is not true of the hyperbolic plane.

5.2 Decomposing via the Trace

Now that some of the oddness of hyperbolic geometry has been illuminated, we shift our focus. Our goal for this section is straightforward: we aim to find a reasonable way of dividing up the orientation-preserving isometries of the hyperbolic plane and hyperbolic space into some number of easily understood families. Our main tool to do this is the trace of a matrix. Given any matrix in $SL(2, \mathbb{R})$ or $SL(2, \mathbb{C})$, we know how to define the trace—can this be done for $PSL(2, \mathbb{R})$ and $PSL(2, \mathbb{C})$? On the face, this is not possible: after all, for any $\phi \in PSL(2, \mathbb{R})$, we know that there exist two different matrices $\pm M$ that correspond to ϕ. However, we might notice that switching between the two matrices only changes the trace by a factor of ± 1—therefore, the square of the trace is perfectly well-defined.

Definition 5.1 Let $\mathbb{F}$ be either $\mathbb{R}$ or $\mathbb{C}$. For any $\phi \in PSL(2, \mathbb{F})$, we define $\mathrm{tr}^2(\phi) = \mathrm{tr}(M)^2$ where $M \in SL(2, \mathbb{F})$ is a matrix that maps to ϕ under the standard group homomorphism $SL(2, \mathbb{F}) \to PSL(2, \mathbb{F})$.

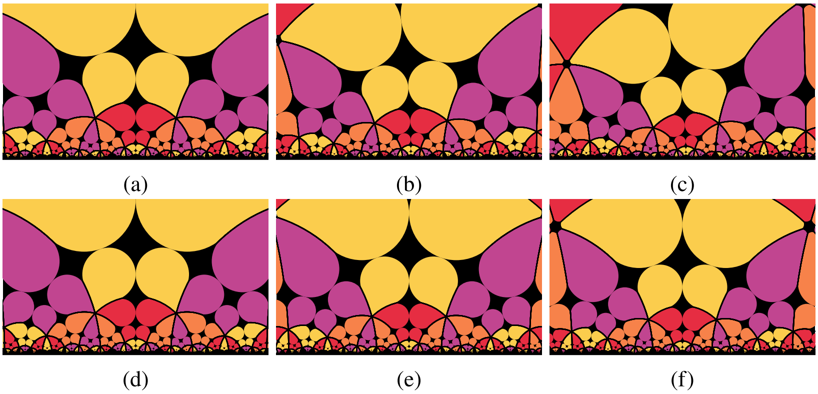


Fig. 5.6 (a), (b), (c) show the effect of a hyperbolic isometry; (d), (e), (f) show the effect of a hyperbolic isometry conjugate to the first, but which is easier to understand.

One of the key properties of the trace is that it is invariant under conjugation.

Definition 5.2 We say that $\phi, \psi \in PSL(2,\mathbb{C})$ are *conjugate* if there exists $\gamma \in PSL(2,\mathbb{C})$ such that $\phi = \gamma \circ \psi \circ \gamma^{-1}$.

Two conjugate isometries are shown in Figure 5.6. Since we know that $\text{tr}(M) = \text{tr}(UMU^{-1})$ for any $U \in GL(2,\mathbb{C}$ and any 2×2 complex matrix M, it is immediate that if ϕ, ψ are conjugate, then $\text{tr}^2(\phi) = \text{tr}^2(\psi)$. Why is conjugation so important? We have seen it crop up countless times in the previous chapters whenever we had to do a change of coordinates—if γ was a transformation of a space, then $\gamma \circ \psi \circ \gamma^{-1}$ was the right way to express ψ in this new setting. So, the square of the trace is some property that persists even under such changes.

Let's start by using the trace to classify elements in $PSL(2,\mathbb{R})$. Note that since the trace of any matrix in $SL(2,\mathbb{R})$ is real, $\text{tr}^2(\phi) \geq 0$ for all $\phi \in PSL(2,\mathbb{R})$.

Definition 5.3 We say that $\phi \in PSL(2,\mathbb{C})$ is

1. *elliptic* if $0 \leq \text{tr}^2(\phi) < 4$,
2. *hyperbolic* if $\text{tr}^2(\phi) > 4$, and
3. *parabolic* if $\text{tr}^2(\phi) = 4$.

Obviously, any element in $PSL(2,\mathbb{R})$ is one of these three, purely by definition. However, this split into three distinct categories is not arbitrary. We first need a lemma.

Lemma 5.2 *Let $\phi \in PSL(2,\mathbb{C})$. There exists $z \in \mathbb{C}P^1$ such that $\phi(z) = z$.*

Remark 5.1 This lemma can be generalized a substantial amount using tools from topology. It is true, for example, that any map $\mathbb{C}P^1 \to \mathbb{C}P^1$ that is differentiable, has a differentiable inverse, and is orientation-preserving, must have at least one fixed

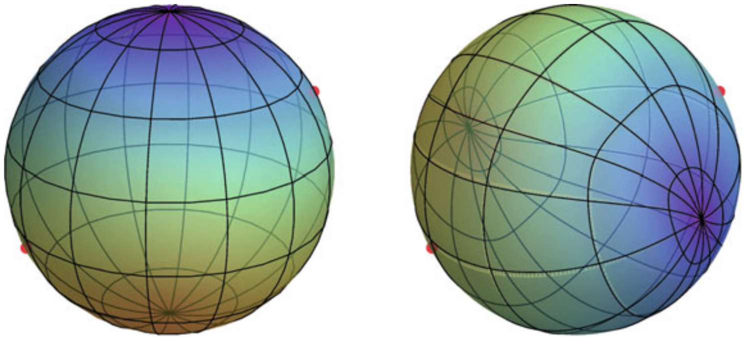


Fig. 5.7 An illustration of a map from the sphere back to itself which is differentiable, has a differentiable inverse, and is orientation-preserving; it is not, however, angle-preserving like any element in $PSL(2, \mathbb{C})$ would be. This map has two fixed points, drawn in red.

point—this is a consequence of the celebrated Brouwer fixed-point theorem. We draw an illustration of such a map in Figure 5.7 (we use the fact that stereographic projection lets us use the sphere and $\mathbb{C}P^1$ interchangeably), but we do not pursue anything so complicated.

Proof Write $\phi(z) = M.z$ for some

$$M = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL(2, \mathbb{C}).$$

If $c = 0$, the claim is obvious: $\phi(\infty) = \infty$. Otherwise, we see that we are trying to solve

$$\phi(z) = \frac{az + b}{cz + d} = z,$$

which is really just a quadratic equation $cz^2 + (d - a)z - b = 0$. Any quadratic equation always has at least one solution in $\mathbb{C}$. □

Theorem 5.2 (Classification of the Orientation-Preserving Isometries of the Hyperbolic Plane)

For any non-identity $\phi \in PSL(2, \mathbb{R})$, exactly one of the following is true.

1. *ϕ is elliptic, it fixes exactly one point in the hyperbolic plane (and none on the boundary), and it is conjugate to some transformation $z \mapsto e^{i\theta}z$.*
2. *ϕ is hyperbolic, it fixes exactly two points on the boundary of the hyperbolic plane (and none in the plane itself), and it is conjugate to some transformation $z \mapsto \lambda^2 z$.*
3. *ϕ is parabolic, it fixes exactly one point on the boundary of the hyperbolic plane (and none in the plane itself), and it is conjugate to some transformation $z \mapsto z + x$.*

Proof By the lemma, we know that ϕ must fix at least one point in the hyperbolic plane plus its boundary. We also know that it can fix at most two points—if it fixes three, then it is the identity, due to the properties of linear fractional transformations. Suppose that ϕ fixes a point z_0 in the Poincaré plane. Extending γ to act on all of $\mathbb{C}P^1$, it is easy to see that ϕ also fixes $\overline{z_0}$—this is because

$$\overline{z_0} = \overline{M.z_0} = \overline{M}.\overline{z_0} = M.\overline{z_0},$$

where $M \in SL(2,\mathbb{R})$ is a matrix such that $\phi(z) = M.z$. But this means that γ cannot possibly fix any more points in either the hyperbolic plane or its boundary. Therefore, we see that we really do have exactly three possibilities: ϕ fixes one point in the plane; ϕ fixes two points on the boundary; γ fixes one point on the boundary. It remains to show that these three cases match with everything else attributed to them.

Suppose that ϕ fixes one point z_0 in the hyperbolic plane. We can find an isometry $\gamma \in PSL(2,\mathbb{C})$ which sends the hyperbolic plane as a whole to $\mathbb{D}^2$ and sends z_0 to 0 in particular. If we define $\tilde{\phi} = \gamma \circ \phi \circ \gamma^{-1}$, then $\tilde{\phi}(0) = 0$. We have already worked out what all such transformations look like in the exercise at the end of Section 4.6: it must be that $\tilde{\phi}(z) = e^{i\theta}z$ for some $0 < \theta < 2\pi$. Of course, ϕ is conjugate to $\tilde{\phi}$ by definition, so we just need to show that ϕ is elliptic. Indeed, $\tilde{\phi}$ corresponds to the matrix

$$\begin{pmatrix} e^{i\theta/2} & 0 \\ 0 & e^{-i\theta/2} \end{pmatrix},$$

so $\mathrm{tr}^2(\phi) = \mathrm{tr}^2(\tilde{\phi}) = 4\cos(\theta/2)^2$. Since $0 < \theta < 2\pi$, we are done.

Suppose that ϕ fixes two points z_0, z_1 on the boundary. We can find an isometry $\gamma \in PSL(2,\mathbb{C})$ which sends the hyperbolic plane as a whole to $\mathbb{H}^2$ and sends $z_0 \mapsto 0$, $z_1 \mapsto \infty$ in particular. If we define $\tilde{\phi} = \gamma \circ \phi \circ \gamma^{-1}$, then $\tilde{\phi}(0) = 0$ and $\tilde{\phi}(\infty) = \infty$. Since $\tilde{\phi}(\infty) = \infty$, $\tilde{\phi}(z) = az + b$; since $\tilde{\phi}(0) = 0$, $b = 0$. Such a transformation preserves $\mathbb{H}^2$ if and only if $a > 0$, hence $\tilde{\phi}(z) = \lambda^2 z$ for some $\lambda \in \mathbb{R}\backslash\{\pm 1\}$, and can be represented by a matrix

$$\begin{pmatrix} \lambda & 0 \\ 0 & \lambda^{-1} \end{pmatrix} \in SL(2,\mathbb{R}),$$

whence $\mathrm{tr}^2(\phi) = \mathrm{tr}^2(\tilde{\phi}) = (\lambda + \lambda^{-1})^2$. The only thing missing is proving that this must necessarily be larger than 4, which is a simple calculus argument. (See Exercise 5.3.1.) Therefore, ϕ is hyperbolic.

If ϕ fixes one point z_0 on the boundary, then by the same argument as above, it must be conjugate to $\tilde{\phi}(z) = az + b$, with $a, b \in \mathbb{R}$. If $z \neq 1$, then this transformation will also fix $b/(1-a)$, which contradicts the fact that it only fixes one point. Therefore, $\tilde{\phi}$ is represented by some matrix

$$\begin{pmatrix} 1 & b \\ 0 & 1 \end{pmatrix}$$

and $\mathrm{tr}^2(\phi) = \mathrm{tr}^2(\tilde{\phi}) = 4$, so ϕ is parabolic. □

This is already thought-provoking. We might analyze each of these three basic types of elements in $PSL(2, \mathbb{R})$ and try to find their Euclidean analogs—for instance, it is reasonably clear that the elliptic elements are nothing more than rotations in hyperbolic space. Indeed, we will do precisely this in the next sections of this chapter. But, first, it is a good idea to see if there is a similar decomposition for $PSL(2, \mathbb{C})$.

There is, and it is very similar to the one for $PSL(2, \mathbb{R})$—we just need to add one other kind of transformation.

Definition 5.4 We say that $\phi \in PSL(2, \mathbb{C})$ is *loxodromic* if $\text{tr}^2(\phi) \notin [0, \infty)$.

Remark 5.2 Some authors instead define loxodromic transformations as those where the square trace is not in $[0, 4]$, which includes the hyperbolic elements.

Theorem 5.3 (Classification of the Orientation-Preserving Isometries of Hyperbolic Space)

For any non-identity $\phi \in PSL(2, \mathbb{C})$, exactly one of the following is true.

1. *ϕ is elliptic, it fixes a hyperbolic line in $\mathbb{H}^3$ (including the endpoints on the boundary), and it is conjugate to some transformation $z \mapsto e^{i\theta} z$ with $0 < \theta < 2\pi$.*
2. *ϕ is hyperbolic, it fixes exactly two points on the boundary of hyperbolic space (and none in $\mathbb{H}^3$), and it is conjugate to some transformation $z \mapsto \lambda^2 z$ with $\lambda > 0$ and $\lambda \neq 1$.*
3. *ϕ is parabolic, it fixes exactly one point on the boundary of hyperbolic space (and none in $\mathbb{H}^3$), and it is conjugate to some transformation $z \mapsto z + b$ for some $b \in \mathbb{C}$.*
4. *ϕ is loxodromic, it fixes exactly two points on the boundary of hyperbolic space (and none in $\mathbb{H}^3$), and it is conjugate to some transformation $z \mapsto \lambda^2 e^{i\theta} z$ for some $1 \neq \lambda > 0$ and $0 < \theta < 2\pi$.*

Proof We know that ϕ has to fix at least one point in $\mathbb{C}P^1$—without loss of generality, (since we only care about ϕ up to conjugation) we may assume that point is ∞. Thus, $\phi(z) = az + b$. If $a = 1$, then $\phi(z) = z + b$ is obviously parabolic and only fixes the point ∞ in both $\mathbb{H}^3$ and $\mathbb{C}P^1$.

Therefore, we may assume that $a \neq 1$. In that case, ϕ has to have a second fixed point in $\mathbb{C}P^1$, namely $b/(1 - a)$. Again, since we only care about ϕ up to conjugation, we may assume that this fixed point is 0, so indeed $\phi(z) = az$ for some $a \neq 1$. Exactly one of the following must be true.

1. $a = e^{i\theta}$ for some $0 < \theta < 2\pi$.
2. $a = \lambda^2$ for some $\lambda > 0$ and $\lambda > 0$.
3. $a = \lambda^2 e^{i\theta}$ for some $1 \neq \lambda > 0$ and $0 < \theta < 2\pi$.

In the first case, ϕ is elliptic—it corresponds to the matrix

$$\begin{pmatrix} e^{i\theta/2} & 0 \\ 0 & e^{-i\theta/2} \end{pmatrix} \in SL(2, \mathbb{C}).$$

To find the fixed points in $\mathbb{H}^3$, we just choose any $\rho = z + tj$ where $z \in \mathbb{C}$ and $t > 0$, and compute

$$\begin{pmatrix} e^{i\theta/2} & 0 \\ 0 & e^{-i\theta/2} \end{pmatrix} .(z + tj) = e^{i\theta/2}(z + tj)e^{i\theta/2} = e^{i\theta}z + tj,$$

which is equal to ρ if and only if $z = 0$—therefore, ϕ fixes the hyperbolic line $z = 0$ in $\mathbb{H}^3$.

In the second case, ϕ is hyperbolic—it corresponds to the matrix

$$\begin{pmatrix} \lambda & 0 \\ 0 & \lambda^{-1} \end{pmatrix} \in SL(2, \mathbb{R}).$$

It's clear that ϕ cannot have any fixed points in $\mathbb{H}^3$ since it will send a point with j-coordinate t to a point with j-coordinate $\lambda^2 t$.

In the final case, we see that ϕ corresponds to a matrix

$$\begin{pmatrix} \lambda e^{i\theta/2} & 0 \\ 0 & \lambda^{-1} e^{-i\theta/2} \end{pmatrix} \in SL(2, \mathbb{C}).$$

If we write $w = \lambda e^{i\theta/2}$, then $\text{tr}^2(\phi) = w + 1/w$. Let's suppose that $\text{tr}^2(\phi) = r \in [0, \infty)$. Then $w + 1/w = r$, which implies that $w^2 - rw + 1 = 0$, so

$$w = \frac{r \pm \sqrt{r^2 - 4}}{2}.$$

If $r \geq 4$, this is a positive real number, which w is not by assumption. On the other hand, if $0 \leq r < 4$, then $\sqrt{r^2 - 4}$ is pure imaginary, which means that

$$|w|^2 = \frac{r + \sqrt{r^2 - 4}}{2} \cdot \frac{r - \sqrt{r^2 - 4}}{2} = 1,$$

which contradicts the fact that w is not on the unit circle. Ergo, we conclude that $\text{tr}^2(\phi) \notin [0, \infty)$, hence ϕ is loxodromic. That it doesn't fix any points in $\mathbb{H}^3$ can be shown as in the hyperbolic case: it will map any point with j-coordinate t to a point with j-coordinate $\lambda^2 t$. □

5.3 Elliptic Elements

We shall now go through each of the four types of elements in $PSL(2, \mathbb{C})$ in turn, starting with the simplest—the elliptic elements. We know by Theorems 5.2 and 5.3 that these are conjugate to transforms $z \mapsto e^{i\theta}z$, so we might suppose that the most natural way to think about them is as rotations of hyperbolic space. This is indeed the usual perspective, and many properties are shared with the more familiar Euclidean setting.

For example, suppose that we choose some elliptic $\varphi \in PSL(2, \mathbb{R})$. It has a unique fixed point $w \in \mathbb{H}^2$, so consider what happens to a hyperbolic circle C with center w and radius r. Since φ is an isometry, $\varphi(C)$ must also be a hyperbolic circle with radius r. In fact, since $\varphi(w) = w$, we see that actually $\varphi(C) = C$. Therefore, we have an infinite family of concentric circles C such that φ moves each of these circles back to itself. Some examples are shown in Figure 5.8. Moreover, we can describe precisely what φ does to the points on each of these circles (Figure 5.9).

Theorem 5.4 *Let $\varphi \in PSL(2, \mathbb{R})$ be elliptic, with unique fixed point $w \in \mathbb{H}^2$. Let C be a hyperbolic circle centered at w with radius R. There exists a constant $k \in (0, 2R]$ such that one of the following is true.*

1. *For each point $z \in C$, $\varphi(z)$ is the unique point that is counterclockwise from z on C and hyperbolic distance k from z.*
2. *For each point $z \in C$, $\varphi(z)$ is the unique point that is clockwise from z on C and hyperbolic distance k from z.*

Proof Note that everything here is expressed in ways that are invariant under isometries, so it is actually sufficient to prove what we want for $\varphi(z) = e^{i\theta} z$ acting on $\mathbb{D}^2$, where C is then just a circle centered at 0 with some Euclidean radius $0 < r < 1$. Since φ is now just a Euclidean rotation, it is going to take move all points on C either clockwise or counterclockwise. The farthest it could possibly move them is $2R$, which is what it would be for antipodal points on the circle. It remains to prove that it moves all points by the same amount, as measured in hyperbolic space. An arbitrary element of C can be written as $re^{i\alpha}$ and so

$$\begin{aligned} d_{\text{hyper}}(z, \varphi(z)) &= d_{\text{hyper}}\left(re^{i\alpha}, re^{i(\alpha+\theta)}\right) \\ &= d_{\text{hyper}}(r, re^{i\theta}), \end{aligned}$$

which we note depends only on φ and C, but not z. □

The situation for an arbitrary element $\varphi \in PSL(2, \mathbb{C})$ is not very different, particularly as we know that any such element must be conjugate to an element in $PSL(2, \mathbb{R})$, as they are both conjugate to elements $z \mapsto e^{i\theta} z$. We may nevertheless want to understand what an elliptic element does to the whole of $\mathbb{C}P^1$ or even $\mathbb{H}^3$, rather than merely $\mathbb{H}^2$ or $\mathbb{D}^2$. In terms of the action on $\mathbb{C}P^1$, we can reason thus: $z \mapsto e^{i\theta} z$ fixes 0 and ∞ and all other elements are rotated around in circular orbits. Since $PSL(2, \mathbb{C})$ preserves generalized circles, we see that the corresponding general statement is this: if ϕ is elliptic, then it has some fixed points $z_1, z_2 \in \mathbb{C}P^1$ and the rest of $\mathbb{C}P^1$ is a union of an infinite family of generalized circles that are each sent back to themselves by ϕ. On any particular circle, ϕ moves points in an orbit counterclockwise (or clockwise, depending on how you look at it). This is illustrated in Figure 5.10.

As for points in $\mathbb{H}^3$, we know that for any elliptic element ϕ, there exists a hyperbolic line in $\mathbb{H}^3$ that is fixed by ϕ. In some sense, every other point is rotated around this line. We can see this in one of two ways. First, we can note that ϕ is conjugate to

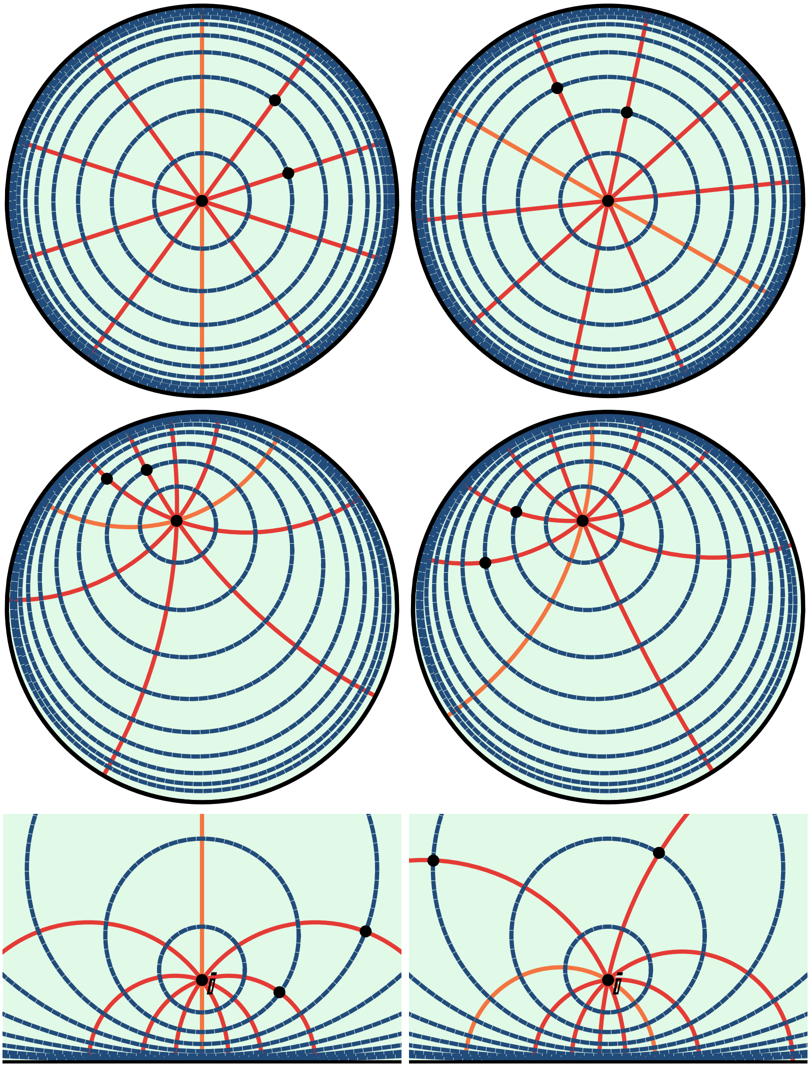


Fig. 5.8 A rotation by an angle of $\pi/3$ seen in three different coordinate frames: in the first row, it is seen in the Poincaré disk, with 0 as the fixed point; in the second row, it is still in the Poincaré disk, but with a different fixed point; in the last row, it is seen in the Poincaré half-plane, with i as the fixed point. In each illustration, the dashed circles are all concentric and are moved back to themselves by the rotation.

some transformation $z \mapsto e^{i\theta}$. We previously saw that $\phi(z + tj) = e^{i\theta}z + tj$ for any $z \in \mathbb{C}$ and $t > 0$, so this is a rotation by an angle θ around the j-axis. Conjugation will replace the j-axis with some other hyperbolic line, but the basic principle that ϕ moves points in circular orbits around the line will remain true.

There is a different way to get the same result, which is summarized in the following theorem.

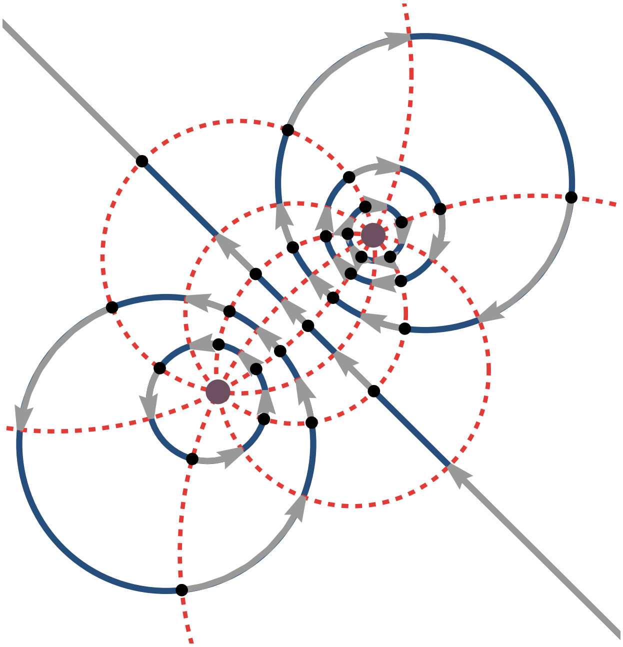


Fig. 5.9 An illustration of the dynamics in $\mathbb{C}P^1$ of an elliptic element conjugate to $z \mapsto e^{2\pi i/5}z$. The fixed points are drawn in purple; all other points are moved along circular paths, drawn in blue. The arrows show precisely where each point is moved.

Theorem 5.5 *Let $\phi \in PSL(2, \mathbb{C})$ be an elliptic element. Let l be the hyperbolic line in $\mathbb{H}^3$ that is fixed by ϕ.*

1. *For any $\rho \in l$, there exists a unique hyperbolic plane P_ρ that is orthogonal to l at ρ.*
2. *$\mathbb{H}^3$ is the disjoint union of the planes P_ρ. (That is, $P_\rho \cap P_{\rho'} = \emptyset$ if $\rho \neq \rho'$, but the union of all of them is $\mathbb{H}^3$.)*
3. *$\phi(P_\rho) = P_\rho$ for all $\rho \in l$. Restricted to P_ρ, ϕ is elliptic, with unique fixed point ρ.*

Proof I leave the proofs of the first two assertions to the reader (see Exercise 5.2.15). For the third part, choose any such plane P_ρ: $\phi(P_\rho)$ must also be a hyperbolic plane that passes through $\phi(\rho) = \rho$ and orthogonal to the fixed hyperbolic line, which is just to say that $\phi(P_\rho) = P_\rho$. Since P_ρ is just a copy of $\mathbb{H}^2$ inside of $\mathbb{H}^3$, we see that we can think of ϕ as an isometry of P_ρ. Specifically, it is an isometry with one fixed point, ρ, so it is elliptic. □

Note that each of the planes P_ρ is isometrically isomorphic to $\mathbb{H}^2$, so Theorem 5.4 applies—in particular, it means that ϕ moves points inside each of these planes in circular orbits, as in Figure 5.10.

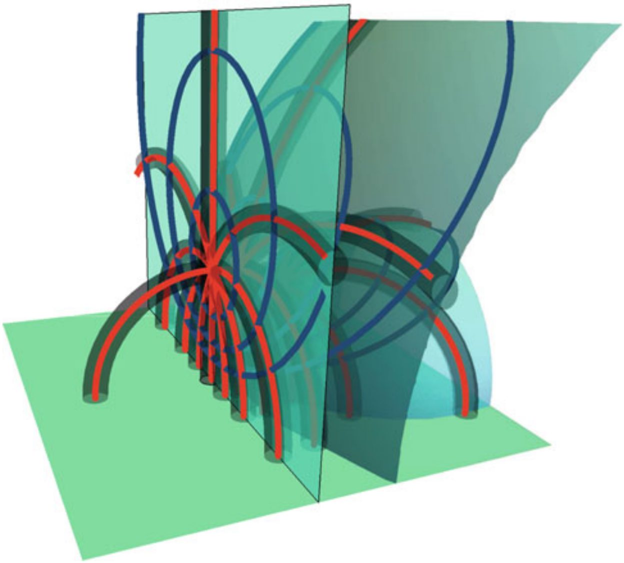


Fig. 5.10 An illustration of the dynamics on $\mathbb{H}^3$ of an elliptic element—a hyperbolic line is shown which is fixed by this element, together with planes orthogonal to said line. Inside each plane, points are moved in circular orbits, drawn in blue.

▶ **Example** *Show that $\phi(z) = -z^{-1}$ is elliptic. Find its fixed points in $\mathbb{C}P^1$ and a $\theta \in \mathbb{R}$ such that it is conjugate $z \mapsto e^{i\theta}z$.*
The matrix that corresponds to ϕ is

$$\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \in SL(2, \mathbb{C}),$$

which has trace 0—therefore, ϕ is elliptic, possessing two fixed points in $\mathbb{C}P^1$. By inspection, $\pm i$ are fixed, so they are the unique fixed points. To get this transformation to be conjugate to $z \mapsto e^{i\theta}z$, we should move these two points to 0 and ∞. Thankfully, we already know, due to the isometric isomorphism between $\mathbb{H}^2$ and $\mathbb{D}^2$ that the map

$$\psi(z) = \frac{iz+1}{z+i} = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{i}{\sqrt{2}} \\ -\frac{i}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}.z$$

does exactly that. So, we simply compute

$$\begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{i}{\sqrt{2}} \\ -\frac{i}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}\begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{i}{\sqrt{2}} \\ -\frac{i}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}^{-1} = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}$$

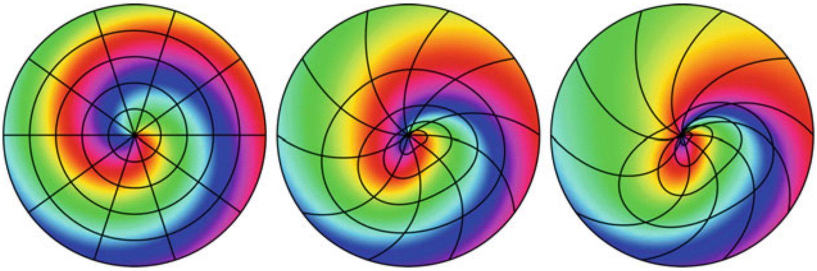


Fig. 5.11 An illustration of a continuous transformation $F : \mathbb{D}^2 \to \mathbb{D}^2$ such that 0 is an attractive fixed point. From left to right, we have the original image, its image under F, and its image under F^2.

to conclude that ϕ is conjugate to $z \mapsto -z = e^{i\pi} z$. Thus, we see it for what it is: it is a half-turn around i in the upper half-plane.

5.4 Hyperbolic Elements

We know that any hyperbolic element is conjugate to $z \mapsto \lambda^2 z$ for some $\lambda > 0$. What is the right geometric interpretation of such a transformation? We might note that this is an isometry of $\mathbb{H}^2$ and that it sends the line $x = 0$ back to itself. Indeed, it isn't hard to see that it has to move all the points on this line in a single direction, either toward 0 or ∞—in other words, one of these points is attractive.

Definition 5.5 Let X be some subset of $\mathbb{R}^n$ and consider a continuous function $F : X \to X$. We say that $x_0 \in X$ is an *attractive fixed point* of F if for all $x \in X$, $\lim_{n\to\infty} F^n(x) = x_0$.

Remark 5.3 Here, F^n should be understood to mean F composed with itself n times, not F raised to some power (whatever that would even mean).

Remark 5.4 See Figure 5.11 for an illustration of a map $\mathbb{D}^2 \to \mathbb{D}^2$ with an attractive fixed point.

We can actually simplify our life a bit and assume that ∞ is the attractive fixed point. This is because

$$\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \lambda & 0 \\ 0 & \lambda^{-1} \end{pmatrix} \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}^{-1} = \begin{pmatrix} \lambda^{-1} & 0 \\ 0 & \lambda \end{pmatrix},$$

so we are always free to replace λ with λ^{-1}. This means that we can always assume that $\lambda > 1$, in which case ∞ is the attractive fixed point. More generally, any hyperbolic element necessarily has an attractive fixed point.

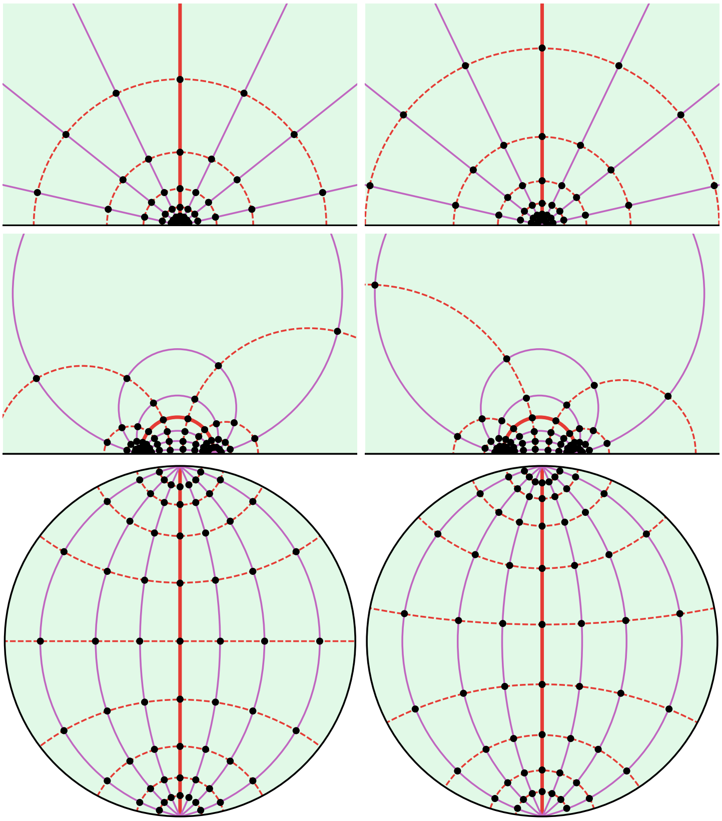


Fig. 5.12 A translation by a hyperbolic distance of about 0.19, seen in three different coordinate frames: in the first row, it is seen in the Poincaré half-plane, with 0, ∞ as the fixed points; in the second row, the fixed points are different; in the last row, it is seen in the Poincaré disk, with $-i, i$ as the fixed points. In each illustration, the solid red curve is the line connecting the fixed points.

Theorem 5.6 *Let $\varphi \in PSL(2,\mathbb{R})$ be hyperbolic. One of its fixed points is attractive—call it z_1 and the other z_0. For every $z \in \mathbb{H}^2$, $\varphi(z)$ lies on the unique generalized circle passing through z_0, z_1, and z. Restricted to any such generalized circle, φ moves points away from z_0 and toward z_1. Exactly one of these generalized circles is a hyperbolic line; restricted to said line, φ translates each point by a constant distance that depends only on $tr(\varphi)$.*

Proof If z_1 is an attractive fixed point of φ, then $\psi(z_1)$ will be an attractive fixed point of $\psi \circ \varphi \circ \psi^{-1}$. Consequently, the fact that we know that φ is conjugate to $z \mapsto \lambda^2 z$ for some $\lambda > 1$ immediately implies that it has an attractive fixed point. Indeed, we know that φ has two fixed points, but it cannot have two attractive fixed points. So, let $z_0, z_1 \in \partial\mathbb{H}^2$ be those two fixed points, with z_1 the attractive one. Any point $z \in \mathbb{H}^2$ will lie on some unique generalized circle C through z, z_0, z_1. The corresponding statement for $z \mapsto \lambda^2 z$ is that every point in $\mathbb{H}^2$ lies on some line

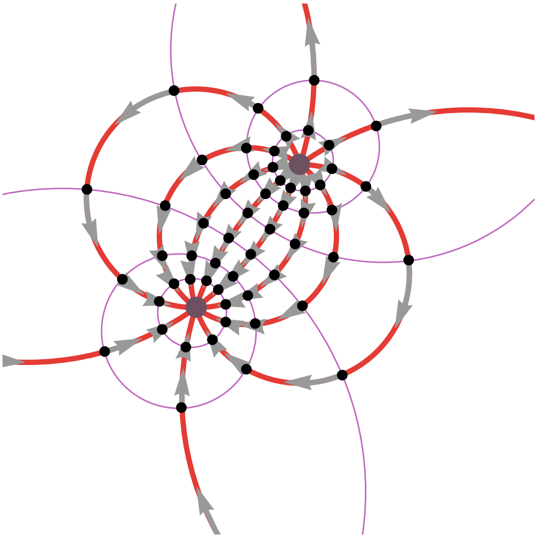


Fig. 5.13 An illustration of the dynamics in $\mathbb{C}P^1$ of an hyperbolic element conjugate to $z \mapsto \sqrt{2}z$. The fixed points are drawn in purple; all other points are moved along circular paths, drawn in red. The arrows show precisely where each point is moved.

through the origin, and $z \mapsto \lambda^2 z$ sends all such lines back to themselves. Therefore, φ must do the same to the generalized circles C. Given a point $z \in C$, where will φ send it? It must be sent closer to z_1; the corresponding picture for $z \mapsto \lambda^2 z$ is that it sends all elements in $\mathbb{H}^2$ closer to ∞.

Out of all these generalized circles through z_0 and z_1, only one of them is orthogonal to $\partial\mathbb{H}^2$ and is, therefore, a line—for $z \mapsto \lambda^2 z$, this is the line $x = 0$. Let $z \in \mathbb{H}^2$ be any point on this line. Then

$$d_{\text{hyper}}(z, \varphi(z)) = d_{\text{hyper}}\left(ti, \lambda^2 ti\right) = \ln\left|\frac{\lambda^2 ti}{ti}\right| = 2\ln(\lambda).$$

Notice that $\text{tr}(\varphi) = \lambda + 1/\lambda$, so in fact $d_{\text{hyper}}(z, \varphi(z))$ is some constant that only depends on $\text{tr}(\varphi)$ and not on the choice of $z \in \mathbb{H}^2$ on the line between the two fixed points. □

What is the right Euclidean analog to this hyperbolic isometry? No description is going to be a perfect match, but we might note that restricted to the unique line between the two fixed points, a hyperbolic element is just a translation. Indeed, such transformations are typically called (hyperbolic) translations. Off the unique translation line, hyperbolic translations are a little weird in that they move points along paths that are not lines at all, but always away from one fixed point toward the other fixed point, as can be seen from Figure 5.12.

It is clear how to extend this action to the whole of $\mathbb{C}P^1$: a hyperbolic element has two fixed points in $\mathbb{C}P^1$, one of which is an attractive fixed point; every element

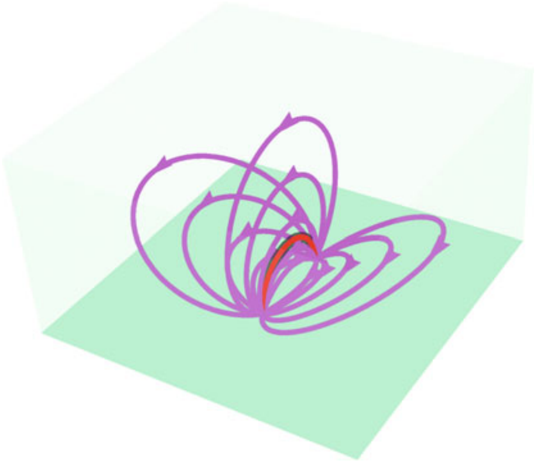


Fig. 5.14 An illustration of the dynamics on $\mathbb{H}^3$ of a hyperbolic element—the line connecting the two fixed points is drawn in red. Other paths connecting those two points and sent back to themselves by the transformation are drawn in purple. Arrows indicate the direction in which the transformation moves points along these curves.

other than the two fixed ones is moved along generalized circles through the two fixed points, toward the attractive fixed point. This is shown in Figure 5.13. We can easily extend to $\mathbb{H}^3$ as well, as in Figure 5.14.

Theorem 5.7 *Let $\varphi \in PSL(2, \mathbb{C})$ be hyperbolic. Let $z_0, z_1 \in \partial\mathbb{H}^3$ be the fixed points of φ, where z_1 is the attractive fixed point. Let l be the line between them. For every point $\rho \in \mathbb{H}^3$, φ sends the generalized circle through z_0, z_1, ρ back to itself, and moving ρ away from z_0 and toward z_1. Restricted to l, φ is a translation: it moves every point away from z_0 and toward z_1 by a fixed distance that depends only on $tr(\varphi)$.*

Proof I leave this proof to the reader. (See Exercise 5.2.16.) □

5.5 Parabolic Elements

The astute reader might have found it surprising that we labeled hyperbolic elements as analogs of Euclidean translations where there is seemingly a more natural candidate: the transformations conjugate to $z \mapsto z + z_0$, which is to say the parabolic elements. After all, these elements *are* Euclidean translations. Counter-intuitively, parabolic elements are absolutely nothing like Euclidean translations—indeed, there isn't a Euclidean analog for them at all. The reason for this is that while elliptic elements can be characterized as those which fix a point and sent circles/spheres centered at that point back on themselves, and hyperbolic elements can be charac-

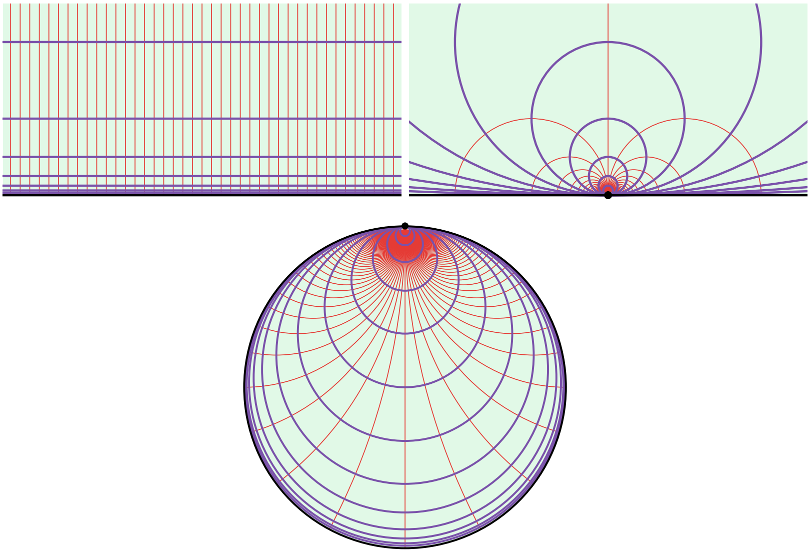


Fig. 5.15 Families of horocycles in the Poincaré half-plane and disk.

terized as those that map a particular line back on itself, parabolic elements preserve an entirely different kind of curve/surface.

Definition 5.6 Let z be a point on the boundary of the hyperbolic plane. A *horocycle* is a curve that is orthogonal to every line in the plane that passes through z.

Similarly, if $z \in \partial\mathbb{H}^3$, then a *horosphere* is a surface that is orthogonal to all lines in $\mathbb{H}^3$ that pass through z.

Some examples of horocycles are drawn in Figure 5.15. Intuitively, they are like circles "at infinity." Indeed, if one were to replace $z \in \partial\mathbb{H}^2$ with $z \in \mathbb{H}^2$, then we would exactly be describing a circle. (See Exercise 5.2.1.) As depicted in Figure 5.16, they can be obtained as "limits" of hyperbolic circles as well. They can also be characterized in a different way that is easier to visualize.

Theorem 5.8 (Characterization of Horocycles) *Horocycles are exactly the generalized circles that are tangent to the boundary of the hyperbolic plane.*

Proof Suppose that we are working in the Poincaré half-plane model and that we take $z = \infty$. What are the hyperbolic lines that pass through this point? Well, these are just all of the vertical lines. What are the curves that are orthogonal to all vertical lines? Horizontal lines! And a horizontal line is nothing more than a generalized circle that is tangent at infinity to the real line—i.e., the boundary. However, we know that all of our isometries preserve both angles and generalized circles. Furthermore, we know that we can always find an isometry that will take ∞ to any other point on the boundary, in either the Poincaré half-plane or the disk. Therefore, all horocycles are generalized circles that are tangent to the boundary of hyperbolic space. Conversely,

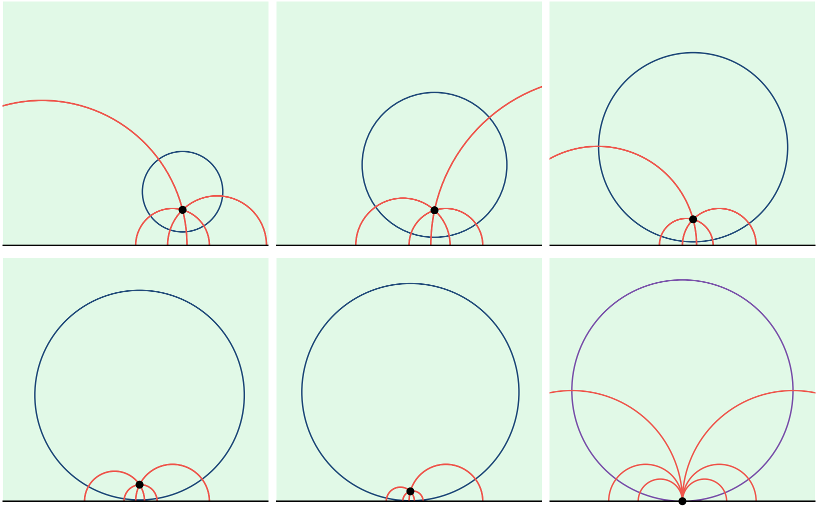


Fig. 5.16 A horocycle envisioned as the limit of a family of hyperbolic circles.

for any generalized circle that is tangent to the boundary of hyperbolic space, we can use an isometry to move that tangency point to ∞, at which point this circle becomes a horizontal line—i.e., a horocycle. □

Theorem 5.9 (Characterization of Horospheres) *Horospheres are exactly the generalized spheres that are tangent to the boundary of hyperbolic space.*

Proof The idea is the same as before: we can reduce to the case where the chosen point is $z = \infty$. The lines that pass through this point are simply the vertical lines, hence the surfaces that are orthogonal to such lines are the horizontal planes. The horizontal planes are exactly those generalized spheres that are tangent to $\partial\mathbb{H}^3$ at ∞. □

Here's the kicker: parabolic elements are exactly those isometries that send horocycles/horospheres back on themselves.

Theorem 5.10 *Choose any $\varphi \in PSL(2, \mathbb{R})$. The following properties are equivalent.*

1. *φ is parabolic.*
2. *There exists a horocycle H such that $\varphi(H) = H$.*
3. *There exists $z \in \partial\mathbb{H}^2$ such that for all horocycles H tangent to the boundary at z, $\varphi(H) = H$.*

Proof The third condition implies the second. We can show that the second condition also implies the first. This is because if H is a horocycle, then it is tangent to $\partial\mathbb{H}^2$ at some one point z, and since φ preserves $\partial\mathbb{H}^2$ and tangency, the fact that $\varphi(H) = H$ implies $\varphi(z) = z$. Therefore, φ has at least one fixed point on the boundary—can it have more? If it does, then it is hyperbolic, and there is some line passing through z

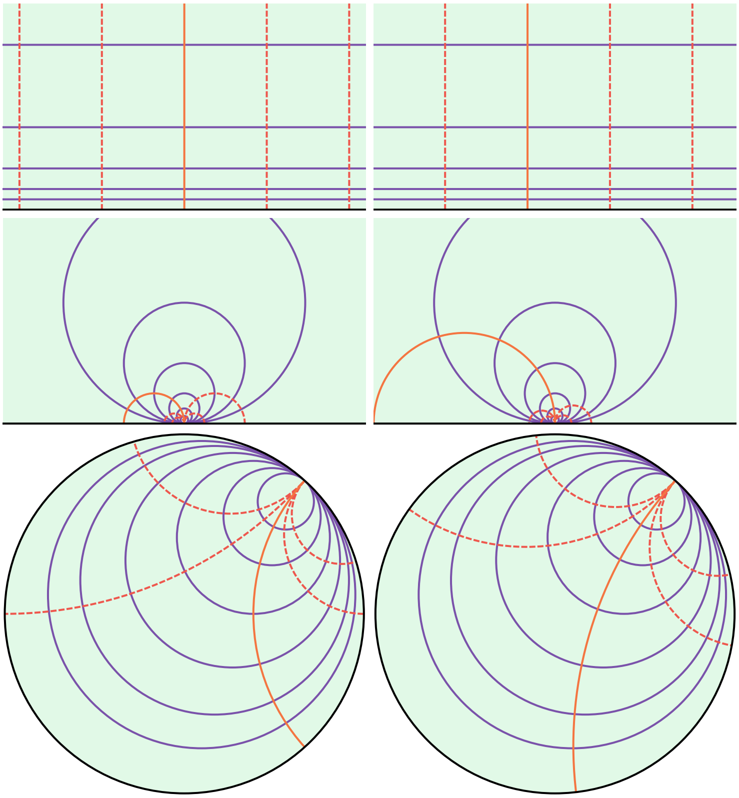


Fig. 5.17 Parabolic elements conjugate to $z \mapsto z - 1/3$, seen in three different coordinate frames: in the first row, it is seen in the Poincaré half-plane, with ∞ as the fixed point; in the second row, the fixed point is $3/2$; in the last row, it is seen in the Poincaré disk. In each illustration, horocycles tangent to the fixed point are drawn in purple.

along which φ simply translates points. But this line has a unique intersection with H, so this contradicts the fact that $\varphi(H) = H$. Therefore, φ has only one fixed point—it is parabolic.

To finish the proof of the theorem, it shall suffice to prove that if φ is parabolic, then there is some point $z \in \partial\mathbb{H}^2$ such that φ preserves all horocycles tangent at that point. Since φ is parabolic, we may assume without loss of generality (because we can always conjugate everything if need be) that $\varphi(z) = z + z_0$, which certainly preserves all the horocycles that are tangent at ∞, as these are just horizontal (Euclidean) lines. □

This action of parabolic elements on families of tangent horocycles can be seen in Figure 5.17. The same is true in $\mathbb{H}^3$ as well—we simply have to replace tangent families of horocycles with tangent families of horospheres.

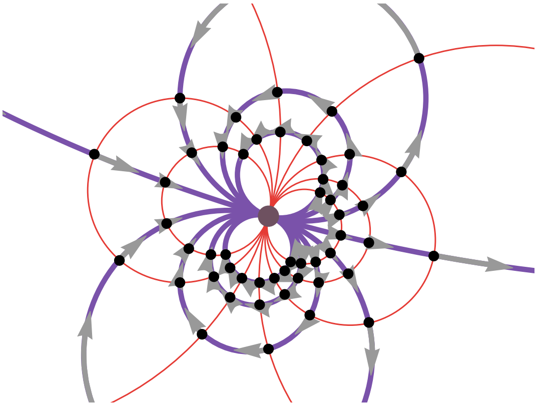


Fig. 5.18 An illustration of the dynamics in $\mathbb{C}P^1$ of a parabolic element conjugate to $z \mapsto z + 1/2$. The fixed point is drawn in purple, as is the family of generalized circles tangent at that point which is preserved by the action of this element. Arrows show the direction of the flow.

Theorem 5.11 *Choose any $\varphi \in PSL(2, \mathbb{C})$. The following properties are equivalent.*

1. *φ is parabolic.*
2. *There exists horosphere H such that $\varphi(H) = H$.*
3. *There exists $z \in \partial\mathbb{H}^3$ such that for all horospheres H tangent to the boundary at z, $\varphi(H) = H$.*

Proof I leave this proof as an exercise to the reader. (See Exercise 5.2.17.) □

The wonderful thing is that these results allow us to immediately understand the dynamics of parabolic elements on hyperbolic space.

Theorem 5.12 *Let $\varphi \in PSL(2, \mathbb{R})$ be parabolic. Let $z_0 \in \partial\mathbb{H}^2$ be its unique fixed point.*

1. *For any $w \in \mathbb{H}^2$, there exists a unique horocycle that passes through both w and z_0.*
2. *Restricted to any horocycle H tangent to the boundary at z_0, φ moves points on H by a fixed distance either clockwise or counterclockwise.*

Proof We can reduce to the case where $z_0 = \infty$ and $\varphi(z) = z + b$ for some $b \in \mathbb{R}$. In this case, $y = \Im(w)$ will be the unique horocycle that passes through z and w and φ either moves all points on this curve to the left, or to the right. Furthermore,

$$d_{\text{hyper}}(w, \varphi(z)) = d_{\text{hyper}}(w, w + b) = d_{\text{hyper}}(\Im(w)i, \Im(w)i + b),$$

which only depends on the particular horocycle H, and not $w \in H$. □

It is easy to extend this to all of $\mathbb{C}P^1$: if $\varphi \in PSL(2, \mathbb{C})$ is parabolic, then it has some fixed point $z_0 \in \mathbb{C}P^1$, and there exists some generalized circle C_0 passing through z_0 such that

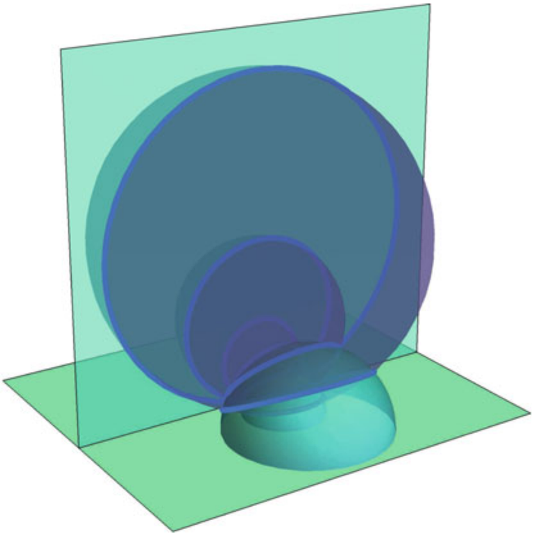


Fig. 5.19 An illustration of the dynamics on $\mathbb{H}^3$ of a parabolic element.

1. $\varphi(C_0) = C_0$,
2. if C is a generalized circle tangent to C_0 at z_0, then $\varphi(C) = C$, and
3. if φ is restricted to any such generalized circle C, then it moves all points along C, away from z_0 and back toward z_0.

This flow on $\mathbb{C}P^1$ is depicted in Figure 5.18. Why is this so? Well, any parabolic element is conjugate to $z \mapsto z + z_0$, which preserves the family of Euclidean lines parallel to $z = z_0 t$—this is precisely a family of generalized circles all tangent to the fixed point of the parabolic element. Since such things are preserved by conjugation, this must be true of all parabolic elements. Of course, once we understand what happens on $\mathbb{C}P^1$, we can understand what happens on $\mathbb{H}^3$. The associated drawing is Figure 5.19.

Theorem 5.13 *Let $\varphi \in PSL(2, \mathbb{C})$ be parabolic, with fixed point $z \in \partial\mathbb{H}^3$. There exists a hyperbolic plane P passing through z such that $\varphi(P) = P$. Furthermore:*

1. *If P' is a plane tangent to P at z, then $\varphi(P') = P'$, and the restriction of φ to P' is parabolic.*
2. *For all $p \in \mathbb{H}^3$, there exists a unique plane P' tangent to P at z that passes through p.*
3. *If H is a horosphere passing through z and P' is a plane tangent to P at z, then their intersection is a horocycle in P'.*

Proof Without loss of generality, we may assume that $\varphi(\rho) = \rho + b$ for some $b \in \mathbb{C}$ and so the fixed point is ∞. Choose any point $z_0 \in \mathbb{C}$—then the Euclidean plane P_{z_0} passing through the points z_0, $z_0 + b$, $z_0 + j$ will be a hyperbolic plane passing through z_0 such that $\varphi(P_{z_0}) = P_{z_0}$. Hyperbolic planes tangent to P_{z_0} at ∞

will exactly be the planes P_{z_1} for other choices of $z_1 \in \mathbb{C}$. From this, it is easy to see that any point $\rho \in \mathbb{H}^3$ is contained in exactly one such plane—specifically, they will be contained in the plane $P_{\rho-\pi_j(\rho)j}$. Restricted to any such plane, φ is certainly parabolic—it has exactly one fixed point, ∞. The horospheres passing through ∞ are just horizontal planes—the intersection of a vertical and a horizontal plane is a horizontal line in one of the planes P_{z_1}, which is just a horocycle. □

5.6 Loxodromic Elements

Loxodromic elements are quite interesting in that they are the only type of transformation that does not occur in $PSL(2, \mathbb{R})$ but does appear in $PSL(2, \mathbb{C})$. On the other hand, their action on $\mathbb{H}^3$ is not hard to understand: they are just a combination of a translation and a rotation along a common line.

Theorem 5.14 *Let $\varphi \in PSL(2, \mathbb{C})$ be loxodromic, with fixed points $z_0, z_1 \in \partial\mathbb{H}^3$. Let l be the unique line that intersects the boundary at z_0 and z_1. There exists a unique hyperbolic element τ and an elliptic element ϕ such that $\varphi = \tau \circ \phi$, τ is a translation along l, and ϕ is a rotation around l.*

Proof Without loss of generality, $z_0 = 0$, $z_1 = \infty$, and $\varphi(z) = re^{i\theta}z$ for some $1 \neq r > 0$ and $0 < \theta < 2\pi$. If τ is a translation along l, then it is of the form $z \mapsto r'z$ for some $1 \neq r' > 0$; if ϕ is a rotation around l, then it is of the form $z \mapsto e^{i\theta'}z$ for some $0 < \theta' < 2\pi$. Then $(\tau \circ \phi)(z) = r'e^{i\theta'}z$, which is equal to φ if and only if $r' = r, \theta' = \theta$. □

A consequence of this is that loxodromic elements move points in $\mathbb{H}^3$ along infinite spiral paths. We can also see this as follows: let $\varphi \in PSL(2, \mathbb{C})$ be loxodromic, with fixed points $z_0, z_1 \in \partial\mathbb{H}^3$. Choose any $\psi \in PSL(2, \mathbb{C})$ such that $\psi(0) = z_0$ and $\psi(\infty) = z_1$. Then $(\psi^{-1} \circ \varphi \circ \psi)(z) = L.z$ for some

$$\begin{pmatrix} re^{i\theta} & 0 \\ 0 & \frac{1}{r}e^{-i\theta} \end{pmatrix} \in SL(2, \mathbb{C}).$$

Now, for any $t \in \mathbb{R}$, define

$$L^t := \begin{pmatrix} r^t e^{it\theta} & 0 \\ 0 & r^{-t}e^{-it\theta} \end{pmatrix} \in SL(2, \mathbb{C}).$$

It is easy to check that $L^{t+s} = L^t L^s$ and that $L^1 = L$. Therefore, for any $\rho \in \mathbb{H}^3$, $L.(L^t.\rho) = L^{t+1}.\rho$. With this in mind, for any $\rho = z + tj \in \mathbb{H}^3$, define

$$\begin{aligned} p_\rho(s) :&= L^s.\rho = r^s e^{is\theta} (z + tj) r^s e^{is\theta} \\ &= r^{2s} \left(e^{2is\theta} z + tj\right). \end{aligned}$$

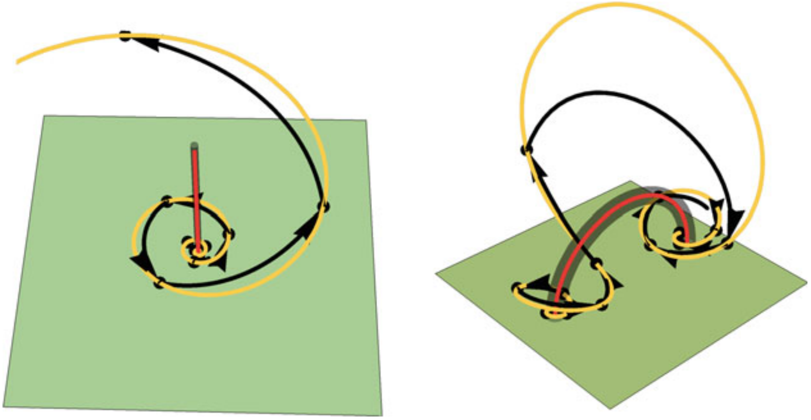


Fig. 5.20 An illustration of the spiraling paths (drawn in yellow) preserved by loxodromic transformations. The lines along which the loxodromic transformation acts as a translation is drawn in red. Arrows show the direction in which points are moved along.

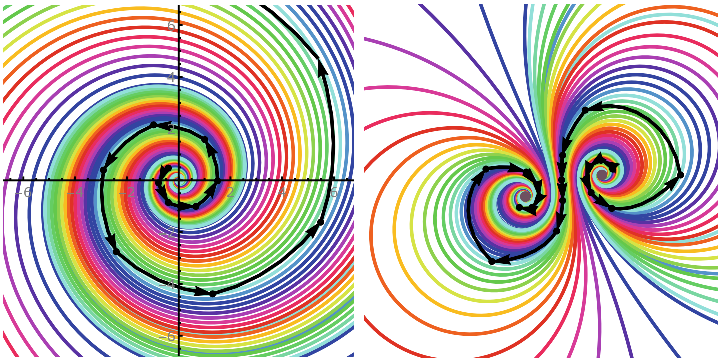


Fig. 5.21 A visualization of the action of a loxodromic element on $\mathbb{C}P^1$. On the left-hand side, the fixed points are 0 and ∞; points are moved outward along the spirals. On the right-hand side, the fixed points are both points in the plane, but still one of them is attractive and points move along spirals from one to the other.

This is a spiraling path in $\mathbb{H}^3$, but notice that due to the way that we have defined it, $\psi^{-1} \circ \varphi \circ \psi$ simply moves points on p_ρ back to points on p_ρ, but further up the spiral. An immediate corollary is that if we define $q_\rho(t) := \psi^{-1}(L.\rho)$, then this is a spiraling path that is preserved by φ. Both of these are illustrated in Figure 5.20.

We can apply this same idea to understand the dynamics on $\mathbb{C}P^1$. In the special case where the fixed points 0 and ∞, we have spirals $p_z(s) = r^{2s}e^{2is\theta}z$—in general, it will instead be some curve $q_z(s) = \psi^{-1}(r^{2s}e^{2is\theta}z)$ instead. In any case, one of the two fixed points will be attractive, and the effect of the loxodromic element

$\varphi \in PSL(2, \mathbb{C})$ is to move points along these spiraling curves away from one of the fixed points and toward the other. This is illustrated in Figure 5.21.

5.7 Other Decomposition Theorems

We finish this chapter by giving a few other ways that we can decompose the groups $PSL(2, \mathbb{R})$ and $PSL(2, \mathbb{C})$—the arguments will be a mixture of algebra and geometry, using a bit of everything that we have learned.

Theorem 5.15 *Every element $\varphi \in PSL(2, \mathbb{C})$ is conjugate to exactly one element in the set*

$$\{z \mapsto z + 1\} \cup \left\{az \middle| a \in \mathbb{C}^\times,\ |a| > 1\right\} \cup \left\{e^{i\theta} z \middle| 0 \le \theta \le \pi\right\}.$$

Proof We know that any element in $\varphi \in PSL(2, \mathbb{C})$ is either the identity, elliptic, hyperbolic, parabolic, or loxodromic. The identity is conjugate to itself and nothing else, so it suffices to note that it is in our defined set. If $\varphi \in PSL(2, \mathbb{C})$ is parabolic, then we know it is conjugate to $z \mapsto z + b$ for some non-zero $b \in \mathbb{C}$. However, we can actually take $b = 1$, because

$$\begin{pmatrix} b^{-1/2} & 0 \\ 0 & b^{1/2} \end{pmatrix} \begin{pmatrix} 1 & b \\ 0 & 1 \end{pmatrix} \begin{pmatrix} b^{-1/2} & 0 \\ 0 & b^{1/2} \end{pmatrix}^{-1} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}.$$

So, every parabolic element is conjugate to $z \mapsto z + 1$, which is in our set. As the other elements in the set are not parabolic (they have two fixed points), φ is only conjugate to one such element.

Now, if φ is not parabolic or the identity, then it is elliptic, hyperbolic, or loxodromic. In all of these cases, we know that it is conjugate to $z \mapsto az$ for some $a \in \mathbb{C}^\times \backslash \{1\}$. However, since

$$\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \begin{pmatrix} a^{1/2} & 0 \\ 0 & a^{-1/2} \end{pmatrix} \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}^{-1} = \begin{pmatrix} a^{-1/2} & 0 \\ 0 & a^{1/2} \end{pmatrix},$$

we can always replace a with $1/a$ instead, which means that we can assume that $|a| \ge 1$. Moreover, if $|a| = 1$, then we may assume that $a = e^{i\theta}$ with $0 \le \theta \le \pi$, since $1/a = e^{-i\theta}$. Therefore, φ is conjugate to an element in our set. Why can't it be conjugate to more than one? Well, note that if it is conjugate to $z \mapsto az$, then $\operatorname{tr}^2(\varphi) = \operatorname{tr}^2(z \mapsto az)$. Since

$$\operatorname{tr} \begin{pmatrix} a^{1/2} & 0 \\ 0 & a^{-1/2} \end{pmatrix} = a^{1/2} + a^{-1/2},$$

$\mathrm{tr}^2(\varphi) = 2 + a + 1/a$. So, if $\mathrm{tr}^2(z \mapsto az) = \mathrm{tr}^2(\varphi) = \mathrm{tr}^2(z \mapsto a'z)$, then either $a = a'$ or $a = 1/a'$. If $|a| \neq 1$, then exactly one of a, a' has norm greater than 1; otherwise, we use the fact that exactly one is of the form $e^{i\theta}$ with $0 \leq \theta \leq \pi$. □

Corollary 5.1 (Jordan Decomposition) *For every matrix $M \in SL(2, \mathbb{C})$, there exists $U \in SL(2, \mathbb{C})$ such that $M = UJU^{-1}$, where J is one of the matrices in the set*

$$\left\{ \pm \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \right\} \cup \left\{ \begin{pmatrix} \lambda & 0 \\ 0 & \lambda^{-1} \end{pmatrix} \middle| |\lambda| \geq 1 \right\}.$$

Remark 5.5 This is just Jordan's normal form theorem, as applied to $SL(2, \mathbb{C})$.

Proof Define $\varphi(z) = M.z$. By the previous theorem, there exists $\psi \in PSL(2, \mathbb{C})$, which we shall write as $\psi(z) = U.z$ for some $U \in SL(2, \mathbb{C})$, such that $\varphi = \psi \circ \gamma \circ \psi^{-1}$ for some $\gamma \in PSL(2, \mathbb{C})$ of a special form. To be precise, $\gamma(z) = J.z$ for some matrix J in the set defined in the statement of the corollary. But this means that $M = \pm UJU^{-1}$. Since $-J$ is the aforementioned set if and only if J is, we see that without loss of generality, $M = UJU^{-1}$. □

Theorem 5.16 *For every $\varphi \in PSL(2, \mathbb{R})$, there exists*

1. *an element $n_x(z) = z + x$ for some $x \in \mathbb{R}$,*
2. *an element $a_r(z) = r^2 z$ for some $r > 0$, and*
3. *an element $k_\theta(z) = \begin{pmatrix} \cos(\theta/2) & -\sin(\theta/2) \\ \sin(\theta/2) & \cos(\theta/2) \end{pmatrix} .z$ for some $0 \leq \theta < 2\pi$*

such that $\varphi = k_\theta \circ a_r \circ n_x$. Moreover, θ, r, n are uniquely determined by φ.

Proof Write $x_0 + y_0 i = \varphi^{-1}(i)$ for some $x_0 \in \mathbb{R}$, $y_0 > 0$. If we define $\varphi' = \varphi \circ n_{x_0} \circ a_{y_0}$, then $\varphi'(i) = \varphi(x_0 + y_0 i) = i$. Since φ' has a fixed point in $\mathbb{H}^2$, it is either the identity or an elliptic element—in either case, it can be written as

$$\varphi'(z) = \begin{pmatrix} \cos(\theta/2) & -\sin(\theta/2) \\ \sin(\theta/2) & \cos(\theta/2) \end{pmatrix} .z$$

for some $0 \leq \theta < 2\pi$—see the exercise at the end of Section 4.6. Therefore, $\varphi = k_\theta \circ a_{y_0^{-1}} \circ n_{-x_0}$, establishing the existence of the desired decomposition. Why is it unique? Well, $\varphi^{-1} = n_{-x} \circ a_{1/r} \circ k_{2\pi - \theta}$, so $\varphi^{-1}(i) = (n_{-x} \circ a_{1/r})(i) = -x + i/r$, so x and r are certainly uniquely determined by φ. The uniqueness of θ follows immediately. □

Corollary 5.2 (Iwasawa Decomposition) *Consider the subgroups*

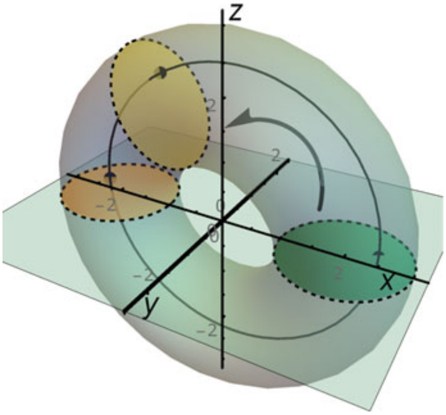


Fig. 5.22 An illustration of how to obtain the solid torus by rotating an open disk around the y-axis.

$$K = \left\{ \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix} \middle| \theta \in \mathbb{R} \right\}$$
$$A = \left\{ \begin{pmatrix} r & 0 \\ 0 & 1/r \end{pmatrix} \middle| r > 0 \right\}$$
$$N = \left\{ \begin{pmatrix} 1 & x \\ 0 & 1 \end{pmatrix} \middle| x \in \mathbb{R} \right\}$$

of $SL(2, \mathbb{C})$. For every $M \in SL(2, \mathbb{C})$, there exist unique matrices $k \in K$, $a \in A$, $n \in N$ such that $M = kan$.

Remark 5.6 Iwasawa decompositions are far more broadly applicable than just for $SL(2, \mathbb{R})$—to start with, every group $SL(n, \mathbb{R})$ has an Iwasawa decomposition, but even this is barely scratching the surface.

Proof I leave this one to the reader. (See Exercise 5.2.18.) □

There are many nice algebraic consequences for the Iwasawa decomposition. In the general spirit of this book, we conclude with a pretty geometric application.

Corollary 5.3 *There exists a bi-continuous map (that is, continuous with a continuous inverse) between $SL(2, \mathbb{R})$ and a solid torus.*

Proof We shall prove that there is a bi-continuous map $S^1 \times \mathbb{R}^2 \to SL(2, \mathbb{R})$ (where S^1 is the unit circle) and a bi-continuous map from $S^1 \times \mathbb{R}^2$ to the solid torus—the desired bi-continuous map can then be obtained as a composition. For the first part, we define a function

$$\begin{aligned} S^1 \times \mathbb{R} \times \mathbb{R} &\to SL(2,\mathbb{R}) \\ (e^{i\theta}, t, x) &\mapsto \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix} \begin{pmatrix} e^t & 0 \\ 0 & e^{-t} \end{pmatrix} \begin{pmatrix} 1 & x \\ 0 & 1 \end{pmatrix}. \end{aligned}$$

That this is bijective is simply the content of the Iwasawa decomposition. That it is continuous is clear—it is a composition of continuous functions like addition, multiplication, and exponentiation. Is the inverse continuous? We could compute it (see Exercise 5.1.6) and see this directly, but here's a different argument. Suppose that we change the entries of a matrix $M \in SL(2,\mathbb{R})$ by a small amount, giving a matrix $\tilde{M} \in SL(2,\mathbb{R})$. Then $\tilde{w} = \tilde{M}^{-1}.i$ must be close to $w = M^{-1}.i$, hence in the Iwasawa decomposition of $\tilde{M}$, the constants r and x are close to what they are for M. This in turn implies that the constant θ is close for M and $\tilde{M}$, which allows us to conclude that the map is continuous.

Now, why is $S^1 \times \mathbb{R}^2$ bi-continuous with a solid torus? Well, one way to think about the solid torus T is that it is what you get if you rotate an open disk like $(x-2)^2 + y^2 < 1$ around the y-axis—see Figure 5.22 for an illustration. If we can find a bi-continuous function $F : \mathbb{R}^2 \to \mathbb{D}^2$, we will be done, because then we get a bi-continuous function $S^1 \times \mathbb{R}^2 \to S^1 \times \mathbb{D}^2$, which we compose with

$$\begin{aligned} S^1 \times \mathbb{D}^2 &\mapsto T \\ (e^{i\theta}, x, y) &\mapsto ((x+2)\cos(\theta), y, \sin(\theta)) \end{aligned}$$

to get the desired bi-continuous function. (This latter map is just what we get by rotating a point in the open disk around the y-axis by θ radians.) Can you map $\mathbb{R}^2 \to \mathbb{D}^2$ in a bi-continuous way? Of course—identify $\mathbb{R}^2$ with $\mathbb{C}$ as we have throughout this book, and define

$$\begin{aligned} f : \mathbb{C} &\to \mathbb{D}^2 \\ z &\mapsto \frac{z}{\sqrt{1+|z|^2}}. \end{aligned}$$

Does this really have the image that we say it does? Yes:

$$|f(z)|^2 = \frac{|z|^2}{1+|z|^2} < 1.$$

Is this function invertible? Yes: define

$$\begin{aligned} g : \mathbb{D}^2 &\to \mathbb{C} \\ z &\mapsto \frac{z}{\sqrt{1-|z|^2}}. \end{aligned}$$

Then

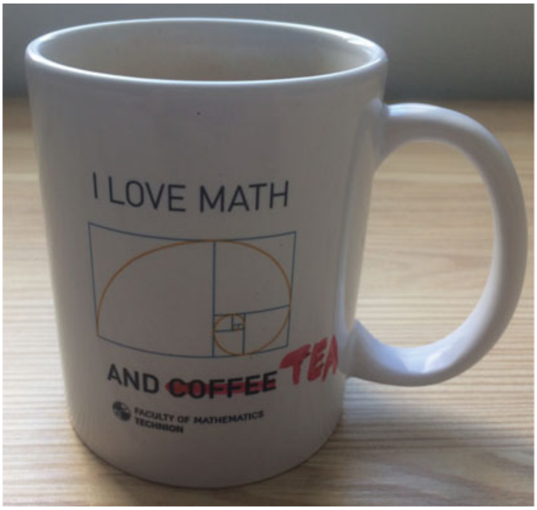


Fig. 5.23 In some sense, this is a picture of $SL(2, \mathbb{R})$.

$$(f \circ g)(z) = f\left(\frac{z}{\sqrt{1-|z|^2}}\right) = \frac{\frac{z}{\sqrt{1-|z|^2}}}{\sqrt{1+\frac{|z|^2}{1-|z|^2}}} = \frac{z}{\sqrt{1-|z|^2+|z|^2}} = z$$

and

$$(g \circ f)(z) = g\left(\frac{z}{\sqrt{1+|z|^2}}\right) = \frac{\frac{z}{\sqrt{1+|z|^2}}}{\sqrt{1-\frac{|z|^2}{1+|z|^2}}} = \frac{z}{\sqrt{1+|z|^2-|z|^2}} = z,$$

so they are inverses. That they are both continuous is also clear, and so we are done. □

Naturally, this result implies that $SL(2, \mathbb{R})$ is bi-continuous with anything that is bi-continuous with a solid torus, which includes anything that you could get by continuously deforming it. A tea mug—like in Figure 5.23—would be an example.

Problems

5.1 COMPUTATIONAL EXERCISES

1. Consider the points $A = ei$, $B = e^{\pi i/6}$, $C = i$ in $\mathbb{H}^2$, along with the triangle ΔABC.

   a) Show that ΔABC is a right triangle, in that the angle at C has angle measure $\pi/2$.
   b) Let a be the length of $\overline{BC}$; b be the length of $\overline{CA}$; c be the length of $\overline{AB}$. Compute a, b, c.
   c) How does $a^2 + b^2$ compare with c^2? What does this tell you about the Pythagorean theorem in hyperbolic geometry?

2. Find a regular hexagon P in the hyperbolic plane such that the sum of the measures of its angles is 3π.
3. Let P be a regular hexagon in the hyperbolic plane, such that the sum of the measures of its angles is 3π. Compute its area.
4. Suppose that we compute an element $SL(2, \mathbb{R})$ randomly, as follows. First, roll a 6-sided dice and divide the result by 3—do this three times to obtain three real numbers a, b, d. Then, define $c = (ad - 1)/b$ so that $M = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in SL(2, \mathbb{R})$.

   a) What is the probability that M is elliptic?
   b) What is the probability that M is hyperbolic?
   c) What is the probability that M is parabolic?

5. Consider the horocycle H in $\mathbb{H}^2$ defined by $x^2 + (y - 1)^2 = 1$. Determine the set of elements $\varphi \in PSL(2, \mathbb{R})$ such that $\varphi(H) = H$.
6. Consider the function

$$S^1 \times \mathbb{R} \times \mathbb{R} \mapsto SL(2, \mathbb{R})$$
$$\left(e^{i\theta}, t, x\right) \mapsto \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix} \begin{pmatrix} e^t & 0 \\ 0 & e^{-t} \end{pmatrix} \begin{pmatrix} 1 & x \\ 0 & 1 \end{pmatrix};$$

   here, S^1 denotes the unit circle. Compute the inverse of this function.

5.2 PROOFS

1. Choose any point z in the hyperbolic plane and consider the set of lines passing through l. Prove that the curves that are orthogonal to each of the aforementioned lines are hyperbolic circles centered at z. *(Hint: you may wish to choose z to be some particularly convenient point.)*

2. Let ΔABC, $\Delta A'B'C'$ be idealized hyperbolic triangles with all of their vertices on the boundary. Prove that there exists a unique isometry φ such that $\varphi(A) = A'$, $\varphi(B) = B'$, and $\varphi(C) = C'$.
3. Let ΔABC be a hyperbolic triangle. Let l be an angle bisector of the angle at the vertex A—that is, a hyperbolic line l that passes through A such that the angle from AB to l is equal to the angle from l to CA. Prove that l intersects BC. *(Hint: the fact that l is an angle bisector is mostly irrelevant—we just need to know that l passes through A and is between AB and CA.)*
4. Let ΔABC be a hyperbolic triangle. Let l be the angle bisector of A and l' the angle bisector of B. Prove that l and l' intersect at some point p. This point is known as the *incenter*.
5. Let ΔABC be a hyperbolic triangle in $\mathbb{H}^2$. Prove that there exists a unique isometry $\varphi \in \mathrm{Isom}(\mathbb{H}^2, d_{\text{hyper}})$ such that $\varphi(A) = i$, $\varphi(B) = it$ with $t > 1$, and $\varphi(C) = x + iy$ with $x, y > 0$. We shall say that φ puts ΔABC into *standard position*.
6. We prove a result known as the hyperbolic law of cosines. Let ΔABC be a hyperbolic triangle. Denote its angle measures by $\angle A$, $\angle B$, and $\angle C$ and the lengths of its sides opposite to those angles to be a, b, and c.
   a) Prove that there exists an isometry φ sending ΔABC into $\mathbb{D}^2$ such that $\varphi(A) = 0$, $0 < \varphi(C) < 1$, and $\Im(\varphi(B)) > 0$. Since hyperbolic isometries preserve both angles and distances, we shall henceforth simply assume that $A = 0$, $0 < C < 1$, and $\Im(B) > 0$.
   b) Prove that $C = \tanh(b/2)$ and $B = \tanh(c/2)e^{i\angle A}$.
   c) Prove that

   $$\cosh(a) = \cosh(b)\cosh(c) - \cos(\angle A)\sinh(b)\sinh(c).$$

   (Hint: $\cosh(a) = \cosh(d_{\text{hyper}}(B, C)$. *You will need to use a lot of identities of hyperbolic functions to simplify everything.)*

7. Let ΔABC be a hyperbolic triangle. Denote its angle measures by $\angle A$, $\angle B$, and $\angle C$ and the lengths of its sides opposite to those angles to be a, b, and c. From the hyperbolic law of cosines, we know that

   $$\cos(\angle A) = \frac{\cosh(b)\cosh(c) - \cosh(a)}{\sinh(b)\sinh(c)}.$$

   a) Prove that
   $$\frac{\sin^2(\angle A)}{\sinh(a)^2} = \frac{1 - \alpha_a^2 - \alpha_b^2 - \alpha_c^2 + 2\alpha_a\alpha_b\alpha_c}{\sinh(a)^2\sinh(b)^2\sinh(c)^2},$$
   where $\alpha_a = \cosh(a)$, $\alpha_b = \cosh(b)$, and $\alpha_c = \cosh(c)$.

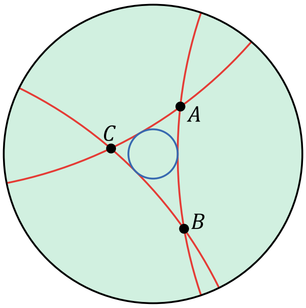
b) Prove that

$$\frac{\sin(\angle A)}{\sinh(a)} = \frac{\sin(\angle B)}{\sinh(b)} = \frac{\sin(\angle C)}{\sinh(c)}.$$

This is known as the hyperbolic law of sines.

8. Let ΔABC, $\Delta A'B'C'$ be hyperbolic triangles such that the length of AB is equal to the length of $A'B'$, $\angle A = \angle A'$, and $\angle B = \angle B'$. Prove that there exists a hyperbolic isometry φ such that $\varphi(A) = A'$, $\varphi(B) = B'$, and $\varphi(C) = C'$. This is the ASA theorem. *(Hint: put ΔABC and $\Delta A'B'C'$ into standard position.)*
9. Let ΔABC, $\Delta A'B'C'$ be hyperbolic triangles such that the length of AB is the equal to the length of $A'B'$, the length of BC is equal to the length of $B'C'$, and $\angle B = \angle B'$. Prove that there exists a hyperbolic isometry φ such that $\varphi(A) = A'$, $\varphi(B) = B'$, and $\varphi(C) = C'$. This is the SAS theorem. *(Hint: put ΔABC and $\Delta A'B'C'$ into standard position.)*
10. Let ΔABC, $\Delta A'B'C'$ be hyperbolic triangles such that the length of AB is the equal to the length of $A'B'$, the length of BC is equal to the length of $B'C'$, and the length of CA is equal to the length of $C'A'$. Prove that there exists a hyperbolic isometry φ such that $\varphi(A) = A'$, $\varphi(B) = B'$, and $\varphi(C) = C'$. This is the SSS theorem. *(Hint: put ΔABC and $\Delta A'B'C'$ into standard position.)*
11. Let ΔABC, $\Delta A'B'C'$ be hyperbolic triangles such that $\angle A = \angle A'$, the length of AB is equal to the length of $A'B'$, and $\angle C = \angle C'$. Prove that there exists an isometry φ such that $\varphi(A) = A'$, $\varphi(B) = B'$, and $\varphi(C) = C'$. This is the AAS theorem. *(Hint: the law of sines and the law of cosines might be useful here.)*
12. Prove that for any hyperbolic triangle ΔABC, there exists a unique circle that is tangent to AB, BC, and CA. This is known as the *incircle* of ΔABC. *(Hint: the center of this circle is the incenter. The proof is exactly the same as in Euclidean geometry; you will need to use the AAS theorem.)*
13. Let ΔABC, $\Delta A'B'C'$ be hyperbolic triangles, and suppose that $\angle A = \angle A'$, $\angle B = \angle B'$, and $\angle C = \angle C'$. We aim to prove that there exists an isometry φ such that $\varphi(A) = A'$, $\varphi(B) = B'$, and $\varphi(C) = C'$—that is, the AAA theorem.

a) Using an isometry if necessary, we can assume that ΔABC, $\Delta A'B'C'$ are hyperbolic triangles in $\mathbb{D}^2$, their incenters are 0, and AB is orthogonal to $\mathbb{R}$. That is, they are in the following standard configuration.



Let r be the hyperbolic radius of the incircle. Prove that the inversive coordinates of AB, CA, and BC are

$$(\sinh(r), \sinh(r), \cosh(r))$$
$$\left(\sinh(r), \sinh(r), \cosh(r)e^{i\theta}\right)$$
$$\left(\sinh(r), \sinh(r), \cosh(r)e^{i\phi}\right)$$

for some θ, ϕ.

b) Prove that

$$\cos(\theta)\cosh(r)^2 - \sinh(r)^2 = \cos(\angle A)$$
$$\cos(\phi)\cosh(r)^2 - \sinh(r)^2 = \cos(\angle B)$$
$$\cos(\theta-\phi)\cosh(r)^2 - \sinh(r)^2 = \cos(\angle C).$$

(Hint: use the inversive product.)

c) Define $\lambda_a = 1-\cos(\angle A)$, $\lambda_b = 1-\cos(\angle B)$, $\lambda_c = 1-\cos(\angle C)$. Prove that

$$\cos(\theta) = 1 + \frac{\lambda_a}{\cosh(r)^2}$$
$$\cos(\phi) = 1 + \frac{\lambda_b}{\cosh(r)^2}$$
$$\cos(\theta-\phi) = 1 + \frac{\lambda_c}{\cosh(r)^2}.$$

d) Prove that

$$\cosh(r)^2 = \frac{2\lambda_a\lambda_b\lambda_c}{\lambda_a^2+\lambda_b^2+\lambda_c^2-(\lambda_a-\lambda_b)^2+(\lambda_b-\lambda_c)^2+(\lambda_c-\lambda_a)^2}.$$

e) Prove that $\Delta ABC = \Delta A'B'C'$. *(Hint: the previous part shows that* $\cosh(r)$ *is uniquely determined by the angles. Use this to show that the sides are also uniquely determined.)*

14. Prove that a hyperbolic circle is the incircle of some hyperbolic triangle if and only if its hyperbolic radius is less than $\ln(3)/2$. *(Hint: use an isometry to reduce to the case where the center of the circle is* $0 \in \mathbb{D}^2$*.)*
15. Complete the proof of Theorem 5.5.
16. Prove Theorem 5.7.
17. Prove Theorem 5.11.
18. Prove the existence and uniqueness of the Iwasawa decomposition for $SL(2, \mathbb{R})$—that is, Corollary 5.2. *(Hint: look at how Corollary 5.1 was proved from the corresponding result about* $PSL(2, \mathbb{C})$*. Mimic this proof strategy.)*

5.3 PROOFS (Calculus)

1. Consider the function $f(x) = x + 1/x$ defined on $(0, \infty)$. Prove that $f'(x) > 0$ if $x > 1$, $f'(x) < 0$ if $x < 1$, and $f'(x) = 0$ if $x = 1$. Use this to show that $f(x)$ has a global minimum at $x = 1$.
2. We wish to define what the hyperbolic area for a region in $\mathbb{H}^2$ is. Consider a Euclidean rectangle with vertices $A = (x, y)$, $B = (x+\Delta x, y)$, $C = (x, y+\Delta x)$, $D = (x + \Delta x, y + \Delta y)$. If Δx and Δy are very small, the hyperbolic area of this region should be approximately the hyperbolic length from A to B times the hyperbolic length from B to C.

   a) Prove that

$$\lim_{\Delta y \to 0} \left(d_{\text{hyper}}\left(x + iy, x + i(y + \Delta)y\right) - \frac{1}{y} d_{\text{Euclid}}\left(x + iy, x + i(y + \Delta)y\right) \right) = 0$$

$$\lim_{\Delta x \to 0} \left(d_{\text{hyper}}\left(x + iy, (x + \Delta x) + iy\right) - \frac{1}{y} d_{\text{Euclid}}\left(x + iy, x + \Delta x + iy\right) \right) = 0.$$

   b) Given the result of the preceding part, why is the definition that the area of $R \subset \mathbb{H}^2$ is

$$\int_R \frac{\mathrm{d}x\, \mathrm{d}y}{y^2}$$

   sensible?

3. We compute the hyperbolic area of idealized triangles.

   a) For any $R \subset \mathbb{H}^2$ and any isometry $\varphi \in \operatorname{Isom}(\mathbb{H}^2, d_{\text{hyper}})$, why is

   $$\int_R \frac{\mathrm{d}x\,\mathrm{d}y}{y^2} = \int_{\varphi(R)} \frac{\mathrm{d}x\,\mathrm{d}y}{y^2}?$$

   (Hint: there are two ways to go about this. You can argue by thinking about what happens to small rectangles under hyperbolic isometries. If you are familiar with the basics of differential forms, it can also be done simply via the usual change of coordinates formulas. In either case, it might be helpful to consider basic kinds of elements in $\operatorname{Isom}(\mathbb{H}^2, d_{\text{hyper}})$*.)*

   b) Let R be the subset of $\mathbb{H}^2$ bounded by the lines $x = -1$, $x = 1$ and the circle $x^2 + y^2 = 1$. Prove that

   $$\int_R \frac{\mathrm{d}x\,\mathrm{d}y}{y^2} = \pi.$$

   (Hint: the hardest part here is to set up the bounds of the integral. Simplify your life by dividing R into three pieces by cutting along $y = 1$.)

   c) Prove that any idealized hyperbolic triangle with all of its vertices on the boundary has area π. *(Hint: use the result of Exercise 5.2.2.)*

5.4 PROOFS (Group Theory)

1. Let G be a group acting on a set X. A *fixed point* of G is a point $x \in X$ such that $g.x = x$ for all $g \in G$.

   a) Prove that the action of S^1 on $\mathbb{C}$ via matrix multiplication has exactly one fixed point. What is it?

   b) Prove that the action of $SL(2, \mathbb{R})$ on $\mathbb{R}^2$ via matrix multiplication has exactly one fixed point. What is it?

   c) Prove that the action of $SL(2, \mathbb{R})$ on $\mathbb{H}^2$ (via hyperbolic isometries) has no fixed points.

2. Let G be a group acting on a set X. For any point $x \in X$, the *stabilizer subgroup* G_x is the set of all $g \in G$ such that $g.x = x$.

   a) Prove that G_x is a group.

   b) Prove that G_x is the largest subgroup of G for which x is a fixed point.

   c) Determine the stabilizer subgroup of $(1, 0) \in \mathbb{R}^2$ under the action of $SL(2, \mathbb{R})$ via matrix multiplication.

   d) Recall that $\operatorname{Sym}(X)$ is the group of bijective functions $f : X \to X$ with composition as the operation. Show that for any $x \in X$, $\operatorname{Sym}(X)_x$ is isomorphic as a group to $\operatorname{Sym}(X \backslash \{x\})$.

e) Prove that if the action is transitive, then all of the stabilizer subgroups are isomorphic to one another.

3. Consider $PSL(2, \mathbb{R})$ acting on $\mathbb{H}^2$ via hyperbolic isometries. Prove that the stabilizer subgroup of any point in $\mathbb{H}^2$ is the set of elliptic elements in $PSL(2, \mathbb{R})$ that are rotations around that point.
4. Consider $PSL(2, \mathbb{C})$ acting on $\mathbb{H}^3$ via hyperbolic isometries. What is the stabilizer subgroup of a point in $\mathbb{H}^3$?

Set Theory

A

Throughout this book, I make heavy use of set theoretic notation. For our purposes, we do not need to dig into the abstract formalism of set theory; instead, we will make use of what is commonly known as naive set theory. Those interested in learning more might consult Halmos' *Naive Set Theory* [5] or Jech's *Set Theory* [7].

A.1 Basic Constructions

To begin with, what is a set? Roughly speaking, a *set* is a collection. Its defining characteristics are what its *elements* are. For instance, one can have sets like

$$S_1 = \{1, 2, 5, 11\}$$
$$S_2 = \{`b', `e', `f'\},$$

where the first set has four elements 1, 2, 5, 11, and the second has three elements 'a', 'e', 'f'. The elements are not ordered, so, for example,

$$\{1, 2, 5, 11\} = \{5, 2, 11, 1\}.$$

Two sets are the same if and only if they have the same elements. If x is an element of a set S, then we express this as $x \in S$. If x is not an element of S, then we write this as $x \notin S$. Thus, $1 \in S_1$, $11 \in S_1$, but $13 \notin S_1$. These elements can be anything whatsoever, including other sets—so, for instance,

$$S_3 = \{0, 2, 3, \{1, 3, 7\}\}$$

is a perfectly kosher set which has four elements in it, namely 0, 2, 3 and $\{1, 3, 7\}$. There are various standard operations on sets—for example, we can take the *union* of two sets, which produces a new set that contains precisely all of the elements of both sets. So, for example,

A. Sheydvasser, *Linear Fractional Transformations*, Undergraduate Texts in Mathematics, https://doi.org/10.1007/978-3-031-25002-6

$$\begin{aligned} S_1 \cup S_2 &= \{1, 2, 5, 11, \text{‘}b\text{’}, \text{‘}e\text{’}, \text{‘}f\text{’}\} \\ S_1 \cup S_3 &= \{0, 1, 2, 3, 5, \{1, 3, 7\}\}\,. \end{aligned}$$

Another common operation is taking intersections—the *intersection* of two sets is a set containing precisely all of the elements that are in both sets, such as

$$\begin{aligned} S_1 \cap S_3 &= \{2\} \\ S_2 \cap S_3 &= \{\}, \end{aligned}$$

where $\{\}$ is the *empty set*, also denoted by $\emptyset$, which is the unique set that has no elements. Given a set S, we can always form a set $\{S\}$ that just has S as an element. For instance,

$$\{\emptyset\} = \{\{\}\}$$

is an entirely acceptable set which, it must be stressed, is not the same as $\emptyset$—$\emptyset$ contains no elements, but $\{\emptyset\}$ has exactly one. Intuitively, if you think about a set like a basket, then $\emptyset$ is an empty basket, but $\{\emptyset\}$ is a basket that has a basket inside of it, which is quite different.

Given any set S, we say that another set T is a *subset*—and we write $T \subset S$—if every element in T is an element of S. We say that T is a *proper subset* of S if $T \subset S$ and $T \neq S$. (Note that if $S = T$ then $S \subset T$ is trivially true.) Thus, if

$$\begin{aligned} X_1 &= \{\triangle, \square, \bigcirc\} \\ X_2 &= \{\triangle, \bigcirc\} \\ X_3 &= \{\triangle\}, \end{aligned}$$

then $X_3 \subset X_2 \subset X_1$ (in fact, they are all proper subsets), but X_1 is not a subset of X_2 and X_2 is not a subset of X_3. It is easy to see that $S = T$ if and only if $S \subset T$ and $T \subset S$—this is also one of the most common ways to prove that two sets are in fact the same. There are two very important operations involving subsets. First, given any set S, we can produce a new subset $T \subset S$ that consists precisely out of all elements in S satisfying some kind of condition. This is most commonly written using what is known as "set-builder" notation, where the set that we are taking a subset of is on the left, and the condition we wish to impose is on the right. For example,

$$\{x \in S_1 | x \leq 6\} = \{1, 2, 5\}.$$

That is, we started with the set $S_1 = \{1, 2, 5, 11\}$ and imposed the condition that we only keep those elements $x \in S_1$ such that $x \leq 6$. Here is another example:

$$\begin{aligned} E &= \{\{0, 1\}, \{1, 3\}, \{2\}, \{4, 5, 6\}, \{1\}\} \\ F &= \{S \in E | 1 \in S\} \\ &= \{\{0, 1\}, \{1, 3\}, \{1\}\}\,. \end{aligned}$$

To start, E is a set of some sets of integers; F is just the subset of all sets that contain 1. This process of producing subsets by imposing some condition on an existing set is known as *restricted comprehension*. One special instance of this is when we remove elements in a set from a larger set that contains it—that is, if $S_1 \subset S_2$ we define

$$S_2 \backslash S_1 = \{x \in S_2 | x \notin S_1\} .$$

Another important construction involving subsets is the *power set* of a set S—typically denoted by $P(S)$ or 2^S—which is the set of all subsets of S. The rationale for the notation 2^S is simple: if S is a finite set with n elements, then 2^S will have 2^n elements. For example,

$$2^{X_1} = \{\{\triangle, \square, \bigcirc\}, \{\triangle, \square\}, \{\triangle, \bigcirc\}, \{\square, \bigcirc\}, \{\triangle\}, \{\square\}, \{\bigcirc\}, \emptyset\}$$

has $2^3 = 8$ elements, whereas $X_1 = \{\triangle, \square, \bigcirc\}$ has 3. It is a good exercise for the reader to show that this is true in general.

A.2 Some Common Sets

Thus far, I have only listed examples of finite sets. However, most of the sets that we will be interested in are not finite. Here are some examples, together with the symbols typically used to denote those sets.

$\mathbb{N}$	the set of natural numbers
$\mathbb{Z}$	the set of integers
$\mathbb{Q}$	the set of rational numbers
$\mathbb{Q}^+$	the set of positive rational numbers
$\mathbb{Q}^\times$	the set of non-zero rational numbers
$\mathbb{R}$	the set of real numbers
$\mathbb{R}^+$	the set of positive real numbers
$\mathbb{R}^\times$	the set of non-zero real numbers
$\mathbb{C}$	the set of complex numbers
$\mathbb{C}^\times$	the set of non-zero complex numbers

Here, by the natural numbers, I mean the set $\{0, 1, 2, 3, \ldots\}$. I wish I could honestly say that this was completely standard notation, but sadly some authors exclude 0 as a natural number. Consequently, to avoid confusion, I will largely try not to mention natural numbers. Each of the above-mentioned sets has a standard construction in set theory. For instance, using von Neumann ordinals, one could define the natural numbers by nesting sets inside of one another, starting with the empty set. This is done as follows:

$$\begin{aligned}
0 &= \emptyset \\
1 &= 0 \cup \{0\} = \{\emptyset\} \\
2 &= 1 \cup \{1\} = \{\emptyset, \{\emptyset\}\} \\
3 &= 2 \cup \{2\} = \{\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}\} \\
&\vdots \\
n &= (n-1) \cup \{n-1\}.
\end{aligned}$$

One can then use this to define addition and multiplication entirely in terms of sets. Constructions such as these are useful in that they allow us to reduce vast swathes of mathematics to just set theory. On the other hand, they tend to be both somewhat unwieldy and just weird—with the von Neumann definition, it is true that $n - 1 \in n$ which has never sat right with me. Other approaches are available: one could, for example, treat everything axiomatically. For the natural numbers, the standard choice of axioms is something like Peano's axioms. One could also apply other branches of mathematics such as category theory to give descriptions of these objects. However, I will not worry about these foundational issues. For our purposes, it shall suffice to know that these basic objects exist so that we can build on top of them. I think that the exact details of how to define them are best left to a different text.

A.3 Ordered Pairs and Relations

By definition, sets are unordered. However, we will often need to consider objects like sets but where the order of the elements matters, and where an element can appear multiple times. There is a very clever definition due to Kuratowski that just makes use of set theoretic language—specifically, in 1921 he defined an ordered pair as

$$(a, b) = \{\{a\}, \{a, b\}\}.$$

Kuratowski's was not the first definition; Weiner and Hausdorff had made their own definitions in 1914. It just happens that Kuratowski's definition is the easiest to work with using standard formulations of set theory for the purposes of proving things. Regardless, all definitions of ordered pairs share in that they essentially add some kind of asymmetry to the set, which allows you to differentiate a "first" element and a "second" element, allowing you to prove the characteristic property of ordered pairs, namely that $(a, b) = (c, d)$ if and only if $a = c$ and $b = d$. One can extend this to ordered triples, ordered quadruples and the like by just nesting ordered pairs—e.g.,

$$(a, b, c, d) := (((a, b), c), d).$$

Again, the exact mechanics of how we define things doesn't really matter for our purposes—the important thing is the characteristic property that $(x_1, x_2, \ldots x_n) =$

$(y_1, y_2, \ldots y_n)$ if and only if $x_i = y_i$ for all $1 \leq i \leq n$. With this, given a collection of sets $S_1, S_2, \ldots S_n$, we can define their *Cartesian product* to be the set

$$\begin{aligned}\bigtimes_{i=1}^{n} S_i &= S_1 \times S_2 \times \ldots S_n \\ &= \{(x_1, x_2, \ldots x_n) | x_1 \in S_1,\ x_2 \in S_2, \ldots x_n \in S_n\}.\end{aligned}$$

It is customary to write S^n to denote the Cartesian product of a set S with itself n times. As an example, the set of Cartesian pairs of real numbers is $\mathbb{R}^2$—this is essentially the set of Cartesian coordinates of the plane, whence the name.

A *binary relation* R between two sets S_1, S_2 is formally defined as a subset of $S_1 \times S_2$. However, in practice, we usually think about relations in a slightly different way: given two elements $x \in S_1$, $y \in S_2$, we say that the relation holds for x, y if $(x, y) \in R$. We also usually use a slightly different notation: instead of writing $(x, y) \in R$, we instead write xRy. Thus, you should essentially think of the R of $S_1 \times S_2$ as the collection of pairs for which the relation R is true. Here is an example: take any set S and define a subset

$$R = \left\{(x, y) \in S^2 \middle| x = y\right\}.$$

The relation R is secretly just =. Here is another example:

$$R = \left\{(x, y) \in \mathbb{Z}^2 \middle| x = y + z,\ \text{for some } z \in \mathbb{N}\right\}.$$

Given two integers x, y, xRy if and only if $x = y + z$ for some natural number z—that is, if $x \geq y$. Here is one last example:

$$R = \left\{(p_1, p_2) \in \text{Humans}^2 \middle| p_1 \text{ is a child of } p_2\right\}.$$

For any binary relation on two sets S_1, S_2, we refer to S_1 as the *domain* and S_2 as the *codomain*.

Functions

A *function* f on two sets S_1, S_2 is a special type of relation with the following additional requirement: for every $x \in S_1$, there exists a unique $y \in S_2$ such that $(x, y) \in f$. Based on the fact that it is unique, we usually write this in the more familiar form $f(x) = y$. To signify that a relation is a function, we typically employ a slightly different notation and write $f : S_1 \rightarrow S_2$—here f is the name of the function, S_1 is its domain, and S_2 is its codomain. We often want to specify a function by some type of rule. The usual notation for this is

$$\begin{aligned} f : S_1 &\to S_2 \\ x &\mapsto \varphi(x). \end{aligned}$$

Here, the first line is telling us the name of the function, its domain and codomain; the second line is telling us that for any element $x \in S_1$, $f(x)$ returns whatever statement $\varphi(x)$ is. For example, given any set S, we might want to define the *identity function*

$$\begin{aligned} id_S : S &\to S \\ x &\mapsto x \end{aligned}$$

which has the following behavior: it accepts as inputs elements $x \in S$, and simply returns them back, unchanged. We might also have less trivial functions such as

$$\begin{aligned} f : \mathbb{Z} &\to \mathbb{R} \\ n &\mapsto \sqrt{1 + n^2}. \end{aligned}$$

This should be understood as follows: as a function, f accepts integers and returns real numbers. For any integer n, $f(n) = \sqrt{1 + n^2}$. Here is another example:

$$\begin{aligned} E : \mathbb{R}^+ \times \mathbb{R} &\to \mathbb{R} \\ (x, y) &\mapsto x^y. \end{aligned}$$

Here, the function E accepts ordered pairs (x, y) of real numbers where the first element has to be positive and returns x^y. Sometimes, we will have functions that are easiest to describe piecewise: we might say that they do some simple operation φ_1 if some condition C_1 is satisfied, or they might do some other simple operation φ_2 if some other condition C_2 is satisfied. Formally, given a function $f : S_1 \to S_2$ such that $S_1 = C_1 \cup C_2 \cup \ldots C_n$ and the sets C_i do not intersect one another, we will write

$$\begin{aligned} f : S_1 &\to S_2 \\ x &\mapsto \begin{cases} \varphi_1(x) & \text{if } x \in C_1 \\ \varphi_2(x) & \text{if } x \in C_2 \\ \vdots & \vdots \\ \varphi_n(x) & \text{if } x \in C_n \end{cases} \end{aligned}$$

to mean that $f(x) = \varphi_i(x)$ if $x \in C_i$. There are many common examples of this, such as

$$\begin{aligned} |\cdot| : \mathbb{R} &\to \mathbb{R} \\ x &\mapsto \begin{cases} x & \text{if } x \geq 0 \\ -x & \text{if } x < 0 \end{cases} \end{aligned}$$

which we all know as the absolute value function. We will have our own examples of such functions, such as

$$\begin{aligned}|\kappa| : \left\{\text{circles, lines in } \mathbb{R}^2\right\} &\mapsto \mathbb{R} \\ C &\mapsto \begin{cases} \frac{1}{R} & \text{if } C \text{ is a circle with radius } R \\ 0 & \text{if } C \text{ is a line.} \end{cases}\end{aligned}$$

Another common way of building new functions from old ones is to compose them together. Specifically, suppose that I have a function $f : X \to Y$ and a function $g : Y \to Z$. Then I can define a new function

$$\begin{aligned}h : X &\to Z \\ x &\mapsto g(f(x)).\end{aligned}$$

We call h the *composition* of f and g; we usually denote this relationship by writing $h = g \circ f$.

Finally, it is common to produce new functions via *restriction*. Suppose that we have a function $\Omega : X \to Y$ and $Z \subset X$. Then we can define a new function

$$\begin{aligned}\Omega|_Z : Z &\to Y \\ x &\mapsto \Omega(x).\end{aligned}$$

Intuitively, this is just the same function as before; we have simply shrunk its domain a little.

A.4 Injections, Surjections, and Bijections

A function $f : X \to Y$ is called *injective* (or *one-to-one*, or sometimes *monic*) if for every $x_1, x_2 \in X$, $f(x_1) = f(x_2)$ implies that $x_1 = x_2$. A function $f : X \to Y$ is called *surjective* (or *onto*, or sometimes *epic*) if for every $y \in Y$, there exists $x \in X$ such that $f(x) = y$. Intuitively, injections are functions that never map to the same element in Y twice; surjections are functions that map to every element in Y. Let's look at some examples. The function

$$\begin{aligned}f : \mathbb{Z} &\to \mathbb{Z} \\ n &\mapsto 2n\end{aligned}$$

is injective, since $2n = 2m$ implies that $n = m$. However, it is not surjective, since there is no integer n such that $2n = 1$, for instance. The function

$$\begin{aligned}g : \mathbb{R} \times \mathbb{R} &\mapsto \mathbb{R} \\ (x, y) &\mapsto x + y\end{aligned}$$

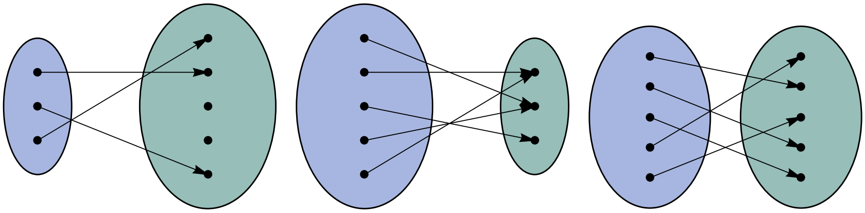


Fig. A.1 Visual representations of three functions; the first one is injective, the second one is surjective, and the third is bijective.

is surjective, since for any $z \in \mathbb{R}$, $g((z, 0)) = z$. However, it is not injective, since it is also true that $g((z-1, 1)) = z$, among infinitely other options. A good exercise is to go through the functions listed in the previous section and to decide whether they are injective, surjective, both, or neither.

A function that is both injective and surjective is called *bijective*. Such a function is one such that for every $y \in Y$ in the codomain, there exists a unique $x \in X$ in the domain such that $f(x) = y$. This is to say that any bijective function $f : X \to Y$ is *invertible*—there exists another function $f^{-1} : Y \to X$ such that $f^{-1} \circ f = id_X$ and $f^{-1} \circ f = id_Y$. In fact, this goes the other way as well.

Theorem A.1 *Let $f : X \to Y$ be a function; it is invertible if and only if it is bijective.*

Proof We have already demonstrated that if f is bijective, then it is invertible. It remains to show that if there exists a function $f^{-1} : Y \to X$ such that $f^{-1} \circ f = id_X$ and $f^{-1} \circ f = id_Y$ then f is bijective. Well, choose any two $x_1, x_2 \in X$ and suppose that $f(x_1) = f(x_2)$. It follows that

$$\begin{aligned} f^{-1}(f(x_1)) &= f^{-1}(f(x_2)) \\ id_X(x_1) &= id_X(x_2) \\ x_1 &= x_2. \end{aligned}$$

Therefore, f is injective. Next, choose any $y \in Y$. Note that if we choose $x = f^{-1}(y) \in X$, then $f(x) = f(f^{-1}(y)) = id_Y(y) = y$. Therefore, f is surjective. We conclude that it is bijective. □

Intuitively, a bijection matches up all of the elements in the domain and all of the elements in the codomain in a unique fashion; if there is a bijection between two sets, then you can essentially think of one set as just being the other one, but with all of its elements relabeled, where the bijection is precisely what keeps track of the labeling. Some examples of injective, surjective, and bijective functions are drawn in Figure A.1.

References

1. Bourgain, J., Kontorovich, A.: On the local-global conjecture for integral apollonian gaskets. Inventiones mathematicae **196**, 589–650 (2013)
2. Coxeter, H.S.M.: Inversive geometry. Annali di Matematica **71**, 73–83 (1966)
3. Guettler, G., Mallows, C.: A generalization of apollonian packing of circles. J. Comb. **1**(1), 1–27 (2010)
4. Halmos, P.: I Want to Be a Mathematician: An Automathography. Springer (1985)
5. Halmos, P.R.: Naive Set Theory. Springer (1974)
6. Hartshorne, R.: Foundations of Projective Geometry. Harvard University. Lecture Notes. W. A. Benjamin (1967)
7. Jech, T.: Set Theory. Academic Press (1978)
8. Klein, F.: Vergleichende betrachtungen über neuere geometrische forschungen. Mathematische Annalen **43**, 63–100 (1893)
9. Levrie, P.: A straightforward proof of descartes's circle theorem. Math. Intell. **41**(3), 24–27 (2019)
10. Lockhart, P.: A Mathematician's Lament. Bellevue Literary Press (2009)
11. Needham, T.: Visual Complex Analysis. Clarendon Press (1997)
12. Pólya, G.: How to Solve It: A New Aspect of Mathematical Method. Princeton University Press (1945)
13. Rosen, K.: Elementary Number Theory and Its Applications. Addison-Wesley (2011)
14. Sheydvasser, A.: Quaternion orders and sphere packings. J. Number Theory **204**, 41–98 (2019)
15. Soddy, F.: The kiss precise. Nature **137**, 1021 (1936)
16. Soddy, F.: The bowl of integers and the hexlet. Nature **139**, 77–79 (1937)
17. Stange, K.E.: Visualising the arithmetic of imaginary quadratic fields. Int. Math. Res. Not. (2017)

A. Sheydvasser, *Linear Fractional Transformations*, Undergraduate Texts in Mathematics, https://doi.org/10.1007/978-3-031-25002-6

Index

A. Sheydvasser, *Linear Fractional Transformations*, Undergraduate Texts in Mathematics, https://doi.org/10.1007/978-3-031-25002-6