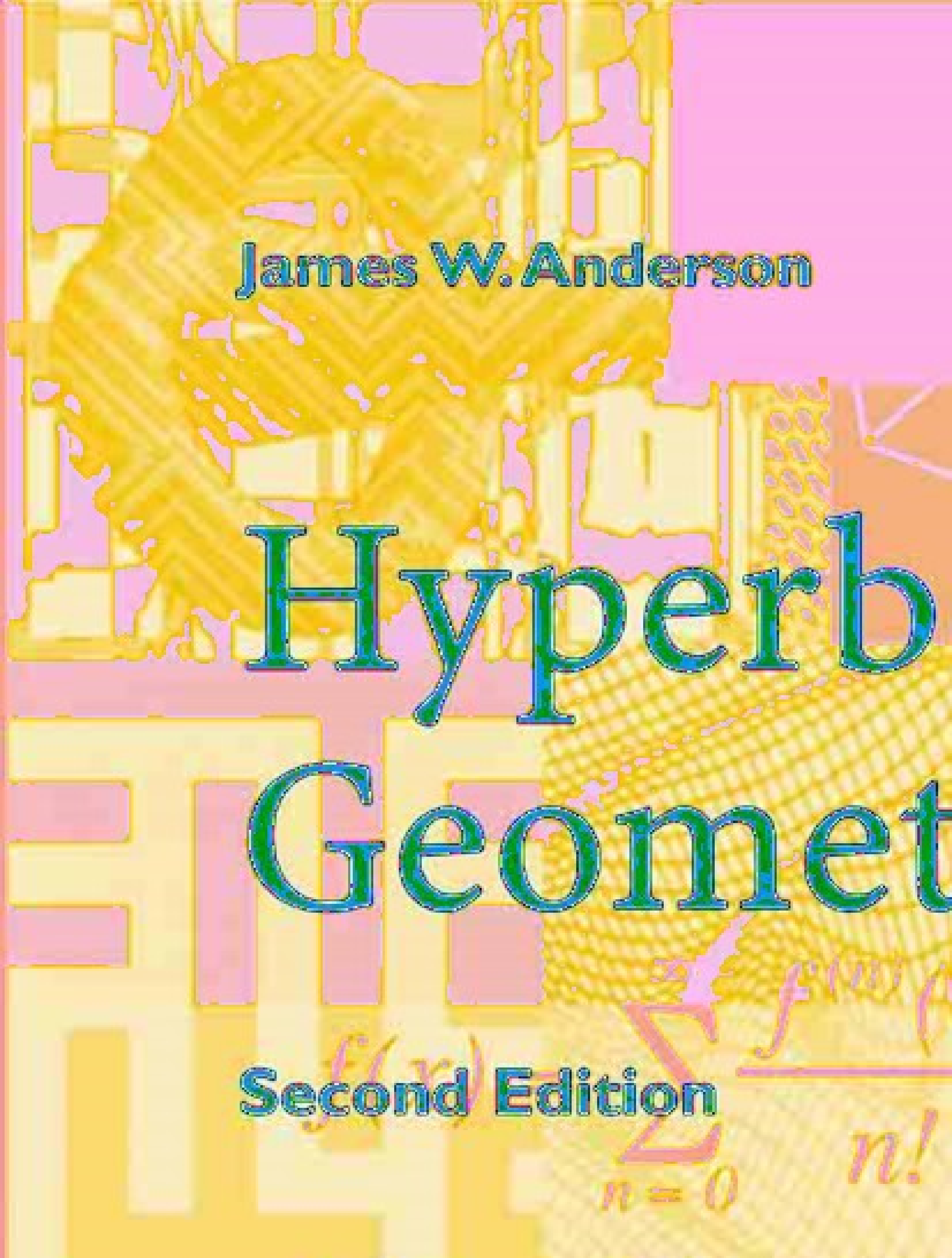

James W. Anderson

Hyperb Geomet

Second Edition

Springer Undergraduate Mathematics Series

Advisory Board

Other books in this series

James W. Anderson

Hyperbolic Geometry

Second Edition

With 21 Figures

Springer

James W. Anderson, BA, PhD
School of Mathematics, University of Southampton, Southampton SO17 1BJ, UK

Cover ilustration elements reproduced by kind permission of:
Aptech Systems, Inc., Publishers of the GAUSS Mathematical and Statistical System, 23804 S.E. Kent-Kangley Road, Maple Valley, WA 98038, USA. Tel: (206) 432-7855 Fax (206) 432-7832 email: info@aptech.com URL: www.aptech.com.
American Statistical Association: Chance Vol 8 No 1, 1995 article by KS and KW Heiner 'Tree Rings of the Northern Shawangunks' page 32 fig 2.
Springer-Verlag: Mathematica in Education and Research Vol 4 Issue 3 1995 article by Roman E Maeder, Beatrice Amrhein and Oliver Gloor 'Illustrated Mathematics: Visualization of Mathematical Objects' page 9 fig 11, originally published as a CD ROM 'Illustrated Mathematics' by TELOS: ISBN 0-387-14222-3, German edition by Birkhauser: ISBN 3-7643-5100-4.
Mathematica in Education and Research Vol 4 Issue 3 1995 article by Richard J Gaylord and Kazume Nishidate 'Traffic Engineering with Cellular Automata' page 35 fig 2. Mathematica in Education and Research Vol 5 Issue 2 1996 article by Michael Trott 'The Implicitization of a Trefoil Knot' page 14.
Mathematica in Education and Research Vol 5 Issue 2 1996 article by Lee de Cola 'Coins, Trees, Bars and Bells: Simulation of the Binomial Process' page 19 fig 3. Mathematica in Education and Research Vol 5 Issue 2 1996 article by Richard Gaylord and Kazume Nishidate 'Contagious Spreading' page 33 fig 1. Mathematica in Education and Research Vol 5 Issue 2 1996 article by Joe Buhler and Stan Wagon 'Secrets of the Madelung Constant' page 50 fig 1.

Mathematics Subject Classification (2000): 51-01

Anderson, James W., 1964–
Hyberbolic geometry. — 2nd ed. — (Springer undergraduate mathematics series)
1. Geometry, Hyperbolic
I. Title
516.9
ISBN 1852339349

Library of Congress Control Number: 2005923338

Springer Undergraduate Mathematics Series ISSN 1615-2085
ISBN 1-85233-934-9
Springer Science+Business Media
springeronline.com

The use of registered names, trademarks etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant laws and regulations and therefore free for general use.

The publisher makes no representation, express or implied, with regard to the accuracy of the information contained in this book and cannot accept any legal responsibility or liability for any errors or omissions that may be made.

Typesetting: Camera ready by the author
Printed in the United States of America
12/3830-543210   Printed on acid-free paper

Contents

Preamble to the Second Edition

Welcome to the second edition of *Hyperbolic Geometry*, and thanks for reading. I have tried to keep the basic structure of the book relatively unchanged, so that it can still be used by the reader either for self-study or as a classroom text. I have also tried to maintain the self-contained aspect of the book. A few new exercises and small bits of new material have been added to most chapters, and a few exercises and small bits of material have been removed. Overall, Chapters 1, 2, 3, and 5 are essentially the same as they were in the first edition.

In addition to this tinkering with exercises and material, there have been two major changes from the first edition of this book.

First, I have tightened the focus of Chapter 4 to just planar models of hyperbolic plane that arise from complex analysis. This has resulted in the introduction of some more advanced material from complex analysis, but not so much that the self-contained aspect of the book is seriously threatened. I have tried to make more clear the connections between planar hyperbolic geometry and complex analysis.

Second, I have changed Chapter 6 completely. Gone is the material on discrete subgroups of $\text{Möb}(\mathbb{H})$. In its place is an introduction to the hyperboloid model of the hyperbolic plane. Unfortunately, I did not feel that I had space to do justice to the Klein model as well, and so I haven't built the bridge from the Poincaré disc model to the hyperboloid model via the Klein model, but this has been done elsewhere by others. I close Chapter 6, and the book, with a very brief look at higher dimensional hyperbolic geometry.

The prerequisites for reading the book haven't significantly changed from the first edition. The book is written primarily for a third- or fourth-year undergraduate student who has encountered some calculus (univariate and multivariate), particularly the definition of arc-length, integration over regions in Euclidean space, and the change of variables theorem; some analysis, particularly continuity, open and closed sets in the plane, and infimum and supremum; and some basic complex analysis, such as the arithmetic of the complex numbers $\mathbb{C}$ and the basics of holomorphic functions.

I would like to close this introduction by adding some acknowledgements to the list given in the Preamble to the First Edition. I would like to thank Karen Borthwick at Springer for giving me the opportunity to write this second edition, and for being patient with me. I have continued to teach the course of Hyperbolic Geometry at the University of Southampton on which this book is based, and I would like to thank the students who have been a part of the course over the past several years, and who have pointed out the occasional mistake. The errors that remain are of course mine. And I would like to thank my wife Barbara, who once again put up with me through the final stages of writing.

Preamble to the First Edition

What you have in your hands is an introduction to the basics of planar hyperbolic geometry. Writing this book was difficult, not because I was at any point at a loss for topics to include, but rather because I continued to come across topics that I felt should be included in an introductory text. I believe that what has emerged from the process of writing gives a good feel for the geometry of the hyperbolic plane.

This book is written to be used either as a classroom text or as more of a self-study book, perhaps as part of a directed reading course. For that reason, I have included solutions to all the exercises. I have tried to choose the exercises to give reasonable coverage to the sorts of calculations and proofs that inhabit the subject. The reader should feel free to make up their own exercises, both proofs and calculations, and to make use of other sources.

I have also tried to keep the exposition as self-contained as possible, and to make as little use of mathematical machinery as possible. The book is written for a third or fourth year student who has encountered some Calculus, particularly the definition of arc-length, integration over regions in Euclidean space, and the change of variables theorem; some Analysis, particularly continuity, open and closed sets in the plane, and infimum and supremum; has a familiarity with Complex Numbers, as most of the book takes place in the complex plane $\mathbb{C}$, but need not have taken a class in Complex Analysis; and some Abstract Algebra, as we make use of some of the very basics from the theory of groups.

Non-Euclidean geometry in general, and hyperbolic geometry in particular, is an area of mathematics which has an interesting history and which is still being actively studied by researchers around the world. One reason for the continuing interest in hyperbolic geometry is that it touches on a number of different fields, including but not limited to Complex Analysis, Abstract Algebra and Group Theory, Number Theory, Differential Geometry, and Low-dimensional Topology.

This book is not written as an encyclopedic introduction to hyperbolic geometry but instead offers a single perspective. Specifically, I wanted to write a hyperbolic geometry book in which very little was assumed, and as much as possible was derived from following Klein's view that geometry, in this case hyperbolic geometry, consists of the study of those quantities invariant under a group. Consequently, I did not want to write down, without what I felt to be reasonable justification, the hyperbolic element of arc-length, or the group of hyperbolic isometries, but instead wanted them to arise as naturally as possible. And I think I have done that in this book.

There is a large number of topics I have chosen not to include, such as the hyperboloid and Klein models of the hyperbolic plane. Also, I have included nothing of the history of hyperbolic geometry and I have not taken the axiomatic approach to define the hyperbolic plane. One reason for these omissions is that there are already a number of excellent books on both the history of hyperbolic geometry and on the axiomatic approach, and I felt that I would not be able to add anything of note to what has already been done. There is an extensive literature on hyperbolic geometry. The interested reader is directed to the list of sources for Further Reading at the end of the book.

And now, a brief outline of the approach taken in this book. We first develop a model of the hyperbolic plane, namely the upper half-plane model $\mathbb{H}$, and define what we mean by a hyperbolic line in $\mathbb{H}$. We then try to determine a reasonable group of transformations of $\mathbb{H}$ that takes hyperbolic lines to hyperbolic lines, which leads us to spend some time studying the group Möb^+ of Möbius transformations and the general Möbius group Möb.

After determining the subgroup $\text{Möb}(\mathbb{H})$ of Möb preserving $\mathbb{H}$, we derive an invariant element of arc-length on $\mathbb{H}$. That is, we derive a means of calculating the hyperbolic length of a path $f : [a, b] \to \mathbb{H}$ in such a way that the hyperbolic length of a path is invariant under the action of $\text{Möb}(\mathbb{H})$, which is to say that the hyperbolic length of a path $f : [a, b] \to \mathbb{H}$ is equal to the hyperbolic length of its translate $\gamma \circ f : [a, b] \to \mathbb{H}$ for any element γ of $\text{Möb}(\mathbb{H})$. We are then able to define a natural metric on $\mathbb{H}$ in terms of the shortest hyperbolic length of a path joining a pair of points.

After exploring calculations of hyperbolic length, we move onto a discussion of convexity and of hyperbolic polygons, and then to the trigonometry of polygons in the hyperbolic plane and the three basic laws of trigonometry in the hyperbolic plane. We also determine how to calculate hyperbolic area, and state and prove the Gauss-Bonnet formula for hyperbolic polygons, which gives the hyperbolic area of a hyperbolic polygon in terms of its interior angles. In the course of this analysis, we introduce other models of the hyperbolic plane, particularly the Poincaré disc model $\mathbb{D}$. We close by describing and exploring very briefly what it means for a subgroup of Möb($\mathbb{H}$) to be well-behaved.

I would like to close this introduction with some acknowledgements. I would like to start by thanking Susan Hezlet for suggesting that I write this book, and David Ireland, who watched over its completion. Part of the writing of this book was done while I was visiting the Mathematics Department of Rice University during the 1998-1999 academic year, and I offer my thanks to the department there, particularly Frank Jones, who was chairman at the time and helped arrange my visit. This book is based on lectures from a class on hyperbolic geometry at the University of Southampton in the Fall terms of the 1996-97 and 1997-98 academic years, and I would like to thank the students in those classes, as well as the students at Rice whose sharp eyes helped in the final clean up of the text. The errors that remain are mine.

I would also like to thank all my mathematics teachers from over the years, particularly Ted Shifrin and Bernie Maskit; my parents, Wyatt and Margaret, and my sisters, Elizabeth and Karen, for all their love and support over the years; and to Barbara, who put up with me through the final stages of the writing.

1
The Basic Spaces

In this chapter, we set the stage for what is to come. Namely, we define the *upper half-plane model* $\mathbb{H}$ of the hyperbolic plane, which is where much of the action in this book takes place. We define *hyperbolic lines* and talk a bit about *parallelism.* To aid our construction of a reasonable group of transformations of $\mathbb{H}$, we expand our horizons to consider the *Riemann sphere* $\overline{\mathbb{C}}$ and close the chapter by considering how $\mathbb{H}$ sits as a subset of $\overline{\mathbb{C}}$.

1.1 A Model for the Hyperbolic Plane

We begin our investigation by describing a model of the hyperbolic plane. By a *model*, we mean a choice of an underlying space, together with a choice of how to represent basic geometric objects, such as points and lines, in this underlying space.

As we shall see over the course of the book, there are many possible models for the hyperbolic plane, each useful in its context. We focus our attention primarily, although not exclusively, on models of the hyperbolic plane whose underlying space is contained in the complex plane $\mathbb{C}$. We have chosen the models we work with and describe in this book for their convenience. To give as concrete a description of its geometry as possible, we begin by working in a single specific model.

The model of the hyperbolic plane we begin with is the *upper half-plane* model. The underlying space of this model is the upper half-plane $\mathbb{H}$ in the complex plane $\mathbb{C}$, defined to be

$$\mathbb{H} = \{z \in \mathbb{C} \mid \operatorname{Im}(z) > 0\}.$$

We use the usual notion of point that $\mathbb{H}$ inherits from $\mathbb{C}$. We also use the usual notion of angle that $\mathbb{H}$ inherits from $\mathbb{C}$; that is, the *angle* between two curves in $\mathbb{H}$ is defined to be the angle between the curves when they are considered to be curves in $\mathbb{C}$, which in turn is defined to be the angle between their tangent lines.

As we will define hyperbolic lines in $\mathbb{H}$ in terms of Euclidean lines and Euclidean circles in $\mathbb{C}$, we begin with a couple of calculations in $\mathbb{C}$.

Exercise 1.1

Express the equations of the Euclidean line $ax + by + c = 0$ and the Euclidean circle $(x-h)^2+(y-k)^2 = r^2$ in terms of the complex coordinate $z = x + iy$ in $\mathbb{C}$.

Exercise 1.2

Let $\mathbb{S}^1 = \{z \in \mathbb{C} \mid |z| = 1\}$ be the *unit circle* in $\mathbb{C}$. Let A be a Euclidean circle in $\mathbb{C}$ with Euclidean centre $re^{i\theta}$, $r > 1$, and Euclidean radius $s > 0$. Show that A is perpendicular to $\mathbb{S}^1$ if and only if $s = \sqrt{r^2 - 1}$.

We are now ready to define a *hyperbolic line* in $\mathbb{H}$.

Definition 1.1

There are two seemingly different types of *hyperbolic line*, both defined in terms of Euclidean objects in $\mathbb{C}$. One is the intersection of $\mathbb{H}$ with a Euclidean line in $\mathbb{C}$ perpendicular to the real axis $\mathbb{R}$ in $\mathbb{C}$. The other is the intersection of $\mathbb{H}$ with a Euclidean circle centred on the real axis $\mathbb{R}$.

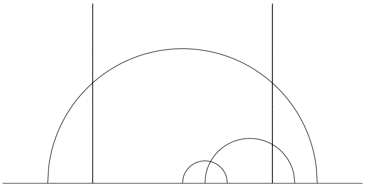

Figure 1.1: Hyperbolic lines in $\mathbb{H}$

Some examples of hyperbolic lines in $\mathbb{H}$ are shown in Figure 1.1.

We will see in Section 1.2 a way of unifying these two different types of hyperbolic line. For the moment, though, we content ourselves with an exploration of some basic properties of hyperbolic geometry with this definition of hyperbolic line.

Working in analogy with what we know from Euclidean geometry, there exists one property that hyperbolic lines in $\mathbb{H}$ should have, namely that there should always exist one and only one hyperbolic line passing through any pair of distinct points of $\mathbb{H}$. That this property holds in $\mathbb{H}$ with hyperbolic lines as defined above is a fairly straightforward calculation.

Proposition 1.2

For each pair p and q of distinct points in $\mathbb{H}$, there exists a unique hyperbolic line ℓ in $\mathbb{H}$ passing through p and q.

Proof

We begin by showing existence. There are two cases to consider. Suppose first that $\text{Re}(p) = \text{Re}(q)$. Then, the Euclidean line L given by the equation $L = \{z \in \mathbb{C} \mid \text{Re}(z) = \text{Re}(p)\}$ is perpendicular to the real axis and passes through both p and q. So, the hyperbolic line $\ell = \mathbb{H} \cap L$ is the desired hyperbolic line through p and q.

Suppose now that $\text{Re}(p) \neq \text{Re}(q)$. As the Euclidean line through p and q is no longer perpendicular to $\mathbb{R}$, we need to construct a Euclidean circle centred on the real axis $\mathbb{R}$ passing though p and q.

Let L_{pq} be the Euclidean line segment joining p and q, and let K be the perpendicular bisector of L_{pq}. Then, every Euclidean circle that passes through

p and q has its centre on K. As p and q have nonequal real parts, the Euclidean line K is not parallel to $\mathbb{R}$, and so K and $\mathbb{R}$ intersect at a unique point c.

Let A be the Euclidean circle centred at this point of intersection c with radius $|c-p|$, so that A passes through p. As c lies on K, we have that $|c-p| = |c-q|$, and so A passes through q. The intersection $\ell = \mathbb{H} \cap A$ is then the desired hyperbolic line passing through p and q.

The uniqueness of this hyperbolic line passing through p and q comes from the uniqueness of the Euclidean lines and Euclidean circles used in its construction. This completes the proof of Proposition 1.2. **QED**

We note here that the argument used to prove Proposition 1.2 contains more information. For any pair of distinct points p and q in $\mathbb{C}$ with nonequal real parts, there exists a unique Euclidean circle centred on $\mathbb{R}$ passing through p and q. The crucial point is that the centre of any Euclidean circle passing through p and q lies on the perpendicular bisector K of the Euclidean line segment L_{pq} joining p and q, and K is not parallel to $\mathbb{R}$.

As we have chosen the underlying space $\mathbb{H}$ for this model of the hyperbolic plane to be contained in $\mathbb{C}$, and as we have chosen to define hyperbolic lines in $\mathbb{H}$ in terms of Euclidean lines and Euclidean circles in $\mathbb{C}$, we can use whatever facts about Euclidean lines and Euclidean circles we already know to analyze the behaviour of hyperbolic lines. We have in effect given ourselves familiar coordinates on $\mathbb{H}$ to work with.

For instance, if ℓ is the hyperbolic line in $\mathbb{H}$ passing through p and q, we can express ℓ explicitly in terms of p and q. When p and q have equal real parts, we have already seen that $\ell = \mathbb{H} \cap L$, where L is the Euclidean line $L = \{z \in \mathbb{C} \mid \text{Re}(z) = \text{Re}(p)\}$. The expression of ℓ in terms of p and q in the case in which $\text{Re}(p) \neq \text{Re}(q)$ is left as an exercise.

Exercise 1.3

Let p and q be distinct points in $\mathbb{C}$ with nonequal real parts, and let A be the Euclidean circle centred on $\mathbb{R}$ and passing through p and q. Express the Euclidean centre c and the Euclidean radius r of A in terms of $\text{Re}(p)$, $\text{Im}(p)$, $\text{Re}(q)$, and $\text{Im}(q)$.

A legitimate question to raise at this point is whether hyperbolic geometry in $\mathbb{H}$, with this definition of hyperbolic line, is actually different from the usual

Euclidean geometry in $\mathbb{C}$ to which we are accustomed. The answer to this question is an emphatic Yes, hyperbolic geometry in $\mathbb{H}$ behaves very differently from Euclidean geometry in $\mathbb{C}$.

One way to see this difference is to consider the behaviour of parallel lines. Recall that Euclidean lines in $\mathbb{C}$ are parallel if and only if they are disjoint, and we adopt this definition in the hyperbolic plane as well.

Definition 1.3

Two hyperbolic lines in $\mathbb{H}$ are *parallel* if they are disjoint.

In Euclidean geometry, parallel lines exist, and in fact, if L is a Euclidean line and if a is a point in $\mathbb{C}$ not on L, then there exists one and only one Euclidean line K through a that is parallel to L.

Additionally, in Euclidean geometry, parallel lines are equidistant; that is, if L and K are parallel Euclidean lines and if a and b are points on L, then the Euclidean distance from a to K is equal to the Euclidean distance from b to K.

In hyperbolic geometry, parallelism behaves much differently. Although we do not yet have a means of measuring hyperbolic distance, we can consider parallel hyperbolic lines qualitatively.

Theorem 1.4

Let ℓ be a hyperbolic line in $\mathbb{H}$, and let p be a point in $\mathbb{H}$ not on ℓ. Then, there exist infinitely many distinct hyperbolic lines through p that are parallel to ℓ.

Proof

There are two cases to consider. First, suppose that ℓ is contained in a Euclidean line L. As p is not on L, there exists a Euclidean line K through p that is parallel to L. As L is perpendicular to $\mathbb{R}$, we have that K is perpendicular to $\mathbb{R}$ as well. So, one hyperbolic line in $\mathbb{H}$ through p and parallel to ℓ is the intersection $\mathbb{H} \cap K$.

To construct another hyperbolic line through p and parallel to ℓ, take a point x on $\mathbb{R}$ between K and L, and let A be the Euclidean circle centred on $\mathbb{R}$ that passes through x and p. We know that such a Euclidean circle A exists because $\mathrm{Re}(x) \neq \mathrm{Re}(p)$.

By construction, A is disjoint from L, and so the hyperbolic line $\mathbb{H} \cap A$ is disjoint from ℓ. That is, $\mathbb{H} \cap A$ is a second hyperbolic line through p that is parallel to ℓ. As there exist infinitely many points on $\mathbb{R}$ between K and L, this construction gives infinitely many distinct hyperbolic lines through p and parallel to ℓ. A picture of this phenomenon is given in Figure 1.2.

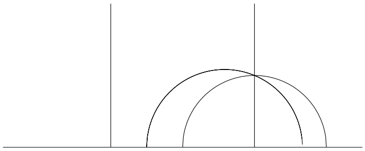

Figure 1.2: Several parallel hyperbolic lines

Exercise 1.4

Give an explicit description of two hyperbolic lines in $\mathbb{H}$ through i and parallel to the hyperbolic line $\ell = \mathbb{H} \cap \{z \in \mathbb{C} \mid \mathrm{Re}(z) = 3\}$.

Now, suppose that ℓ is contained in a Euclidean circle A. Let D be the Euclidean circle that is concentric to A and that passes through p. As concentric circles are disjoint and have the same centre, one hyperbolic line through p and parallel to ℓ is the intersection $\mathbb{H} \cap D$.

To construct a second hyperbolic line through p and parallel to ℓ, take any point x on $\mathbb{R}$ between A and D. Let E be the Euclidean circle centred on $\mathbb{R}$ that passes through x and p. Again, by construction, E and A are disjoint, and so $\mathbb{H} \cap E$ is a hyperbolic line through p parallel to ℓ.

As above, because there exist infinitely many points on $\mathbb{R}$ between A and D, there exist infinitely many distinct hyperbolic lines through p parallel to ℓ. A picture of this phenomenon is given in Figure 1.3.

This completes the proof of Theorem 1.4. **QED**

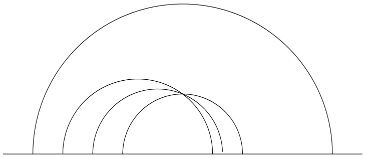

Figure 1.3: Several parallel hyperbolic lines

Exercise 1.5

Give an explicit description of two hyperbolic lines in $\mathbb{H}$ through i and parallel to the hyperbolic line $\ell = \mathbb{H} \cap A$, where A is the Euclidean circle with Euclidean centre -2 and Euclidean radius 1.

We now have a model to play with. The bulk of this book is spent exploring this particular model of the hyperbolic plane, although we do spend some time developing and exploring other models. Although we focus primarily on models of hyperbolic plane for which the underlying space is a subset of $\mathbb{C}$, we note that there are many others. We explore one such other model, the hyperboloid model, in Section 6.1.

We close this section with a few words to put what we are doing in a historical context. We are proceeding almost completely backward in our development of hyperbolic geometry from the historical development of the subject. A much more common approach is to begin with the axiomization of Euclidean geometry. One of the axioms is the statement about parallel lines mentioned above, namely, that given a line Euclidean L and a point p not on L, there exists a unique Euclidean line through p and parallel to L. This axiom is often referred to as the parallel postulate; the form we give here is credited to Playfair.

Hyperbolic geometry is then defined using the same set of axioms as Euclidean geometry, with the hyperbolic variant of the parallel postulate, namely, that given a hyperbolic line ℓ and a point p not on ℓ, there exist at least two hyperbolic lines through p and parallel to ℓ.

It is then shown that the upper half-plane model, with hyperbolic lines as we have defined them, is a model of the resulting non-Euclidean geometry. For instance, see the books of Stahl [31] and Greenberg [16].

In this book, we are less concerned with the axiomatic approach to hyperbolic geometry, preferring to make use of the fact that we have reasonable coordinates in the upper half-plane $\mathbb{H}$, which allow us to calculate fairly directly.

Our first major task is to determine whether we have enough information in this description of hyperbolic geometry to define the notions of hyperbolic length, hyperbolic distance, and hyperbolic area in $\mathbb{H}$. We do this using the group of transformations of $\mathbb{H}$ taking hyperbolic lines to hyperbolic lines.

For the history of the subject, see Rosenfeld [29], Greenberg [16], and Bonola [11]. Also of interest are the translations by Stillwell [33] of some of the original papers of Beltrami, Klein, and Poincaré, all of whom were instrumental in the development of hyperbolic geometry.

A far from complete list of books discussing various aspects of the hyperbolic geometry of the plane, listed in no particular order, are the books of Trudeau [36], Stahl [31], Rédei [28], Wylie [37], Iversen [22], Coxeter [14], Kelly and Matthews [24], Thurston [35], Fenchel [15], and Pedoe [27], as well as the articles of Beardon [10], Stillwell [32], and von Helmholz [19]. We will refer to some of these at various points of the text, where they are especially relevant.

1.2 The Riemann Sphere $\overline{\mathbb{C}}$

To determine a reasonable group of transformations of $\mathbb{H}$ that take hyperbolic lines to hyperbolic lines, we first fulfil our earlier promise of unifying the two seemingly different types of hyperbolic line, namely, those contained in a Euclidean line and those contained in a Euclidean circle. We take as our stepping off point the observation that a Euclidean circle can be obtained from a Euclidean line by adding a single point.

To be explicit, let $\mathbb{S}^1$ be the unit circle in $\mathbb{C}$, and consider the function

$$\xi : \mathbb{S}^1 - \{i\} \to \mathbb{R}$$

defined as follows: given a point z in $\mathbb{S}^1 - \{i\}$, let K_z be the Euclidean line passing through i and z, and set $\xi(z) = \mathbb{R} \cap K_z$. This function is well defined, because K_z and $\mathbb{R}$ intersect in a unique point as long as $\mathrm{Im}(z) \neq 1$. See Figure 1.4.

This function ξ is referred to as *stereographic projection.* In terms of the usual cartesian coordinates on the plane, the real axis $\mathbb{R}$ in $\mathbb{C}$ corresponds to the x-axis, and so $\xi(z)$ is the x-intercept of K_z. Calculating, we see that K_z has slope

$$m = \frac{\mathrm{Im}(z) - 1}{\mathrm{Re}(z)}$$

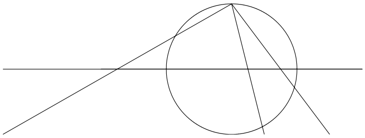

Figure 1.4: Stereographic projection

and y-intercept 1. Hence, the equation for K_z is

$$y - 1 = \frac{\mathrm{Im}(z) - 1}{\mathrm{Re}(z)}\, x.$$

In particular, the x-intercept of K_z is

$$\xi(z) = \frac{\mathrm{Re}(z)}{1 - \mathrm{Im}(z)}.$$

Exercise 1.6

Give an explicit formula for $\xi^{-1} : \mathbb{R} \to \mathbb{S}^1 - \{i\}$.

Exercise 1.7

Consider the three points $z_k = \exp\left(\frac{2\pi k}{3} i\right)$, $0 \le k \le 2$, of $\mathbb{S}^1$ that form the vertices of an equilateral triangle in $\mathbb{C}$. Calculate their images under ξ.

In fact, ξ is a bijection between $\mathbb{S}^1 - \{i\}$ and $\mathbb{R}$. Geometrically, this follows from the fact that a pair of distinct points in $\mathbb{C}$ determines a unique Euclidean line: If z and w are points of $\mathbb{S}^1 - \{i\}$ for which $\xi(z) = \xi(w)$, then the Euclidean lines K_z and K_w both pass through both i and $\xi(z) = \xi(w)$, which forces the two lines K_z and K_w to be equal, and so $z = w$.

As we obtain $\mathbb{R}$ from $\mathbb{S}^1$ by removing a single point of $\mathbb{S}^1$, namely, i, we can think of constructing the Euclidean circle $\mathbb{S}^1$ by starting with the Euclidean line $\mathbb{R}$ and adding a single point.

Motivated by this, one possibility for a space that contains $\mathbb{H}$ and in which the two seemingly different types of hyperbolic line are unified is the space that is obtained from $\mathbb{C}$ by adding a single point. This is the classical construction from complex analysis of the *Riemann sphere* $\overline{\mathbb{C}}$.

As a set of points, the Riemann sphere is the union

$$\overline{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$$

of the complex plane $\mathbb{C}$ with a point not contained in $\mathbb{C}$, which we denote ∞. To explore the basic properties of $\overline{\mathbb{C}}$, we first define what it means for a subset of $\overline{\mathbb{C}}$ to be open.

We begin by recalling that a set X in $\mathbb{C}$ is *open* if for each $z \in X$, there exists some $\varepsilon > 0$ so that $U_\varepsilon(z) \subset X$, where

$$U_\varepsilon(z) = \{w \in \mathbb{C} \mid |w - z| < \varepsilon\}$$

is the open Euclidean disc of radius ε centred at z. (In general, ε will depend on both z and X.)

A set X in $\mathbb{C}$ is *closed* if its complement $\mathbb{C} - X$ in $\mathbb{C}$ is open.

A set X in $\mathbb{C}$ is *bounded* if there exists some constant $\varepsilon > 0$ so that $X \subset U_\varepsilon(0)$.

Exercise 1.8

Prove that $\mathbb{H}$ is open in $\mathbb{C}$. For each point z of $\mathbb{H}$, calculate the maximum ε so that $U_\varepsilon(z)$ is contained in $\mathbb{H}$.

To extend this definition to $\overline{\mathbb{C}}$, we need only define what $U_\varepsilon(z)$ means for each point z of $\overline{\mathbb{C}}$ and each $\varepsilon > 0$. As all but one point of $\overline{\mathbb{C}}$ lies in $\mathbb{C}$, it makes sense to use the definition we had above wherever possible, and so for each point z of $\mathbb{C}$, we define

$$U_\varepsilon(z) = \{w \in \mathbb{C} \mid |w - z| < \varepsilon\}.$$

It remains only to define $U_\varepsilon(\infty)$, which we take to be

$$U_\varepsilon(\infty) = \{w \in \mathbb{C} \mid |w| > \varepsilon\} \cup \{\infty\}.$$

Definition 1.5

A set X in $\overline{\mathbb{C}}$ is *open* if for each point x of X, there exists some $\varepsilon > 0$ (which may depend on x and X) so that $U_\varepsilon(x) \subset X$.

One immediate consequence of this definition is that if D is an open set in $\mathbb{C}$, then D is also open in $\overline{\mathbb{C}}$. That is, we are not distorting $\mathbb{C}$ by viewing it as a subset of $\overline{\mathbb{C}}$. For example, because $\mathbb{H}$ is an open subset of $\mathbb{C}$, by Exercise 1.8, we immediately have that $\mathbb{H}$ is open in $\overline{\mathbb{C}}$.

As another example, we show that the set $E = \{z \in \mathbb{C} \,|\, |z| > 1\} \cup \{\infty\} = U_1(\infty)$ is open in $\overline{\mathbb{C}}$. We need to show that for each point z of E, there exists some $\varepsilon > 0$ so that $U_\varepsilon(z) \subset E$. As $E = U_1(\infty)$, we can find a suitable ε for $z = \infty$, namely, $\varepsilon = 1$. For a point z of $E - \{\infty\}$, note that the Euclidean distance from z to $\partial E = \mathbb{S}^1$ is $|z| - 1$, and so we have that $U_\varepsilon(z) \subset E$ for any $0 < \varepsilon < |z| - 1$.

On the other hand, the unit circle $\mathbb{S}^1$ in $\mathbb{C}$ is not open. No matter which point z of $\mathbb{S}^1$ and which $\varepsilon > 0$ we consider, we have that $U_\varepsilon(z)$ does not lie in $\mathbb{S}^1$, as $U_\varepsilon(z)$ necessarily contains the point $(1 + \frac{1}{2}\varepsilon)z$ whose modulus is $|(1 + \frac{1}{2}\varepsilon)z| = (1 + \frac{1}{2}\varepsilon)|z| = 1 + \frac{1}{2}\varepsilon \neq 1$.

Definition 1.6

A set X in $\overline{\mathbb{C}}$ is *closed* if its complement $\overline{\mathbb{C}} - X$ in $\overline{\mathbb{C}}$ is open.

For example, the unit circle $\mathbb{S}^1$ is closed in $\overline{\mathbb{C}}$, because its complement is the union

$$\overline{\mathbb{C}} - \mathbb{S}^1 = U_1(0) \cup U_1(\infty),$$

and it follows from the definition that the union of open sets is open.

Exercise 1.9

Prove that if K is a closed and bounded subset of $\mathbb{C}$, then $X = (\mathbb{C} - K) \cup \{\infty\}$ is open in $\overline{\mathbb{C}}$. Conversely, prove that every open subset of $\overline{\mathbb{C}}$ is either an open subset of $\mathbb{C}$ or is the complement in $\overline{\mathbb{C}}$ of a closed and bounded subset of $\mathbb{C}$.

This construction of the Riemann sphere $\overline{\mathbb{C}}$ from the complex plane $\mathbb{C}$ is an example of a more general construction, known as the *one-point compactification.* For more information on the one-point compactification and related topological constructions, the interested reader is referred to a book on point-set topology, such as Munkres [26].

One major use of open sets is to define *convergence.* Convergence in $\overline{\mathbb{C}}$ is analogous to convergence in $\mathbb{C}$; that is, a sequence $\{z_n\}$ of points in $\overline{\mathbb{C}}$ *converges* to a point z of $\overline{\mathbb{C}}$ if for each $\varepsilon > 0$, there exists N so that $z_n \in U_\varepsilon(z)$ for all $n > N$.

Exercise 1.10

Prove that $\left\{z_n = \frac{1}{n} \,|\, n \in \mathbb{N}\right\}$ converges to 0 in $\overline{\mathbb{C}}$, and that $\{w_n = n \,|\, n \in \mathbb{N}\}$ converges to ∞ in $\overline{\mathbb{C}}$.

Let X be a subset of $\overline{\mathbb{C}}$. Define the *closure* $\overline{X}$ of X in $\overline{\mathbb{C}}$ to be the set

$$\overline{X} = \{z \in \overline{\mathbb{C}} \,|\, U_\varepsilon(z) \cap X \neq \emptyset \text{ for all } \varepsilon > 0\}.$$

Note that every point $x \in X$ lies in $\overline{X}$, because $\{x\} \subset U_\varepsilon(x) \cap X$ for every $\varepsilon > 0$. There may be points in $\overline{X}$ other than the points of X.

As an example, if $X \subset \overline{\mathbb{C}}$ and if $\{x_n\}$ is a sequence of points of X converging to a point x of $\overline{\mathbb{C}}$, then x is necessarily a point of $\overline{X}$.

Exercise 1.11

Determine the closure in $\overline{\mathbb{C}}$ of $X = \left\{\frac{1}{n} \,|\, n \in \mathbb{Z} - \{0\}\right\}$ and of $Y = \mathbb{Q} + \mathbb{Q}i = \{a + ib|\, a, b \in \mathbb{Q}\}$.

Exercise 1.12

If X is a subset of $\overline{\mathbb{C}}$, prove that $\overline{X}$ is closed in $\overline{\mathbb{C}}$.

We are now ready to unify the two notions of Euclidean line and Euclidean circle in $\mathbb{C}$.

Definition 1.7

A *circle in* $\overline{\mathbb{C}}$ is either a Euclidean circle in $\mathbb{C}$ or the union of a Euclidean line in $\mathbb{C}$ with $\{\infty\}$.

That is, we use the point ∞, which we adjoined to $\mathbb{C}$ to obtain $\overline{\mathbb{C}}$, to be the point we add to each Euclidean line to get a circle.

As a bit of notation, for a Euclidean line L in $\mathbb{C}$, let $\overline{L} = L \cup \{\infty\}$ be the circle in $\overline{\mathbb{C}}$ containing L. For example, the *extended real axis* $\overline{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$ is the circle in $\overline{\mathbb{C}}$ containing the real axis $\mathbb{R}$ in $\mathbb{C}$.

Note that this notation for the circle in $\overline{\mathbb{C}}$ containing the Euclidean line L agrees with our earlier notation for the closure of a subset of $\overline{\mathbb{C}}$, as the closure in $\overline{\mathbb{C}}$ of a Euclidean line L in $\mathbb{C}$ is exactly $\overline{L} = L \cup \{\infty\}$.

As might be guessed, there is a generalization of stereographic projection to the Riemann sphere $\overline{\mathbb{C}}$ and the complex plane $\mathbb{C}$. Identify $\mathbb{C}$ with the x_1x_2-plane in $\mathbb{R}^3$, where the coordinates on $\mathbb{R}^3$ are (x_1, x_2, x_3), by identifying the point $z = x + iy$ in $\mathbb{C}$ with the point $(x, y, 0)$ in $\mathbb{R}^3$. Let $\mathbb{S}^2$ be the unit sphere in $\mathbb{R}^3$; that is

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

with north pole $\mathrm{N} = (0, 0, 1)$.

Consider the function $\xi : \mathbb{S}^2 - \{\mathrm{N}\} \to \mathbb{C}$ defined as follows: For each point P of $\mathbb{S}^2 - \{\mathrm{N}\}$, let L_{P} be the Euclidean line in $\mathbb{R}^3$ passing through N and P, and define $\xi(\mathrm{P})$ to be the point of intersection $L_{\mathrm{P}} \cap \mathbb{C}$.

Exercise 1.13

Write out explicit formulae for both ξ and its inverse $\xi^{-1} : \mathbb{C} \to \mathbb{S}^2 - \{\mathrm{N}\}$.

The bijectivity of ξ follows from the fact that we can write down an explicit expression for ξ^{-1}. We could also argue geometrically, as we did for stereographic projection from $\mathbb{S}^1 - \{i\}$ to $\mathbb{R}$.

We can also describe circles in $\overline{\mathbb{C}}$ as the sets of solutions to equations in $\overline{\mathbb{C}}$. Recall that we show in the solution to Exercise 1.1 that every Euclidean circle in $\mathbb{C}$ can be described as the set of solutions of an equation of the form

$$\alpha z\overline{z} + \beta z + \overline{\beta}\overline{z} + \gamma = 0,$$

where α, $\gamma \in \mathbb{R}$, $\alpha \neq 0$, and $\beta \in \mathbb{C}$, and that every Euclidean line in $\mathbb{C}$ can be described as the set of solutions of an equation of the form

$$\beta z + \overline{\beta}\overline{z} + \gamma = 0,$$

where $\gamma \in \mathbb{R}$ and $\beta \in \mathbb{C}$.

Combining these, we see that every circle in $\overline{\mathbb{C}}$ can be described as the set of solutions in $\overline{\mathbb{C}}$ to an equation of the form

$$\alpha z\overline{z} + \beta z + \overline{\beta}\overline{z} + \gamma = 0,$$

where α, $\gamma \in \mathbb{R}$, and $\beta \in \mathbb{C}$.

One subtlety needs considered here, namely, the question of how we consider whether ∞ is a solution of such an equation. For an equation of the form

$$\beta z + \overline{\beta}\overline{z} + \gamma = 0,$$

where $\gamma \in \mathbb{R}$ and $\beta \in \mathbb{C}$, we may consider ∞ to be a solution *by continuity.* That is, there is a sequence $\{z_n\}$ of points in $\mathbb{C}$ that satisfies this equation and that converges to ∞ in $\overline{\mathbb{C}}$. Specifically, let w_0 and w_1 be two distinct solutions, so that every linear combination of the form $w_0 + t(w_1 - w_0)$, $t \in \mathbb{R}$, is also a solution. Consider the sequence

$$\{z_n = w_0 + n(w_1 - w_0) \mid n \in \mathbb{N}\}.$$

For any choice of w_0 and w_1, $w_0 \neq w_1$, we have that

$$\beta z_n + \overline{\beta}\overline{z_n} + \gamma = 0,$$

and that $\{z_n\}$ converges to ∞ in $\overline{\mathbb{C}}$.

However, for an equation of the form

$$\alpha z\overline{z} + \beta z + \overline{\beta}\overline{z} + \gamma = 0,$$

where α, $\gamma \in \mathbb{R}$, $\alpha \neq 0$, and $\beta \in \mathbb{C}$, we cannot view ∞ as a solution by continuity. This follows immediately from the fact that we can rewrite this equation as

$$\alpha z\overline{z} + \beta z + \overline{\beta}\overline{z} + \gamma = \alpha \left| z + \frac{\overline{\beta}}{\alpha} \right|^2 + \gamma - \frac{|\beta|^2}{\alpha} = 0.$$

However, if $\{z_n\}$ is any sequence of points in $\overline{\mathbb{C}}$ converging to ∞, then

$$\lim_{n\to\infty} (\alpha z_n\overline{z_n} + \beta z_n + \overline{\beta}\overline{z_n} + \gamma) = \lim_{n\to\infty} \left(\alpha \left| z_n + \frac{\overline{\beta}}{\alpha} \right|^2 + \gamma - \frac{|\beta|^2}{\alpha} \right) = \infty.$$

Therefore, z_n cannot lie on the circle in $\overline{\mathbb{C}}$

$$A = \{z \in \mathbb{C} \mid \alpha z\overline{z} + \beta z + \overline{\beta}\overline{z} + \gamma = 0\}$$

for n large, and so we cannot consider ∞ to be a point of A.

As we now have a definition of what it means for a subset of $\overline{\mathbb{C}}$ to be open, we can define what it means for a function $f : \overline{\mathbb{C}} \to \overline{\mathbb{C}}$ to be continuous, this time in analogy with the usual definition of continuity of functions from $\mathbb{R}$ to $\mathbb{R}$.

Definition 1.8

A function $f : \overline{\mathbb{C}} \to \overline{\mathbb{C}}$ is *continuous at* $z \in \overline{\mathbb{C}}$ if for each $\varepsilon > 0$, there exists $\delta > 0$ (depending on z and ε) so that $w \in U_\delta(z)$ implies that $f(w) \in U_\varepsilon(f(z))$. A function $f : \overline{\mathbb{C}} \to \overline{\mathbb{C}}$ is *continuous on* $\overline{\mathbb{C}}$ if it is continuous at every point $z \in \overline{\mathbb{C}}$.

One advantage to generalizing this definition of continuity is that we may use exactly the same proofs as with functions from $\mathbb{R}$ to $\mathbb{R}$ to show that constant functions from $\overline{\mathbb{C}}$ to $\overline{\mathbb{C}}$ are continuous, as are products and quotients (when they are defined), sums and differences (when they are defined), and compositions of continuous functions.

However, there exist some slight differences between functions from $\mathbb{R}$ to $\mathbb{R}$ and functions from $\overline{\mathbb{C}}$ to $\overline{\mathbb{C}}$, which arise from the presence of the point ∞. Consider the following example.

Proposition 1.9

The function $J : \overline{\mathbb{C}} \to \overline{\mathbb{C}}$ defined by

$$J(z) = \frac{1}{z} \text{ for } z \in \mathbb{C} - \{0\}, \quad J(0) = \infty, \text{ and } J(\infty) = 0$$

is continuous on $\overline{\mathbb{C}}$.

Proof

To see that J is continuous at 0, take $\varepsilon > 0$ to be given. As we have that $J(0) = \infty$, we need to show that there exists some $\delta > 0$ so that

$$J(U_\delta(0)) \subset U_\varepsilon(J(0)) = U_\varepsilon(\infty).$$

Take $\delta = \frac{1}{\varepsilon}$. For each $w \in U_\delta(0) - \{0\}$, we have that

$$|J(w)| = \frac{1}{|w|} > \frac{1}{\delta} = \varepsilon,$$

and so $J(w) \in U_\varepsilon(\infty)$. As we have that $J(0) = \infty \in U_\varepsilon(\infty)$ by definition, we see that J is continuous at 0.

The argument that J is continuous at ∞ is similar to the argument that J is continuous at 0. Again, given $\varepsilon > 0$, we take $\delta = \frac{1}{\varepsilon}$. Then, for each $w \in U_\delta(\infty) - \{\infty\}$, we have that

$$|J(w)| = \frac{1}{|w|} < \frac{1}{\delta} = \varepsilon,$$

and so $J(w) \in U_\varepsilon(0)$. We have that $J(\infty) = 0 \in U_\varepsilon(0)$ by definition, and so J is continuous at ∞.

To complete the proof, let $z \in \mathbb{C} - \{0\}$ be any point, and let $\varepsilon > 0$ be given. We need to find $\delta > 0$ so that $w \in U_\delta(z)$ implies that $J(w) \in U_\varepsilon(J(z))$. Let $\varepsilon' = \min(\varepsilon, \frac{1}{2|z|})$, so that $U_{\varepsilon'}(z)$ does not contain 0.

For any $\xi \in U_{\varepsilon'}(J(z))$, we have that

$$|\xi| < |J(z)| + \varepsilon' = \frac{1}{|z|} + \varepsilon'.$$

As $\varepsilon' \leq \frac{1}{2|z|}$, we have that

$$|\xi| < \frac{3}{2|z|}.$$

Writing $\xi = \frac{1}{w}$, we have that

$$\frac{1}{|w|} < \frac{3}{2|z|}, \text{ and so } \frac{1}{|zw|} < \frac{3}{2|z|^2}.$$

So, set $\delta = \frac{2}{3}\varepsilon'|z|^2$.

For $|z - w| < \delta$, we then have that

$$|J(z) - J(w)| = \left| \frac{1}{z} - \frac{1}{w} \right| = \frac{|z-w|}{|zw|} < \frac{2}{3}\varepsilon'|z|^2 \frac{3}{2|z|^2} = \varepsilon'.$$

As $\varepsilon' \leq \varepsilon$, we have that J is continuous at $z \in \mathbb{C} - \{0\}$. This completes the proof of Proposition 1.9. **QED**

Exercise 1.14

Let $g(z)$ be a polynomial of degree at least one. Prove that the function $f : \overline{\mathbb{C}} \to \overline{\mathbb{C}}$, defined by

$$f(z) = g(z) \text{ for } z \in \mathbb{C} \text{ and } f(\infty) = \infty,$$

is continuous on $\overline{\mathbb{C}}$.

One useful property, and indeed a defining property, of continuous functions is that they preserve convergent sequences. That is, if $f : \overline{\mathbb{C}} \to \overline{\mathbb{C}}$ is continuous and if $\{x_n\}$ is a sequence in $\overline{\mathbb{C}}$ converging to x, then $\{f(x_n)\}$ converges to $f(x)$.

There is a class of continuous functions from $\overline{\mathbb{C}}$ to itself that are especially well behaved.

Definition 1.10

A function $f : \overline{\mathbb{C}} \to \overline{\mathbb{C}}$ is a *homeomorphism* if f is a bijection and if both f and f^{-1} are continuous.

We have already seen one example of a homeomorphism of $\overline{\mathbb{C}}$.

Proposition 1.11

The function $J : \overline{\mathbb{C}} \to \overline{\mathbb{C}}$ defined by

$$J(z) = \frac{1}{z} \text{ for } z \in \mathbb{C} - \{0\}, \quad J(0) = \infty, \text{ and } J(\infty) = 0$$

is a homeomorphism of $\overline{\mathbb{C}}$.

Proof

As $J \circ J(z) = z$ for all $z \in \overline{\mathbb{C}}$, we immediately have that J is bijective: To see that J is injective, suppose that there exist points z and w for which $J(z) = J(w)$, and note that $z = J(J(z)) = J(J(w)) = w$; to see that J is surjective, note that for any $z \in \overline{\mathbb{C}}$, we have that $z = J(J(z))$.

Moreover, as $J^{-1}(z) = J(z)$ for all $z \in \overline{\mathbb{C}}$ and as J is continuous, by Proposition 1.9, we have that J^{-1} is continuous. This completes the proof of Proposition 1.11. **QED**

The homeomorphisms of $\overline{\mathbb{C}}$ are the transformations of $\overline{\mathbb{C}}$ that are of most interest to us, so set

$$\text{Homeo}(\overline{\mathbb{C}}) = \{f : \overline{\mathbb{C}} \to \overline{\mathbb{C}} \mid f \text{ is a homeomorphism}\}.$$

By definition, the inverse of a homeomorphism is again a homeomorphism. Also, the composition of two homeomorphisms is again a homeomorphism, because the composition of bijections is again a bijection and because the composition of continuous functions is again continuous. As the identity homeomorphism $f : \overline{\mathbb{C}} \to \overline{\mathbb{C}}$ given by $f(z) = z$ is a homeomorphism, we have that $\text{Homeo}(\overline{\mathbb{C}})$ is a group.

Exercise 1.15

Let $g(z)$ be a polynomial. Prove that the function $f : \overline{\mathbb{C}} \to \overline{\mathbb{C}}$, defined by

$$f(z) = g(z) \text{ for } z \in \mathbb{C} \text{ and } f(\infty) = \infty,$$

is a homeomorphism if and only if the degree of g is one.

Exercise 1.16

A subset X of $\overline{\mathbb{C}}$ is *dense* if $\overline{X} = \overline{\mathbb{C}}$. Prove that if X is dense in $\overline{\mathbb{C}}$ and if $f : \overline{\mathbb{C}} \to \overline{\mathbb{C}}$ is a continuous function for which $f(x) = x$ for all x in X, then $f(z) = z$ for all $z \in \overline{\mathbb{C}}$.

1.3 The Boundary at Infinity of $\mathbb{H}$

In Section 1.2, we define a circle in the Riemann sphere $\overline{\mathbb{C}}$ to be either a Euclidean circle in $\mathbb{C}$ or the union of a Euclidean line in $\mathbb{C}$ with $\{\infty\}$. We also have several examples of circles in $\overline{\mathbb{C}}$, including the *unit circle* $\mathbb{S}^1$ in $\mathbb{C}$ and the *extended real axis* $\overline{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$.

In particular, the complement of a circle in $\overline{\mathbb{C}}$ has two components. For $\mathbb{S}^1$, the components of $\overline{\mathbb{C}} - \mathbb{S}^1$ are the Euclidean disc $\mathbb{D} = U_1(0)$ and the Euclidean disc $U_1(\infty)$, whereas for $\overline{\mathbb{R}}$ the components of $\overline{\mathbb{C}} - \overline{\mathbb{R}}$ are the upper half-plane $\mathbb{H}$ and the lower half-plane $\{z \in \mathbb{C} \mid \mathrm{Im}(z) < 0\}$.

Definition 1.12

Define a *disc in* $\overline{\mathbb{C}}$ D to be one of the components of the complement in $\overline{\mathbb{C}}$ of a circle in $\overline{\mathbb{C}}$ A. For such D and A, we refer to A as the *circle determining the disc* D.

Note that every disc in $\overline{\mathbb{C}}$ determines a unique circle in $\overline{\mathbb{C}}$ and that every circle in $\overline{\mathbb{C}}$ determines two disjoint discs in $\overline{\mathbb{C}}$.

For the remainder of this section, we focus our attention on one particular disc in $\overline{\mathbb{C}}$, namely, $\mathbb{H}$, and the circle in $\overline{\mathbb{C}}$ determining it, namely, $\overline{\mathbb{R}}$. We refer to $\overline{\mathbb{R}}$ as the *boundary at infinity* of $\mathbb{H}$, and we refer to points of $\overline{\mathbb{R}}$ as *points at infinity* of $\mathbb{H}$. The reason we use this term will be explained in Section 3.4, after we have developed a means of measuring hyperbolic distance in $\mathbb{H}$.

More generally, for any set X in $\mathbb{H}$, we can make sense of the notion of the *boundary at infinity* of X. Specifically, we form the closure $\overline{X}$ of X in $\overline{\mathbb{C}}$, and then define the *boundary at infinity of* X to be the intersection $\overline{X} \cap \overline{\mathbb{R}}$ of $\overline{X}$ with the boundary at infinity $\overline{\mathbb{R}}$ of $\mathbb{H}$.

As an example, let ℓ be a hyperbolic line in $\mathbb{H}$, and suppose that ℓ is contained in the circle A in $\overline{\mathbb{C}}$. Then, the boundary at infinity of ℓ is the pair of points contained in the intersection $A \cap \overline{\mathbb{R}}$.

There are more complicated examples as well. Let ℓ_1 and ℓ_2 be parallel hyperbolic lines in $\mathbb{H}$, and let H be the closed region in $\mathbb{H}$ that consists of the two lines ℓ_1 and ℓ_2, together with the part of $\mathbb{H}$ that lies between them. There are two possibilities for the boundary at infinity of this region H.

Let C_k be the circle in $\overline{\mathbb{C}}$ containing ℓ_k. As ℓ_1 and ℓ_2 are disjoint, either C_1 and C_2 are disjoint, or C_1 and C_2 intersect in a single point, which is then necessarily contained in $\overline{\mathbb{R}}$.

In the case in which C_1 and C_2 intersect at the point x of $\overline{\mathbb{R}}$, the boundary at infinity of H is the union of a closed arc in $\overline{\mathbb{R}}$ and the set $\{x\}$.

In the case in which C_1 and C_2 are disjoint, the boundary at infinity of H is the union of two closed arcs in $\overline{\mathbb{R}}$. These two possibilities are shown in Figure 1.5.

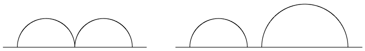

Figure 1.5: Two possibilities for parallel hyperbolic lines

This gives us a way of distinguishing two different types of parallelism for hyperbolic lines in $\mathbb{H}$. Namely, there are parallel hyperbolic lines whose boundaries at infinity intersect, and there are parallel hyperbolic lines whose boundaries at infinity are disjoint. When we need to make the distinction, we refer to a pair of the latter type as *ultraparallel.*

There is another way to see the distinction between parallel and ultraparallel hyperbolic lines.

Exercise 1.17

Let ℓ_1 and ℓ_2 be parallel hyperbolic lines. Show that ℓ_1 and ℓ_2 are ultraparallel if and only if there exists a hyperbolic line perpendicular to both ℓ_1 and ℓ_2.

We saw in Section 1.1, specifically in Proposition 1.2, that two points in $\mathbb{H}$ determine a unique hyperbolic line in $\mathbb{H}$. The key to the proof of this fact is that there exists a unique Euclidean circle or Euclidean line in $\mathbb{C}$ that passes through the given two points and is perpendicular to the real axis $\mathbb{R}$.

This same argument applies to hyperbolic lines determined by points at infinity.

Proposition 1.13

Let p be a point of $\mathbb{H}$ and q a point of $\overline{\mathbb{R}}$. Then, there is a unique hyperbolic line in $\mathbb{H}$ determined by p and q.

Proof

Suppose that $q = \infty$. Of all the hyperbolic lines through p, there is exactly one that contains q in its boundary at infinity, namely, the hyperbolic line contained in the Euclidean line $\{z \in \mathbb{C} \,|\, \mathrm{Re}(z) = \mathrm{Re}(p)\}$. The statement about uniqueness follows from the observation that no hyperbolic line contained in a Euclidean circle contains ∞ in its boundary at infinity.

Suppose that $q \neq \infty$ and that $\mathrm{Re}(p) = \mathrm{Re}(q)$. Then, the hyperbolic line contained in the Euclidean line $\{z \in \mathbb{C} \mid \mathrm{Re}(z) = \mathrm{Re}(p)\}$ is the unique hyperbolic line through p that contains q in its boundary at infinity.

Suppose that $q \neq \infty$ and that $\mathrm{Re}(p) \neq \mathrm{Re}(q)$. Then, we may again use the construction from the proof of Proposition 1.2, using the perpendicular bisector of the Euclidean line segment joining p to q, to find the unique Euclidean circle centred on the real axis $\mathbb{R}$ that passes through both p and q. Intersecting this circle with $\mathbb{H}$ yields the unique hyperbolic line determined by p and q. This completes the proof of Proposition 1.13. **QED**

We refer to the part of the hyperbolic line between p and q as the *hyperbolic ray determined by p and q* or as the *hyperbolic ray through p with endpoint at infinity q*.

The argument for the existence and uniqueness of a hyperbolic line determined by two distinct points at infinity is left as an exercise.

Exercise 1.18

Let p and q be two distinct points of $\overline{\mathbb{R}}$. Prove that p and q determine a unique hyperbolic line whose endpoints at infinity are p and q.

2
The General Möbius Group

As our goal is to study the geometry of the hyperbolic plane by considering quantities invariant under the action of a reasonable group of transformations, we spend this chapter describing just such a reasonable group of transformations of $\overline{\mathbb{C}}$, namely, the *general Möbius group* Möb, which consists of compositions of *Möbius transformations* and *reflections*. We close the chapter by restricting our attention to the transformations in Möb preserving $\mathbb{H}$.

2.1 The Group of Möbius Transformations

As every hyperbolic line in $\mathbb{H}$ is by definition contained in a circle in $\overline{\mathbb{C}}$, we begin the process of determining the transformations of $\mathbb{H}$ taking hyperbolic lines to hyperbolic lines by first determining the group of homeomorphisms of $\overline{\mathbb{C}}$ taking circles in $\overline{\mathbb{C}}$ to circles in $\overline{\mathbb{C}}$.

For the sake of notational convenience, let $\text{Homeo}^{\text{C}}(\overline{\mathbb{C}})$ be the subset of the group $\text{Homeo}(\overline{\mathbb{C}})$ of homeomorphisms of $\overline{\mathbb{C}}$ that contains those homeomorphisms of $\overline{\mathbb{C}}$ taking circles in $\overline{\mathbb{C}}$ to circles in $\overline{\mathbb{C}}$. That is, if $f : \overline{\mathbb{C}} \to \overline{\mathbb{C}}$ is a homeomorphism, then $f \in \text{Homeo}^{\text{C}}(\overline{\mathbb{C}})$ if for every circle in $\overline{\mathbb{C}}$ A, we have that $f(A)$ is also a circle in $\overline{\mathbb{C}}$. Note that, although it is easy to see that the composition of two elements of $\text{Homeo}^{\text{C}}(\overline{\mathbb{C}})$ is again an element of $\text{Homeo}^{\text{C}}(\overline{\mathbb{C}})$ and that the identity homeomorphism is an element of $\mathrm{Homeo}^{\mathrm{C}}(\overline{\mathbb{C}})$, we do not yet know that inverses of elements of $\mathrm{Homeo}^{\mathrm{C}}(\overline{\mathbb{C}})$ lie in $\mathrm{Homeo}^{\mathrm{C}}(\overline{\mathbb{C}})$, and hence we cannot yet conclude that $\mathrm{Homeo}^{\mathrm{C}}(\overline{\mathbb{C}})$ is a group.

Note that there are many homeomorphisms of $\overline{\mathbb{C}}$ that do not lie in $\mathrm{Homeo}^{\mathrm{C}}(\overline{\mathbb{C}})$.

Exercise 2.1

Give an explicit example of an element of $\mathrm{Homeo}(\overline{\mathbb{C}})$ that is not an element of $\mathrm{Homeo}^{\mathrm{C}}(\overline{\mathbb{C}})$.

We begin by considering a class of homeomorphisms of $\overline{\mathbb{C}}$ that we understand, namely, those arising from polynomials. As we saw in Exercise 1.14 and Exercise 1.15, we may extend each polynomial $g(z)$ to the function $f : \overline{\mathbb{C}} \to \overline{\mathbb{C}}$ given by

$$f(z) = g(z) \text{ for } z \in \mathbb{C} \text{ and } f(\infty) = \infty.$$

As we wish to consider homeomorphisms of $\overline{\mathbb{C}}$ that arise from polynomials, we are forced to restrict our attention to polynomials of degree 1.

Proposition 2.1

The element f of $\mathrm{Homeo}(\overline{\mathbb{C}})$ defined by

$$f(z) = az + b \text{ for } z \in \mathbb{C} \text{ and } f(\infty) = \infty,$$

where a, $b \in \mathbb{C}$ and $a \neq 0$, is an element of $\mathrm{Homeo}^{\mathrm{C}}(\overline{\mathbb{C}})$.

Proof

Recall from Section 1.2 that each circle in $\overline{\mathbb{C}}$ A can be described as the set of solutions to an equation of the form

$$\alpha z\overline{z} + \beta z + \overline{\beta}\overline{z} + \gamma = 0,$$

where α, $\gamma \in \mathbb{R}$ and $\beta \in \mathbb{C}$, and where $\alpha \neq 0$ if and only if A is a circle in $\mathbb{C}$.

We begin with the case in which A is a Euclidean line in $\mathbb{C}$. So, consider the Euclidean line A given as the solution to the equation

$$A = \{z \in \mathbb{C} \mid \beta z + \overline{\beta}\overline{z} + \gamma = 0\},$$

where $\beta \in \mathbb{C}$ and $\gamma \in \mathbb{R}$. We wish to show that if z satisfies this equation, then $w = az + b$ satisfies a similar equation.

As $w = az + b$, we have that $z = \frac{1}{a}(w - b)$. Substituting this calculation into the equation for A given above yields

$$\begin{aligned} \beta z + \overline{\beta}\overline{z} + \gamma &= \beta\frac{1}{a}(w-b) + \overline{\beta}\overline{\frac{1}{a}(w-b)} + \gamma \\ &= \frac{\beta}{a}w + \overline{\left(\frac{\beta}{a}\right)}\overline{w} - \frac{\beta}{a}b - \overline{\frac{\beta}{a}b} + \gamma = 0. \end{aligned}$$

As $\frac{\beta}{a}b + \overline{\frac{\beta}{a}b} = 2\,\mathrm{Re}\left(\frac{\beta}{a}b\right)$ is real and as the coefficients of w and $\overline{w}$ are complex conjugates, this shows that w also satisfies the equation of a Euclidean line. Hence, f takes Euclidean lines in $\mathbb{C}$ to Euclidean lines in $\mathbb{C}$. The proof that f takes Euclidean circles to Euclidean circles is similar and is left as an exercise.

Exercise 2.2

Show that the homeomorphism f of $\overline{\mathbb{C}}$ defined by setting

$$f(z) = az + b \text{ for } z \in \mathbb{C} \text{ and } f(\infty) = \infty,$$

where a, $b \in \mathbb{C}$ and $a \neq 0$, takes Euclidean circles in $\mathbb{C}$ to Euclidean circles in $\mathbb{C}$.

Exercise 2.2 completes the proof of Proposition 2.1. **QED**

We can refine this argument to obtain quantitative information about the image circle in $\overline{\mathbb{C}}$ in terms of the coefficients of $f(z) = az + b$ and the equation of the original circle in $\overline{\mathbb{C}}$.

For example, suppose that L is a Euclidean line given by the equation $\beta z + \overline{\beta}\overline{z} + \gamma = 0$, and recall from the solution of Exercise 1.1 that the slope of L is $\frac{\mathrm{Re}(\beta)}{\mathrm{Im}(\beta)}$.

We have seen that f takes L to the Euclidean line $f(L)$ given by the equation

$$\frac{\beta}{a}w + \overline{\left(\frac{\beta}{a}\right)}\overline{w} - \frac{\beta}{a}b - \overline{\frac{\beta}{a}}b + \gamma = 0,$$

which has slope $\frac{\mathrm{Re}(\beta\overline{a})}{\mathrm{Im}(\beta\overline{a})}$.

Exercise 2.3

Determine the Euclidean centre and Euclidean radius of the image of the Euclidean circle A given by the equation $\alpha z\overline{z} + \beta z + \overline{\beta}\overline{z} + \gamma = 0$ under the homeomorphism

$$f(z) = az + b \text{ for } z \in \mathbb{C} \text{ and } f(\infty) = \infty,$$

where a, $b \in \mathbb{C}$ and $a \neq 0$.

There is another homeomorphism of $\overline{\mathbb{C}}$ we considered earlier, in Proposition 1.11, namely, the function $J : \overline{\mathbb{C}} \to \overline{\mathbb{C}}$ defined by setting

$$J(z) = \frac{1}{z} \text{ for } z \in \mathbb{C} - \{0\},\ \ J(0) = \infty,\ \text{ and } J(\infty) = 0.$$

Proposition 2.2

The element J of $\text{Homeo}(\overline{\mathbb{C}})$ defined by

$$J(z) = \frac{1}{z} \text{ for } z \in \mathbb{C} - \{0\},\ \ J(0) = \infty,\ \text{ and } J(\infty) = 0$$

is an element of $\text{Homeo}^{\text{C}}(\overline{\mathbb{C}})$.

Proof

We proceed as before. Let A be a circle in $\overline{\mathbb{C}}$ given by the equation

$$\alpha z\overline{z} + \beta z + \overline{\beta}\overline{z} + \gamma = 0,$$

where α, $\gamma \in \mathbb{R}$ and $\beta \in \mathbb{C}$.

Set $w = \frac{1}{z}$, so that $z = \frac{1}{w}$. Substituting this back into the equation for A gives

$$\alpha \frac{1}{w}\overline{\frac{1}{w}} + \beta\frac{1}{w} + \overline{\beta}\overline{\frac{1}{w}} + \gamma = 0.$$

Multiplying through by $w\overline{w}$, we see that w satisfies the equation

$$\alpha + \beta\overline{w} + \overline{\beta}w + \gamma w\overline{w} = 0.$$

As α and γ are real and as the coefficients of w and $\overline{w}$ are complex conjugates, this is again the equation of a circle in $\overline{\mathbb{C}}$. This completes the proof of Proposition 2.2. **QED**

As we can see from the proof of Proposition 2.1, we can extract some quantitative information from the proof of Proposition 2.2 about the circle in $\overline{\mathbb{C}}$ $J(A)$ in terms of the circle in $\overline{\mathbb{C}}$ A.

As a specific example, let A be the circle in $\overline{\mathbb{C}}$ given by the equation $2z+2\overline{z}+3=0$. Then, $J(A)$ is the circle in $\overline{\mathbb{C}}$ given by the equation $2\overline{w}+2w+3w\overline{w}=0$, which is a Euclidean circle in $\mathbb{C}$ with Euclidean centre $-\frac{2}{3}$ and Euclidean radius $\frac{2}{3}$.

Exercise 2.4

Let A be a Euclidean circle in $\mathbb{C}$ given by the equation $|z-z_0|=r$. Determine conditions on z_0 and r so that $J(A)$ is a Euclidean line in $\mathbb{C}$.

Note that all possible compositions of these two types of homeomorphisms of $\overline{\mathbb{C}}$, namely, the $f(z)=az+b$ with a, $b\in\mathbb{C}$ and $a\neq 0$, and $J(z)=\frac{1}{z}$, have the same form $m(z)=\frac{\alpha z+\beta}{\gamma z+\delta}$, where $\alpha\delta-\beta\gamma\neq 0$. This leads us to the following definition.

Definition 2.3

A *Möbius transformation* is a function $m:\overline{\mathbb{C}}\to\overline{\mathbb{C}}$ of the form

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

where a, b, c, $d\in\mathbb{C}$ and $ad-bc\neq 0$. Let Möb^+ denote the set of all Möbius transformations.

We pause here to insert a remark about the arithmetic of ∞. For any $a\neq 0$, we can unambiguously assign the value of $\frac{a}{0}$ to be ∞ by continuity. That is, we set

$$\frac{a}{0}=\lim_{w\to 0}\frac{a}{w}.$$

As $a\neq 0$, $\frac{a}{w}$ is nonzero, and by considering the modulus $|\frac{a}{w}|$, we can see that $\lim_{w\to 0}\frac{a}{w}=\infty$ in $\overline{\mathbb{C}}$. However, we still cannot make unambiguous sense of the expressions $\frac{0}{0}$ and $\frac{\infty}{\infty}$.

Similarly, we define the image of ∞ under $m(z)=\frac{az+b}{cz+d}$ by continuity. That is, we set

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

The value $m(\infty)$ is well defined because one of a or c has to be nonzero, as from the definition of Möbius transformation we know that $ad - bc \neq 0$.

Observe that, because $m(\infty) = \frac{a}{c}$, we have that $m(\infty) = \infty$ if and only if $c = 0$. Also, because $m(0) = \frac{b}{d}$, we have that $m(0) = 0$ if and only if $b = 0$.

As we will see in Exercise 2.5, we can write an explicit expression for the inverse of a Möbius transformation. As the composition of two Möbius transformations is again a Möbius transformation (which we leave for the interested reader to verify), we have that the set Möb^+ of Möbius transformations is a group under composition with identity element $e : \overline{\mathbb{C}} \to \overline{\mathbb{C}}$, $e(z) = z$.

Exercise 2.5

To prove that Möbius transformations are bijective, give an explicit expression for the inverse of the Möbius transformation $m(z) = \frac{az+b}{cz+d}$.

As we have already mentioned, the form of a Möbius transformation is similar to the forms of the homeomorphisms of $\overline{\mathbb{C}}$ we encountered earlier this section, namely, the

$$f(z) = az + b \text{ for } z \in \mathbb{C} \text{ and } f(\infty) = \infty,$$

where a, $b \in \mathbb{C}$ and $a \neq 0$, and

$$J(z) = \frac{1}{z} \text{ for } z \in \mathbb{C} - \{0\},\ \ J(0) = \infty,\ \text{ and } J(\infty) = 0.$$

In fact, we may write any Möbius transformation $m(z) = \frac{az+b}{cz+d}$ as a composition of such homeomorphisms.

Theorem 2.4

Consider the Möbius transformation $m(z) = \frac{az+b}{cz+d}$, where a, b, c, $d \in \mathbb{C}$ and $ad - bc \neq 0$.

If $c = 0$, then $m(z) = \frac{a}{d}z + \frac{b}{d}$.

If $c \neq 0$, then $m(z) = f(J(g(z)))$, where $g(z) = c^2 z + cd$ and $f(z) = -(ad - bc)z + \frac{a}{c}$ for $z \in \mathbb{C}$, and $f(\infty) = \infty = g(\infty)$.

Proof

The proof of Theorem 2.4 is a direct calculation. If $c = 0$, there is nothing to check. If $c \neq 0$, then

$$m(z) = \frac{az+b}{cz+d} = \frac{(az+b)\,c}{(cz+d)\,c} = \frac{acz+bc}{c^2z+cd}.$$

As $ad - bc \neq 0$, we have that

$$m(z) = \frac{acz+bc}{c^2z+cd} = \frac{acz+ad-(ad-bc)}{c^2z+cd} = \frac{a}{c} - \frac{ad-bc}{c^2z+cd} = f(J(g(z))),$$

where $g(z) = c^2z + cd$ and $f(z) = -(ad-bc)z + \frac{a}{c}$. This completes the proof of Theorem 2.4. **QED**

Theorem 2.4 has several immediate corollaries. First, every Möbius transformation is a homeomorphism of $\overline{\mathbb{C}}$, as it is a composition of homeomorphisms of $\overline{\mathbb{C}}$. That is,

$$\text{Möb}^+ \subset \text{Homeo}(\overline{\mathbb{C}}).$$

Second, every Möbius transformation takes circles in $\overline{\mathbb{C}}$ to circles in $\overline{\mathbb{C}}$, as it is a composition of functions with this property. We combine these observations in the following theorem.

Theorem 2.5

$\text{Möb}^+ \subset \text{Homeo}^{\text{C}}(\overline{\mathbb{C}})$.

We note here that the condition that $ad - bc \neq 0$ in the definition of a Möbius transformation is not spurious.

Exercise 2.6

Consider a function $p : \overline{\mathbb{C}} \to \overline{\mathbb{C}}$ of the form $p(z) = \frac{az+b}{cz+d}$, where a, b, c, $d \in \mathbb{C}$ and $ad - bc = 0$. Prove that p is not a homeomorphism of $\overline{\mathbb{C}}$.

We close this section with a crude classification of Möbius transformations, based on the number of fixed points. A *fixed point* of the Möbius transformation m is a point z of $\overline{\mathbb{C}}$ satisfying $m(z) = z$. Suppose that m is not the identity.

We saw earlier in this section that for $m(z) = \frac{az+b}{cz+d}$, we have that $m(\infty) = \frac{a}{c}$, and so $m(\infty) = \infty$ if and only if $c = 0$.

If $c = 0$, then $m(z) = \frac{a}{d}z + \frac{b}{d}$, and the fixed point of m in $\mathbb{C}$ is the solution to the equation $m(z) = \frac{a}{d}z + \frac{b}{d} = z$. If $\frac{a}{d} = 1$, then $b \neq 0$ (as m is not the identity) and there is no solution in $\mathbb{C}$, whereas if $\frac{a}{d} \neq 1$, then $z = \frac{b}{d-a}$ is the unique solution in $\mathbb{C}$. In particular, if $c = 0$, then m has either one or two fixed points.

If $c \neq 0$, then $m(\infty) \neq \infty$, and so the fixed points of m are the solutions in $\mathbb{C}$ of the equation $m(z) = \frac{az+b}{cz+d} = z$, which are the roots of the quadratic polynomial $cz^2 + (d-a)z - b = 0$. In particular, if $c \neq 0$, then again m has either one or two fixed points.

This analysis has the following important consequence.

Theorem 2.6

Let $m(z)$ be a Möbius transformation fixing three distinct points of $\overline{\mathbb{C}}$. Then, m is the identity transformation. That is, $m(z) = z$ for every point z of $\overline{\mathbb{C}}$.

Exercise 2.7

Calculate the fixed points of each of the following Möbius transformations.

1. $m(z) = \frac{2z+5}{3z-1}$; 2. $m(z) = 7z + 6$; 3. $J(z) = \frac{1}{z}$; 4. $m(z) = \frac{z}{z+1}$.

2.2 Transitivity Properties of Möb$^+$

One of the most basic properties of Möb$^+$ is that it acts *uniquely triply transitively* on $\overline{\mathbb{C}}$. By this we mean that given two triples (z_1, z_2, z_3) and (w_1, w_2, w_3) of distinct points of $\overline{\mathbb{C}}$, there exists a unique element m of Möb$^+$ so that $m(z_1) = w_1$, $m(z_2) = w_2$, and $m(z_3) = w_3$.

As is often done, we begin our proof of existence and uniqueness by first showing uniqueness. We then construct a particular Möbius transformation by whatever means are at hand, and observe that by uniqueness it must be the only one.

So, given two triples (z_1, z_2, z_3) and (w_1, w_2, w_3) of distinct points of $\overline{\mathbb{C}}$, suppose there are two elements m and n of Möb^+ satisfying $n(z_1) = w_1 = m(z_1)$, $n(z_2) = w_2 = m(z_2)$, and $n(z_3) = w_3 = m(z_3)$. By Theorem 2.6, we know that because $m^{-1} \circ n$ fixes three distinct points of $\overline{\mathbb{C}}$, it is the identity, and so $m = n$. This completes the proof of uniqueness.

To demonstrate the existence of a Möbius transformation taking (z_1, z_2, z_3) to (w_1, w_2, w_3), it suffices to show that there is a Möbius transformation m satisfying $m(z_1) = 0$, $m(z_2) = 1$, and $m(z_3) = \infty$. If we can construct such an m, we can also construct a Möbius transformation n satisfying $n(w_1) = 0$, $n(w_2) = 1$, and $n(w_3) = \infty$, and then $n^{-1} \circ m$ is the desired transformation taking (z_1, z_2, z_3) to (w_1, w_2, w_3).

So, it remains only to construct a Möbius transformation m satisfing $m(z_1) = 0$, $m(z_2) = 1$, and $m(z_3) = \infty$. We work in the case in which all the z_k lie in $\mathbb{C}$, and leave the derivation of m in the case in which one of the z_k is ∞ for the interested reader. Explicitly, consider the function on $\overline{\mathbb{C}}$ given by

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

Just by its construction, we have that $m(z_1) = 0$, $m(z_2) = 1$, and $m(z_3) = \infty$. Moreover, because the z_k are distinct,

$$(z_2 - z_3)(-z_3)(z_2 - z_1) - (-z_1)(z_2 - z_3)(z_2 - z_1) = (z_2 - z_3)(z_1 - z_3)(z_2 - z_1) \neq 0,$$

and so m is a Möbius transformation.

Exercise 2.8

Derive the general form of the Möbius transformation taking the triple (∞, z_2, z_3) to the triple $(0, 1, \infty)$.

As is often the case, the actual construction of the specific Möbius transformation taking one triple to another can be fairly unpleasant. For example, let us consider the two triples $(2i, 1 + i, 3)$ and $(0, 2 + 2i, 4)$ and construct the Möbius transformation taking $(2i, 1 + i, 3)$ to $(0, 2 + 2i, 4)$. Warning: This example has not been chosen for its numerical elegance.

Following the proof of existence, we construct the Möbius transformation m taking $(2i, 1 + i, 3)$ to $(0, 1, \infty)$ and the Möbius transformation n taking $(0, 2 + 2i, 4)$ to $(0, 1, \infty)$.

The Möbius transformation m taking $(2i, 1+i, 3)$ to $(0, 1, \infty)$ is given by

$$m(z) = \frac{(z-2i)}{(z-3)} \frac{(1+i-3)}{(1+i-2i)} = \frac{(-2+i)z+2+4i}{(1-i)z-3+3i}.$$

The Möbius transformation n taking $(0, 2+2i, 4)$ to $(0, 1, \infty)$ is given as

$$n(z) = \frac{z}{(z-4)} \frac{(2+2i-4)}{(2+2i)} = \frac{(-2+2i)z}{(2+2i)z-8-8i}.$$

So, the transformation we are looking for is

$$n^{-1} \circ m(z) = \frac{(24+8i)z+16-48i}{(6+6i)z+4-24i}.$$

Up to this point, we've been considering *ordered* triples of distinct points in $\overline{\mathbb{C}}$. If we consider *unordered* triples, and in particular, if we ask about the Möbius transformations taking one unordered triple of distinct points to another, the proof of existence goes through without change, but the proof of uniqueness no longer holds.

Exercise 2.9

Consider the unordered triple $T = \{0, 1, \infty\}$ of points of $\overline{\mathbb{C}}$. Determine all Möbius transformations m satisfying $m(T) = T$.

The action of Möb^+ on the set of triples of distinct points of $\overline{\mathbb{C}}$ is an example of a *group action*.

Definition 2.7

A group G *acts on a set* X if there is a homomorphism from G into the group $\text{bij}(X)$ of bijections of X.

That is, a group G acts on a set X if every element of g gives rise to a bijection of X, and moreover if multiplication of elements of G using the group operation corresponds to composition of the corresponding bijections of X.

We do not do much with group actions in this book, other than making use of some of the basic terminology. For more information, the interested reader should pick up a book on abstract algebra, such as Herstein [20]. Philosophically, considering group actions allows one to view a group not as an abstract object, but as a well behaved collection of symmetries of a set X.

There are many adjectives that one can apply to group actions, and restrictions that one can apply to the types of bijections considered.

For instance, say that G acts *transitively* on X if for each pair x and y of elements of X, there exists some element g of G satisfying $g(x) = y$. This is one of the properties of most interest to us. The following lemma gives a slightly easier condition to check to obtain transitivity, which is merely a generalization of the idea we used in the proof that Möb^+ acts transitively on triples of distinct points of $\overline{\mathbb{C}}$.

Lemma 2.8

Suppose that a group G acts on a set X, and let x_0 be a point of X. Suppose that for each point y of X there exists an element g of G so that $g(y) = x_0$. Then, G acts transitively on X.

Proof

Given two points y and z of X, choose elements g_y and g_z of G so that $g_y(y) = x_0 = g_z(z)$. Then, $(g_z)^{-1} \circ g_y(y) = z$. This completes the proof of Lemma 2.8.

QED

Also inspired by considering the action of Möb^+ on triples of distinct points of $\overline{\mathbb{C}}$, say that a group G acts *uniquely transitively* on a set X if for each pair x and y of elements of X, there exists one and only one element g of G with $g(x) = y$. In this language, we can restate what we know about the action of Möb^+ on triples of distinct points of $\overline{\mathbb{C}}$.

Theorem 2.9

Möb^+ acts uniquely transitively on the set $\mathcal{T}$ of triples of distinct points of $\overline{\mathbb{C}}$.

There are other sets of objects in $\overline{\mathbb{C}}$ on which Möb^+ acts transitively.

Theorem 2.10

Möb^+ acts transitively on the set $\mathcal{C}$ of circles in $\overline{\mathbb{C}}$.

Proof

The first step in proving Theorem 2.10 is to observe that a triple of distinct points in $\overline{\mathbb{C}}$ determines a unique circle in $\overline{\mathbb{C}}$.

To see this, let (z_1, z_2, z_3) be a triple of distinct points of $\overline{\mathbb{C}}$. If all the z_k lie in $\mathbb{C}$ and are not colinear, then there exists a unique Euclidean circle passing through all three. The centre of the circle is the point of intersection of any two of the perpendicular bisectors of the Euclidean line segments joining any two of z_1, z_2, z_3. If all the z_k lie in $\mathbb{C}$ and are colinear, then there exists a unique Euclidean line passing through all three. If one of the z_k is ∞, then there is a unique Euclidean line passing through the other two.

However, although each triple of distinct points of $\overline{\mathbb{C}}$ determines a unique circle in $\overline{\mathbb{C}}$, the converse is not true. Given a circle in $\overline{\mathbb{C}}$ A, there are infinitely many triples of distinct points of $\overline{\mathbb{C}}$ that give rise to A.

So, let A and B be two circles in $\overline{\mathbb{C}}$. Choose a triple of distinct points on A and a triple of distinct points on B, and let m be the Möbius transformation taking the triple of distinct points determining A to the triple of distinct points determining B. As $m(A)$ and B are then two circles in $\overline{\mathbb{C}}$ that pass through the same triple of distinct points and as Möbius transformations take circles in $\overline{\mathbb{C}}$ to circles in $\overline{\mathbb{C}}$, we have that $m(A) = B$. This completes the proof of Theorem 2.10. **QED**

However, the fact that a circle in $\overline{\mathbb{C}}$ does not determine a unique triple of distinct points in $\overline{\mathbb{C}}$ means that this action is not uniquely transitive. That is, given two circles in $\overline{\mathbb{C}}$, there are in fact many Möbius transformations taking one to the other.

For example, we may mimic Exercise 2.9. Let (z_1, z_2, z_3) be a triple of distinct points, and let A be the circle in $\overline{\mathbb{C}}$ determined by (z_1, z_2, z_3). Then, the identity takes A to A. However, the Möbius transformation taking (z_1, z_2, z_3) to (z_2, z_1, z_3) also takes A to A. We will encounter this phenomenon again in

Section 2.8, in which we determine the set of Möbius transformations taking any circle in $\overline{\mathbb{C}}$ A to itself.

We can rephrase this argument as saying that there exists a well defined surjective function from the set $\mathcal{T}$ of triples of distinct points of $\overline{\mathbb{C}}$ to the set $\mathcal{C}$ of circles in $\overline{\mathbb{C}}$. As Möb^+ acts transitively on $\mathcal{T}$, we can use this function from $\mathcal{T}$ to $\mathcal{C}$ to push down the action of Möb^+ from $\mathcal{T}$ to $\mathcal{C}$. The lack of uniqueness in the action of Möb^+ on $\mathcal{C}$ is a reflection of the fact that this function is not injective.

We can also consider the action of Möb^+ on the set $\mathcal{D}$ of discs in $\overline{\mathbb{C}}$.

Theorem 2.11

Möb^+ acts transitively on the set $\mathcal{D}$ of discs in $\overline{\mathbb{C}}$.

Proof

As might be expected, the proof of Theorem 2.11 is similar to the proof of Theorem 2.10. In fact, the proofs differ in only one respect.

Let D and E be two discs in $\overline{\mathbb{C}}$, where D is determined by the circle in $\overline{\mathbb{C}}$ C_D and E is determined by the circle in $\overline{\mathbb{C}}$ C_E. As Möb^+ acts transitively on the set $\mathcal{C}$ of circles in $\overline{\mathbb{C}}$, there is a Möbius transformation m satisfying $m(C_D) = C_E$, and so $m(D)$ is a disc determined by C_E.

However, there are two discs determined by C_E, and we have no way of knowing whether $m(D) = E$ or $m(D)$ is the other disc determined by C_E. If $m(D) = E$, we are done.

If $m(D) \neq E$, we need to find a Möbius transformation taking C_E to itself and interchanging the two discs determined by C_E.

This construction is not too difficult. We first work with a circle in $\overline{\mathbb{C}}$ we understand, and we then use the transitivity of Möb^+ on the set of circles in $\overline{\mathbb{C}}$ to transport our solution for this particular circle to any other circle.

For the circle $\overline{\mathbb{R}}$, we have already seen the answer to this question, namely, the Möbius transformation $J(z) = \frac{1}{z}$. As $J(0) = \infty$, $J(\infty) = 0$, and $J(1) = 1$, we see that J takes $\overline{\mathbb{R}}$ to itself. As $J(i) = \frac{1}{i} = -i$, we see that J does not take $\mathbb{H}$ to itself, and so J interchanges the two discs determined by $\overline{\mathbb{R}}$. Now, let A be any circle in $\overline{\mathbb{C}}$ and let n be a Möbius transformation satisfying $n(A) = \overline{\mathbb{R}}$.

Then, the Möbius transformation $n^{-1} \circ J \circ n$ takes A to itself and interchanges the two discs determined by A.

In particular, there exists a Möbius transformation p so that $p(C_E) = C_E$ and p interchanges the two discs determined by C_E. Hence, in the case in which $m(D) \neq E$, the composition $p \circ m$ satisfies $p \circ m(C_D) = p(C_E) = C_E$ and $p \circ m(D) = E$.

This completes the proof of Theorem 2.11. **QED**

As with determining the Möbius transformation taking one triple of distinct points of $\overline{\mathbb{C}}$ to another triple of distinct points, it can be somewhat messy to write out the Möbius transformation taking one disc in $\overline{\mathbb{C}}$ to another.

Consider the two discs

$$D = \{z \in \mathbb{C} \mid |z| < 2\} \text{ and } E = \{z \in \mathbb{C} \mid |z - (4 + 5i)| < 1\}.$$

Many different Möbius transformations take E to D. We construct one.

Let $m(z) = z - 4 - 5i$. As E is the Euclidean disc with centre $4 + 5i$ and radius 1, we have that $m(E)$ is the Euclidean disc with centre $m(4 + 5i) = 0$ and radius 1. If we now compose m with $n(z) = 2z$, we see that $n \circ m(E)$ is the Euclidean disc with centre 0 and radius 2, so that $n \circ m(E) = D$, as desired. Writing out $n \circ m$ explicitly, we get $n \circ m(z) = n(z - 4 - 5i) = 2z - 8 - 10i$.

Exercise 2.10

Give an explicit Möbius transformation taking the unit disc $\mathbb{D} = U_1(0)$ in $\mathbb{C}$ to $\mathbb{H}$.

2.3 The Cross Ratio

In Section 2.2, we considered the transitivity properties of Möb^+. We saw that Möb^+acts uniquely transitively on the set $\mathcal{T}$ of ordered triples of distinct points of $\overline{\mathbb{C}}$ and acts transitively on both the set $\mathcal{C}$ of circles in $\overline{\mathbb{C}}$ and the set $\mathcal{D}$ of discs in $\overline{\mathbb{C}}$.

In this section, we consider a different sort of question, and we start by asking about functions on $\overline{\mathbb{C}}$ that are invariant under Möb^+. Several variants of this question will occupy our attention at different times throughout the book.

Definition 2.12

A *function invariant under* Möb^+ is a function $f : U \to \overline{\mathbb{C}}$, where U is an open set in $\overline{\mathbb{C}}^k$, of variables $(z_1, \dots, z_k)$ so that

$$f(z_1, \dots, z_k) = f(m(z_1), \dots, m(z_k))$$

for all $m \in \text{Möb}^+$ and all $(z_1, \dots, z_k) \in U$.

Typically, in Definition 2.12, we require that U be invariant under Möb^+, so that if $(z_1, \dots, z_k) \in U$ and $m \in \text{Möb}^+$, then $(m(z_1), \dots, m(z_k)) \in U$.

Exercise 2.11

Show that the function $f : \overline{\mathbb{C}} \to \overline{\mathbb{C}}$ given by $f(z) = z^2$ for $z \in \mathbb{C}$ and $f(\infty) = \infty$, is not invariant under Möb^+. Determine whether there exists a nontrivial subgroup of Möb^+ under which f is invariant.

In fact, the triple transitivity of Möb^+ on $\overline{\mathbb{C}}$ implies that for $1 \le n \le 3$, the only functions f on $\overline{\mathbb{C}}$ of n variables invariant under Möb^+ are the constant functions.

For $n \ge 4$, the situation becomes more interesting. One example of a function of four variables invariant under Möb^+ is the cross ratio.

Definition 2.13

Given four distinct points z_1, z_2, z_3, and z_4 in $\mathbb{C}$, define the *cross ratio* of z_1, z_2, z_3, and z_4 to be

$$[z_1, z_2; z_3, z_4] = \frac{(z_1 - z_4)}{(z_1 - z_2)} \frac{(z_3 - z_2)}{(z_3 - z_4)}.$$

Following our usual pattern, if one of the z_k is ∞, we define the cross ratio by continuity. That is, we set

$$\begin{aligned}[\infty, z_2; z_3, z_4] = \lim_{z\to\infty} (z, z_2; z_3, z_4) &= \lim_{z\to\infty} \frac{(z - z_4)}{(z - z_2)} \frac{(z_3 - z_2)}{(z_3 - z_4)} \\ &= \lim_{z\to\infty} \frac{(1 - \frac{z_4}{z})}{(1 - \frac{z_2}{z})} \frac{(z_3 - z_2)}{(z_3 - z_4)} = \frac{z_3 - z_2}{z_3 - z_4}.\end{aligned}$$

The cross ratios $[z_1, \infty; z_3, z_4]$, $[z_1, z_2; \infty, z_4]$, and $[z_1, z_2; z_3, \infty]$ are defined similarly.

Exercise 2.12

Show that the cross ratio, as a function on the subset U of $\overline{\mathbb{C}}^4$ consisting of quadruples of distinct points of $\overline{\mathbb{C}}$, is invariant under Möb^+.

Exercise 2.13

Determine whether or not there exists a function $F : \overline{\mathbb{C}}^4 \to \overline{\mathbb{C}}$ invariant under Möb^+ so that $F(z_1, z_2, z_3, z_4) = [z_1, z_2; z_3, z_4]$ for all quadruples z_1, z_2, z_3, z_4 of distinct points of $\overline{\mathbb{C}}$.

In some cases, the cross ratio is particularly easy to calculate. For example, consider $[\infty, 0; 1, z]$. From what we have done above, we have that

$$[\infty, 0; 1, z] = \frac{1}{1-z} = \frac{1-\overline{z}}{|1-z|^2}.$$

In particular, we have that $[\infty, 0; 1, z]$ is real if and only if $\overline{z}$, and hence z, is real.

Combining this result with the fact that the cross ratio, as a function on the set of quadruples of distinct points of $\overline{\mathbb{C}}$, is invariant under Möb^+ gives us an easy test to see whether four distinct points of $\overline{\mathbb{C}}$ lie on a circle in $\overline{\mathbb{C}}$.

Proposition 2.14

Let z_1, z_2, z_3, and z_4 be four distinct points in $\overline{\mathbb{C}}$. Then, z_1, z_2, z_3, and z_4 lie on a circle in $\overline{\mathbb{C}}$ if and only if the cross ratio $[z_1, z_2; z_3, z_4]$ is real.

Proof

Let z_1, z_2, z_3, and z_4 be four distinct points in $\overline{\mathbb{C}}$, and let m be a Möbius transformation satisfying $m(z_1) = \infty$, $m(z_2) = 0$, and $m(z_3) = 1$.

Observe that $m(z_1) = \infty$, $m(z_2) = 0$, $m(z_3) = 1$, and $m(z_4)$ lie on a circle in $\overline{\mathbb{C}}$, namely, $\overline{\mathbb{R}}$, if and only if $m(z_4)$, and hence $[m(z_1), m(z_2); m(z_3), m(z_4)]$, is real.

As $[z_1, z_2; z_3, z_4] = [m(z_1), m(z_2); m(z_3), m(z_4)]$ and as Möb^+ takes circles in $\overline{\mathbb{C}}$ to circles in $\overline{\mathbb{C}}$, we have that z_1, z_2, z_3, and z_4 lie on a circle in $\overline{\mathbb{C}}$ if and only if $[z_1, z_2; z_3, z_4]$ is real. This completes the proof of Proposition 2.14. **QED**

Exercise 2.14

Determine whether $2 + 3i$, $-2i$, $1 - i$, and 4 lie on a circle in $\overline{\mathbb{C}}$.

Exercise 2.15

Determine the real values of s for which the points $2 + 3i$, $-2i$, $1 - i$, and s lie on a circle in $\overline{\mathbb{C}}$.

Some amount of choice is inherent in the definition of the cross ratio, as we are free to permute the coordinates. For instance, we can also consider the cross ratios

$$[z_1, z_2; z_3, z_4]_2 = \frac{(z_1 - z_2)}{(z_1 - z_4)} \frac{(z_3 - z_4)}{(z_3 - z_2)}$$

and

$$[z_1, z_2; z_3, z_4]_3 = \frac{(z_2 - z_1)}{(z_2 - z_4)} \frac{(z_3 - z_4)}{(z_3 - z_1)}.$$

We note here that all possible choices of cross ratio, such as the three described in this section, are all closely related.

Exercise 2.16

Express the two cross ratios $[z_1, z_2; z_3, z_4]_2$ and $[z_1, z_2; z_3, z_4]_3$ in terms of the standard cross ratio $[z_1, z_2; z_3, z_4]$.

2.4 Classification of Möbius Transformations

The classification of Möbius transformations given in Section 2.1, in terms of the number of fixed points, is as we wrote at the time crude and can be considerably refined.

Before getting into the refinement of this classification, we introduce a notion of sameness for Möbius transformations.

Definition 2.15

Two Möbius transformations m_1 and m_2 are *conjugate* if there exists some Möbius transformation p so that $m_2 = p \circ m_1 \circ p^{-1}$.

Geometrically, m_1 and m_2 are conjugate by p if the action of m_1 on $\overline{\mathbb{C}}$ is the same as the action of m_2 on $p(\overline{\mathbb{C}}) = \overline{\mathbb{C}}$. That is, conjugacy reflects a change of coordinates on $\overline{\mathbb{C}}$.

Exercise 2.17

Suppose that m and n are Möbius transformations that are conjugate by p, so that $m = p \circ n \circ p^{-1}$. Prove that m and n have the same number of fixed points in $\overline{\mathbb{C}}$.

The basic idea of the classification of Möbius transformations is to conjugate a given Möbius transformation into a standard form and then to classify the possible standard forms. For the remainder of this section, we work with a Möbius transformation m that is not the identity.

Suppose that m has only one fixed point in $\overline{\mathbb{C}}$, and call it x. Let y be any point of $\overline{\mathbb{C}} - \{x\}$, and observe that $(x, y, m(y))$ is a triple of distinct points of $\overline{\mathbb{C}}$. Let p be the Möbius transformation taking the triple $(x, y, m(y))$ to the triple $(\infty, 0, 1)$, and consider the composition $p \circ m \circ p^{-1}$.

By our construction of p, we have that $p \circ m \circ p^{-1}(\infty) = p \circ m(x) = p(x) = \infty$. As $p \circ m \circ p^{-1}$ fixes ∞, we can write it as $p \circ m \circ p^{-1}(z) = az + b$ with $a \neq 0$. As $p \circ m \circ p^{-1}$ has only the one fixed point in $\overline{\mathbb{C}}$, namely, ∞, there is no solution in $\mathbb{C}$ to the equation $p \circ m \circ p^{-1}(z) = z$, and so it must be that $a = 1$.

As $p \circ m \circ p^{-1}(0) = p \circ m(y) = 1$, we see that $b = 1$ as well, and so $p \circ m \circ p^{-1}(z) = z + 1$. Therefore, any Möbius transformation m with only one fixed point is conjugate by a Möbius transformation to $n(z) = z + 1$. We say that m is *parabolic*, and we refer to $p \circ m \circ p^{-1}(z) = z + 1$ as its *standard form*.

To consider a specific example, let $m(z) = \frac{z}{z+1}$. As $m(\infty) = 1 \neq \infty$, the fixed points of m are the solutions in $\mathbb{C}$ to the equation $m(z) = \frac{z}{z+1} = z$, which are the solutions to $z = z^2 + z$. Hence, the only fixed point of m is 0.

To find the Möbius transformation p conjugating m to its standard form, choose some point in $\overline{\mathbb{C}} - \{0\}$, say ∞, and calculate that $m(\infty) = 1$. Then, we take

p to be the Möbius transformation sending the triple $(0, \infty, 1)$ to the triple $(\infty, 0, 1)$, namely, $p(z) = \frac{iz}{i} = \frac{1}{z}$.

In the argument just given, there is some ambiguity in the choice of the conjugating Möbius transformation p, as the specific form of p depends on the choice of the point y not fixed by m. However, this choice does not play an essential role.

Suppose now that m has two fixed points in $\overline{\mathbb{C}}$, and call them x and y. Let q be a Möbius transformation satisfying $q(x) = 0$ and $q(y) = \infty$, and consider the composition $q \circ m \circ q^{-1}$.

By definition, we have that $q \circ m \circ q^{-1}(\infty) = q \circ m(y) = q(y) = \infty$, and that $q \circ m \circ q^{-1}(0) = q \circ m(x) = q(x) = 0$, and so we may write $q \circ m \circ q^{-1}(z) = az$ for some $a \in \mathbb{C} - \{0,\ 1\}$. We refer to a as the *multiplier of* m.

To consider a specific example, let $m(z) = \frac{2z+1}{z+1}$. As $m(\infty) = 2 \neq \infty$, the fixed points of m are the solutions in $\mathbb{C}$ to the equation $m(z) = \frac{2z+1}{z+1} = z$, which are the solutions to $z^2 - z - 1 = 0$. Using the quadratic formula, we see that the fixed points of m are $z = \frac{1}{2}(1 \pm \sqrt{5})$.

To find the Möbius transformation q conjugating m to its standard form, consider a Möbius transformation q taking $\frac{1}{2}(1 + \sqrt{5})$ to 0 and taking $\frac{1}{2}(1 - \sqrt{5})$ to ∞; for instance,

$$q(z) = \frac{z - \frac{1}{2}(1 + \sqrt{5})}{z - \frac{1}{2}(1 - \sqrt{5})}.$$

At this point, instead of calculating out the composition $q \circ m \circ q^{-1}$ explicitly, which we can certainly do, we calculate the multiplier of m by calculating the single value

$$a = q \circ m \circ q^{-1}(1) = q \circ m(\infty) = q(2) = \frac{3 - \sqrt{5}}{3 + \sqrt{5}}.$$

As in the argument for parabolic Möbius transformations, there is some ambiguity in the choice of the conjugating Möbius transformation q, as there is not enough information to specify q uniquely. However, as is seen in the following two exercises, this choice does not play an essential role.

Exercise 2.18

Let m be a Möbius transformation with two fixed points x and y. Prove that if n_1 and n_2 are two Möbius transformations satisfying $n_1(x) = 0 = n_2(x)$ and $n_1(y) = \infty = n_2(y)$, then the multipliers of $n_1 \circ m \circ n_1^{-1}$ and $n_2 \circ m \circ n_2^{-1}$ are equal.

Exercise 2.19

Using the notation of the argument just given for Möbius transformations with two fixed points, prove that if we conjugate m as above by a Möbius transformation s satisfying $s(x) = \infty$ and $s(y) = 0$, the multiplier of $s \circ m \circ s^{-1}$ is $\frac{1}{a}$.

As a corollary to Exercises 2.18 and 2.19, we see that the multiplier of a Möbius transformation with two fixed points is only defined up to taking its inverse. Moreover, the solution to Exercise 2.19 shows that $J(z) = \frac{1}{z}$ conjugates $m(z) = az$ to $m^{-1}(z) = \frac{1}{a}z$.

If the multiplier of m satisfies $|a| = 1$, then we may write $a = e^{2i\theta}$ for some θ in $(0, \pi)$, and $q \circ m \circ q^{-1}(z) = e^{2i\theta}z$ is rotation about the origin by angle 2θ. We say that m is *elliptic*, and we refer to $q \circ m \circ q^{-1}(z) = e^{2i\theta}z$ as its *standard form.*

If, on the other hand, $|a| \neq 1$, then we may write $a = \rho^2 e^{2i\theta}$ for some positive real number $\rho \neq 1$ and some θ in $[0, \pi)$, so that $q \circ m \circ q^{-1}(z) = \rho^2 e^{2i\theta}z$ is the composition of a dilation by ρ^2 (an expansion if $\rho^2 > 1$ or a contraction if $\rho^2 < 1$) and a (possibly trivial) rotation about the origin by angle 2θ. We say that m is *loxodromic*, and we refer to $q \circ m \circ q^{-1}(z) = \rho^2 e^{2i\theta}z$ as its *standard form.*

Exercise 2.20

Determine the type, parabolic, elliptic, or loxodromic, of each Möbius transformation given in Exercise 2.7.

The name loxodromic comes from the word *loxodrome*, which is a curve on the sphere that meets every line of latitude at the same angle. Lines of longitude are loxodromes, but there are also loxodromes that spiral into both poles. The reason these Möbius transformations are called loxodromic is that each one keeps invariant a loxodrome.

2.5 A Matrix Representation

If we examine the formula for the composition of two Möbius transformations, we get a hint that there is a strong connection between Möbius transformations and 2×2 matrices. Consider the Möbius transformations $m(z) = \frac{az+b}{cz+d}$ and $n(z) = \frac{\alpha z+\beta}{\gamma z+\delta}$. Then,

$$\begin{aligned} n \circ m(z) = \frac{\alpha m(z) + \beta}{\gamma m(z) + \delta} &= \frac{\alpha\left(\frac{az+b}{cz+d}\right) + \beta}{\gamma\left(\frac{az+b}{cz+d}\right) + \delta} \\ &= \frac{\alpha(az+b) + \beta(cz+d)}{\gamma(az+b) + \delta(cz+d)} \\ &= \frac{(\alpha a + \beta c)z + \alpha b + \beta d}{(\gamma a + \delta c)z + \gamma b + \delta d}. \end{aligned}$$

If instead we view the coefficients of m and n as the entries in a pair of 2×2 matrices and multiply, we get

$$\begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} \alpha a + \beta c & \alpha b + \beta d \\ \gamma a + \delta c & \gamma b + \delta d \end{pmatrix},$$

and the entries of the product matrix correspond to the coefficients of the composition $n \circ m$.

We will examine the details of this correspondence between Möbius transformations and matrices later in the section. For the moment, let us concentrate on using this similarity to refine further the classification of Möbius transformations we discussed in Section 2.4.

We can associate two main numerical quantities with a 2×2 matrix, the *determinant* and the *trace*. Using this correspondence between matrices and Möbius transformations, we can define similar notions for Möbius transformations.

Define the *determinant* $\det(m)$ of the Möbius transformation $m(z) = \frac{az+b}{cz+d}$ to be the quantity $\det(m) = ad - bc$. Note that the determinant of a Möbius transformation is not a well defined quantity. If we multiply the coefficients of m by any nonzero constant, this has no effect on the action of m on $\overline{\mathbb{C}}$, because

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

for all $\alpha \in \mathbb{C} - \{0\}$ and all $z \in \overline{\mathbb{C}}$. However, the determinants are not equal, because the determinant of $f(z) = \frac{az+b}{cz+d}$ is $\det(f) = ad - bc$ and the determinant of $g(z) = \frac{\alpha az+\alpha b}{\alpha cz+\alpha d}$ is $\det(g) = \alpha^2(ad - bc)$.

Exercise 2.21

Calculate the determinants of the following Möbius transformations:

1. $m(z) = \frac{2z+4}{5z-7}$; 2. $m(z) = \frac{1}{z}$; 3. $m(z) = \frac{-z-3}{z+1}$;

4. $m(z) = \frac{iz+1}{z+3i}$; 5. $m(z) = iz + 1$; 6. $m(z) = \frac{-z}{z+4}$.

However, we can always choose α so that the determinant of m is 1. This still leaves a small amount of ambiguity, because all coefficients of m can be multiplied by -1 without changing the determinant of m, but this is the only remaining ambiguity. We refer to this process as *normalizing* m.

Exercise 2.22

Normalize each of the Möbius transformations from Exercise 2.21.

Having normalized a Möbius transformation m, there is another useful numerical quantity associated to m, which corresponds to taking the trace. Consider the function

$$\tau : \text{Möb}^{+} \to \mathbb{C}$$

defined by setting $\tau(m) = (a+d)^2$, where $m(z) = \frac{az+b}{cz+d}$ is normalized. As the only ambiguity in the definition of a normalized Möbius transformation arises from multiplying all coefficients by -1, we see that $\tau(m)$ is well defined. In fact, this possible ambiguity is why we consider the function τ and not the actual trace $\text{trace}(m) = a + d$.

As with the trace of a matrix, one useful property of τ is that it is invariant under conjugation.

Exercise 2.23

Show that $\tau(m \circ n) = \tau(n \circ m)$.

Exercise 2.24

Show that $\tau(p \circ m \circ p^{-1}) = \tau(m)$.

Using this invariance of τ under conjugation, we can distinguish the different types of Möbius transformations without explicitly conjugating them to their standard forms. namely, let m be a Möbius transformation, and let p be a Möbius transformation conjugating m to its standard form. As

$$\tau(m) = \tau(p \circ m \circ p^{-1}),$$

it suffices to consider the values of τ on the standard forms.

If m is parabolic, then $p \circ m \circ p^{-1}(z) = z + 1$, which is normalized, and so

$$\tau(m) = \tau(p \circ m \circ p^{-1}) = (1+1)^2 = 4.$$

Note that, for the identity Möbius transformation $e(z) = z$, we also have that $\tau(e) = (1+1)^2 = 4$.

If m is either elliptic or loxodromic, we may write $n(z) = p \circ m \circ p^{-1}(z) = \alpha^2 z$, where $\alpha^2 \in \mathbb{C} - \{0,\ 1\}$. When we normalize so that the determinant of n (or equivalently of m) is 1, we need to write

$$n(z) = \frac{\alpha z}{\alpha^{-1}},$$

and so

$$\tau(n) = \tau(p \circ m \circ p^{-1}) = \tau(m) = (\alpha + \alpha^{-1})^2.$$

In the case in which m is elliptic, so that $|\alpha| = 1$, write $\alpha = e^{i\theta}$ for some θ in $(0, \pi)$. Calculating, we see that

$$\tau(m) = (\alpha + \alpha^{-1})^2 = \left(e^{i\theta} + e^{-i\theta}\right)^2 = 4\cos^2(\theta).$$

In particular, we have that $\tau(m)$ is real and lies in the interval $[0, 4)$.

In the case in which m is loxodromic, so that $|\alpha| \neq 1$, we write $\alpha = \rho e^{i\theta}$ for some $\rho > 0$, $\rho \neq 1$, and some θ in $[0, \pi)$. Calculating, we see that

$$\alpha + \alpha^{-1} = \rho e^{i\theta} + \rho^{-1} e^{-i\theta},$$

and so

$$\tau(m) = (\alpha + \alpha^{-1})^2 = \cos(2\theta)(\rho^2 + \rho^{-2}) + 2 + i\sin(2\theta)(\rho^2 - \rho^{-2}).$$

In particular, because $\rho \neq 1$, we see that $\mathrm{Im}(\tau(m)) \neq 0$ for $\theta \neq 0$ and $\theta \neq \frac{\pi}{2}$.

For the two cases in which $\theta = 0$ and $\theta = \frac{\pi}{2}$, we use the following exercise from calculus.

Exercise 2.25

Show that the function $f : (0, \infty) \to \mathbb{R}$ defined by $f(\rho) = \rho^2 + \rho^{-2}$ satisfies $f(\rho) \geq 2$, with $f(\rho) = 2$ if and only if $\rho = 1$.

For $\theta = 0$, we see that $\tau(m) > 4$, whereas for $\theta = \frac{\pi}{2}$, we see that $\tau(m) < 0$.

To summarize, we have shown the following.

Proposition 2.16

Let m be a Möbius transformation other than the identity. Then,

1. m is parabolic if and only if $\tau(m) = 4$.

2. m is elliptic if and only if $\tau(m)$ is real and lies in $[0, 4)$.

3. m is loxodromic if and only if either $\tau(m)$ has nonzero imaginary part or $\tau(m)$ is real and lies in $(-\infty, 0) \cup (4, \infty)$.

To work through a specific example, consider $m(z) = \frac{z+1}{z+3}$. The determinant of m is $3 - 1 = 2$, and so the normalized form of m is

$$m(z) = \frac{\frac{1}{\sqrt{2}}z + \frac{1}{\sqrt{2}}}{\frac{1}{\sqrt{2}}z + \frac{3}{\sqrt{2}}}.$$

Calculating, we see that $\tau(m) = 8$, and so m is loxodromic.

Note that we can determine the multiplier of an elliptic or loxodromic transformation m, up to taking its inverse, knowing only the value of $\tau(m)$. Specifically, if m has multiplier α^2, then

$$\tau(m) = (\alpha + \alpha^{-1})^2 = \alpha^2 + \alpha^{-2} + 2.$$

Multiplying through by α^2 and collecting terms gives

$$\alpha^4 + (2 - \tau(m))\alpha^2 + 1 = 0.$$

Applying the quadratic formula, we obtain

$$\begin{aligned} \alpha^2 &= \frac{1}{2}\left[\tau(m) - 2 \pm \sqrt{(2 - \tau(m))^2 - 4}\right] \\ &= \frac{1}{2}\left[\tau(m) - 2 \pm \sqrt{-4\tau(m) + \tau^2(m)}\right]. \end{aligned}$$

As

$$\frac{1}{2}\left[\tau(m) - 2 + \sqrt{-4\tau(m) + \tau^2(m)}\right] \cdot \frac{1}{2}\left[\tau(m) - 2 - \sqrt{-4\tau(m) + \tau^2(m)}\right] = 1,$$

we may take the multiplier α^2 to satisfy $|\alpha|^2 \geq 1$.

Exercise 2.26

Determine the type of each Möbius transformation from Exercise 2.21. If the transformation is elliptic or loxodromic, determine its multiplier.

Exercise 2.27

Show that if m is a parabolic Möbius transformation with fixed point $x \neq \infty$, then there exists a unique complex number p so that

$$m(z) = \frac{(1+px)z - px^2}{pz + 1 - px}.$$

Exercise 2.28

Show that if m is a Möbius transformation with distinct fixed points $x \neq \infty$ and $y \neq \infty$ and multiplier a, then we can write

$$m(z) = \frac{\left(\frac{x-ya}{x-y}\right) z + \frac{xy(a-1)}{x-y}}{\left(\frac{1-a}{x-y}\right) z + \frac{xa-y}{x-y}}.$$

We close this section by making explicit the correspondence between Möbius transformations and 2 × 2 matrices. To set notation, let

$$\mathrm{GL}_2(\mathbb{C}) = \left\{ \left(\begin{array}{cc} a & b \\ c & d \end{array} \right) \middle| \; a,\, b,\, c,\, d \in \mathbb{C} \text{ and } ad - bc \neq 0 \right\},$$

and let

$$\mathrm{SL}_2(\mathbb{C}) = \left\{ \left(\begin{array}{cc} a & b \\ c & d \end{array} \right) \middle| \; a,\, b,\, c,\, d \in \mathbb{C} \text{ and } ad - bc = 1 \right\},$$

We have already seen, in our discussion of normalization, that a Möbius transformation determines many matrices, so the obvious guess of a function from Möb^+ to $\mathrm{GL}_2(\mathbb{C})$ is not well defined.

So we go the other way and consider the obvious choice of a function from $\mathrm{GL}_2(\mathbb{C})$ to Möb^+. Define $\mu : \mathrm{GL}_2(\mathbb{C}) \to \text{Möb}^+$ by

$$\mu\left(M = \left(\begin{array}{cc} a & b \\ c & d \end{array} \right) \right) = \left(m(z) = \frac{az+b}{cz+d} \right).$$

Note that the calculation performed at the beginning of this section proves that μ is a homomorphism.

Exercise 2.29

Prove that the kernel $\ker(\mu)$ of μ is the subgroup $K = \{\alpha \mathrm{I} \mid \alpha \in \mathbb{C}\}$ of $\mathrm{GL}_2(\mathbb{C})$. Conclude that Möb^+ is isomorphic to $\mathrm{PGL}_2(\mathbb{C}) = \mathrm{GL}_2(\mathbb{C})/K$.

2.6 Reflections

We have seen, in Theorem 2.5, that Möb^+ is contained in the set $\text{Homeo}^{\mathrm{C}}(\overline{\mathbb{C}})$ of homeomorphisms of $\overline{\mathbb{C}}$ that take circles in $\overline{\mathbb{C}}$ to circles in $\overline{\mathbb{C}}$. There is a natural extension of Möb^+ that also lies in $\text{Homeo}^{\mathrm{C}}(\overline{\mathbb{C}})$.

To extend Möb^+ to a larger group, we consider the simplest homeomorphism of $\overline{\mathbb{C}}$ not already in Möb^+, namely, *complex conjugation.* Set

$$C(z) = \overline{z} \text{ for } z \in \mathbb{C} \text{ and } C(\infty) = \infty.$$

Proposition 2.17

The function $C : \overline{\mathbb{C}} \to \overline{\mathbb{C}}$ defined by

$$C(z) = \overline{z} \text{ for } z \in \mathbb{C} \text{ and } C(\infty) = \infty$$

is an element of $\text{Homeo}(\overline{\mathbb{C}})$.

Proof

Note that C is its own inverse; that is,nontrivial $C^{-1}(z) = C(z)$, and hence C is a bijection of $\overline{\mathbb{C}}$. So, we need only check that C is continuous.

The continuity of C follows from the observation that for any point $z \in \overline{\mathbb{C}}$ and any $\varepsilon > 0$, we have that $C(U_\varepsilon(z)) = U_\varepsilon(C(z))$. This completes the proof of Proposition 2.17. **QED**

Exercise 2.30

Show that C is not an element of Möb^+.

Definition 2.18

The *general Möbius group* Möb is the group generated by Möb^+ and C. That is, every (nontrivial) element p of Möb can be expressed as a composition

$$p = C \circ m_k \circ \cdots C \circ m_1$$

for some $k \geq 1$, where each m_k is an element of Möb^+.

Note that because Möb contains Möb^+, all transitivity properties of Möb^+ discussed in Section 2.2 are inherited by Möb. That is, Möb acts transitively on the set $\mathcal{T}$ of triples of distinct points in $\overline{\mathbb{C}}$, on the set $\mathcal{C}$ of circles in $\overline{\mathbb{C}}$, and on the set $\mathcal{D}$ of discs in $\overline{\mathbb{C}}$. Unfortunately, though, Möb does not inherit unique transitively on triples of distinct points, as we saw in the solution to Exercise 2.30.

The proof that $C : \overline{\mathbb{C}} \to \overline{\mathbb{C}}$ lies in $\text{Homeo}^{\text{C}}(\overline{\mathbb{C}})$ is similar to the proof that the elements of Möb^+ lie in $\text{Homeo}^{\text{C}}(\overline{\mathbb{C}})$.

Exercise 2.31

Show that the function $C : \overline{\mathbb{C}} \to \overline{\mathbb{C}}$ lies in $\text{Homeo}^{\text{C}}(\overline{\mathbb{C}})$.

Exercise 2.31 and Theorem 2.5 combine to give the following theorem.

Theorem 2.19

$\text{Möb} \subset \text{Homeo}^{\text{C}}(\overline{\mathbb{C}})$.

We can also write explicit expressions for every element of Möb.

Exercise 2.32

Show that every element of Möb has either the form

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

or the form

$$n(z) = \frac{a\overline{z}+b}{c\overline{z}+d},$$

where a, b, c, $d \in \mathbb{C}$ and $ad - bc \neq 0$.

Geometrically, the action of C on $\overline{\mathbb{C}}$ is *reflection* in the extended real axis $\overline{\mathbb{R}}$. That is, every point of $\overline{\mathbb{R}}$ is fixed by C, and every point z of $\overline{\mathbb{C}} - \overline{\mathbb{R}} = \mathbb{C} - \mathbb{R}$ has the property that $\mathbb{R}$ is the perpendicular bisector of the Euclidean line segment joining z and $C(z)$.

Given that we have defined reflection in the specific circle $\overline{\mathbb{R}}$, and given that Möb acts transitively on the set $\mathcal{C}$ of circles in $\overline{\mathbb{C}}$, we may define *reflection* in any circle in $\overline{\mathbb{C}}$. Specifically, for a circle in $\overline{\mathbb{C}}$ A, we choose an element m of Möb taking $\overline{\mathbb{R}}$ to A, and define *reflection in* A to be the composition $C_A = m \circ C \circ m^{-1}$. Note that there is some potential for ambiguity in this definition of C_A, in that there are many choices for the transformation m. We will show in Section 2.8 that reflection in A is well defined.

For example, consider $A = \mathbb{S}^1$. One element of Möb$^+$ taking $\overline{\mathbb{R}}$ to $\mathbb{S}^1$ is the transformation taking the triple $(0, 1, \infty)$ to the triple $(i, 1, -i)$, namely,

$$m(z) = \frac{\frac{1}{\sqrt{2}}z + \frac{i}{\sqrt{2}}}{\frac{i}{\sqrt{2}}z + \frac{1}{\sqrt{2}}}.$$

Calculating, we see that

$$C_A(z) = m \circ C \circ m^{-1}(z) = \frac{1}{\overline{z}} = \frac{z}{|z|^2}.$$

Exercise 2.33

Write explicit expressions for two distinct elements p and n of Möb taking $\overline{\mathbb{R}}$ to $\mathbb{S}^1$. Show that $p \circ C \circ p^{-1} = n \circ C \circ n^{-1}$.

In the case in which A is the Euclidean circle in $\mathbb{C}$ with centre α and radius ρ, we may conjugate reflection in $\mathbb{S}^1$, namely, $c(z) = \frac{1}{\overline{z}}$, by the Möbius transformation p taking $\mathbb{S}^1$ to A, namely, $p(z) = \rho z + \alpha$, to obtain an explicit expression for the reflection C_A in A, namely,

$$C_A(z) = p \circ c \circ p^{-1}(z) = \frac{\rho^2}{\overline{z} - \overline{\alpha}} + \alpha.$$

Similarly, if A is the Euclidean line in $\mathbb{C}$ passing through α and making angle θ with $\mathbb{R}$, we may conjugate reflection in $\mathbb{R}$, namely, $C(z) = \overline{z}$, by the Möbius transformation p taking $\mathbb{R}$ to A, namely, $p(z) = e^{i\theta}z + \alpha$, to obtain an explicit expression for the reflection C_A in A, namely,

$$C_A(z) = p \circ C \circ p^{-1}(z) = e^{2i\theta}(\overline{z} - \overline{\alpha}) + \alpha.$$

This construction of reflections in circles in $\overline{\mathbb{C}}$ has the following consequence.

Proposition 2.20

Every element of Möb can be expressed as the composition of reflections in finitely many circles in $\overline{\mathbb{C}}$.

Proof

As Möb is generated by Möb^+ and $C(z) = \overline{z}$, and as Möb^+ is generated by $J(z) = \frac{1}{z}$ and the $f(z) = az + b$ for $a, b \in \mathbb{C}$ with $a \neq 0$, we need only verify the proposition for these transformations.

By definition, C is reflection in $\overline{\mathbb{R}}$. We can express J as the composition of $C(z) = \overline{z}$ and the reflection $c(z) = \frac{1}{\overline{z}}$ in $\mathbb{S}^1$. Hence, the proof of Proposition 2.20 is completed by the following exercise.

Exercise 2.34

Express every element of Möb^+ of the form $f(z) = az + b$ for $a, b \in \mathbb{C}$ with $a \neq 0$, as the composition of reflections in finitely many circles in $\overline{\mathbb{C}}$.

This completes the proof of Proposition 2.20. **QED**

Over the past several sections, we have seen that the elements of Möb are homeomorphisms of $\overline{\mathbb{C}}$ that take circles in $\overline{\mathbb{C}}$ to circles in $\overline{\mathbb{C}}$. In fact, this property characterizes Möb.

Theorem 2.21

$\text{Möb} = \text{Homeo}^{\text{C}}(\overline{\mathbb{C}})$.

Proof

We give a sketch of the proof of Theorem 2.21. By Theorem 2.19, we already have that $\text{Möb} \subset \text{Homeo}^{\text{C}}(\overline{\mathbb{C}})$, and so it remains only to show the opposite inclusion, that $\text{Homeo}^{\text{C}}(\overline{\mathbb{C}}) \subset \text{Möb}$.

So, let f be an element of $\text{Homeo}^{\text{C}}(\overline{\mathbb{C}})$. Let p be the Möbius transformation taking the triple $(f(0), f(1), f(\infty))$ to the triple $(0, 1, \infty)$, so that $p \circ f$ satisfies $p \circ f(0) = 0$, $p \circ f(1) = 1$, and $p \circ f(\infty) = \infty$.

As $p \circ f$ takes circles in $\overline{\mathbb{C}}$ to circles in $\overline{\mathbb{C}}$, it must be that $p \circ f(\mathbb{R}) = \mathbb{R}$, because $p \circ f(\infty) = \infty$ and $\overline{\mathbb{R}}$ is the circle in $\overline{\mathbb{C}}$ determined by the triple $(0, 1, \infty)$.

As $p \circ f$ fixes ∞ and takes $\mathbb{R}$ to $\mathbb{R}$, either $p \circ f(\mathbb{H}) = \mathbb{H}$ or $p \circ f(\mathbb{H})$ is the lower half-plane. In the former case, set $m = p$. In the latter case, set $m = C \circ p$, where $C(z) = \overline{z}$ is complex conjugation.

We now have an element m of Möb so that $m \circ f(0) = 0$, $m \circ f(1) = 1$, $m \circ f(\infty) = \infty$, and $m \circ f(\mathbb{H}) = \mathbb{H}$. We show that $m \circ f$ is the identity. We do this by constructing a dense set of points in $\overline{\mathbb{C}}$, each of which is fixed by $m \circ f$.

Before beginning the construction of this dense set, we introduce a bit of notation. Set $Z = \{z \in \overline{\mathbb{C}} \,|\, m \circ f(z) = z\}$ to be the set of points of $\overline{\mathbb{C}}$ fixed by $m \circ f$. By our choice of m, we have that 0, 1, and ∞ are points of Z.

As $m \circ f$ fixes ∞ and lies in $\text{Homeo}^{\text{C}}(\overline{\mathbb{C}})$, we see that $m \circ f$ takes Euclidean lines in $\mathbb{C}$ to Euclidean lines in $\mathbb{C}$, and it takes Euclidean circles in $\mathbb{C}$ to Euclidean circles in $\mathbb{C}$. Also, if X and Y are two Euclidean lines in $\mathbb{C}$ that intersect at some point z_0, and if $m \circ f(X) = X$ and $m \circ f(Y) = Y$, then $m \circ f(z_0) = z_0$ and so z_0 is contained in this set Z of points fixed by $m \circ f$.

For each $s \in \mathbb{R}$, let $V(s)$ be the vertical line in $\mathbb{C}$ through s and let $H(s)$ be the horizontal line in $\mathbb{C}$ through is.

Let H be any horizontal line in $\mathbb{C}$. As $m \circ f(\mathbb{R}) = \mathbb{R}$ and as H and $\mathbb{R}$ are disjoint, we see that $m \circ f(H)$ and $m \circ f(\mathbb{R}) = \mathbb{R}$ are disjoint lines in $\mathbb{C}$, and so H is again a horizontal line in $\mathbb{C}$. Also, as $m \circ f(\mathbb{H}) = \mathbb{H}$, we have that H lies in $\mathbb{H}$ if and only if $m \circ f(H)$ lies in $\mathbb{H}$.

Let A be the Euclidean circle with Euclidean centre $\frac{1}{2}$ and Euclidean radius $\frac{1}{2}$. As $V(0)$ is tangent to A at 0, we see that $m \circ f(V(0))$ is the tangent line to $m \circ f(A)$ at $m \circ f(0) = 0$, and similarly that $m \circ f(V(1))$ is the tangent line to $m \circ f(A)$ at 1.

As $V(0)$ and $V(1)$ are parallel Euclidean lines in $\mathbb{C}$, we see that $m \circ f(V(0))$ and $m \circ f(V(1))$ are also parallel Euclidean lines in $\mathbb{C}$, and so we must have that $m \circ f(V(0)) = V(0)$ and $m \circ f(V(1)) = V(1)$.

In particular, this process forces $m \circ f(A) = A$, as the tangent lines through 0 and 1 to any other Euclidean circle passing through 0 and 1 are not parallel. However, even though $m \circ f(A) = A$, we do not yet know that A contains any points of Z other than 0 and 1.

However, we can run the same argument with the two horizontal tangent lines to A. Consider first the tangent line $H(\frac{1}{2})$ to A at $\frac{1}{2}+\frac{1}{2}i$. As $m\circ f(H(\frac{1}{2}))$ is again a horizontal line in $\mathbb{H}$ tangent to $m\circ f(A) = A$, we see that $m\circ f(H(\frac{1}{2})) = H(\frac{1}{2})$.

We now have more points in Z. namely, the intersections $H(\frac{1}{2})\cap V(0) = \frac{1}{2}i$ and $H(\frac{1}{2})\cap V(1) = 1+\frac{1}{2}i$ lie in Z. The same argument gives that $m\circ f(H(-\frac{1}{2})) = H(-\frac{1}{2})$, and hence that both $H(-\frac{1}{2})\cap V(0) = -\frac{1}{2}i$ and $H(-\frac{1}{2})\cap V(1) = 1-\frac{1}{2}i$ lie in Z.

Each pair of points in Z gives rise to a Euclidean line that is taken to itself by $m\circ f$, and each triple of noncolinear points in Z gives rise to a Euclidean circle that is taken to itself by $m\circ f$. The intersections of these Euclidean lines and Euclidean circles give rise to more points of Z, which in turn give rise to more Euclidean lines and Euclidean circles taken to themselves, and so on.

Continuing on, this process yields that Z contains a dense set of points of $\overline{\mathbb{C}}$, which in turn implies that $m\circ f$ is the identity, by Exercise 1.16. Hence, $f = m^{-1}$ is an element of Möb. This completes the sketch of the proof of Theorem 2.21. **QED**

2.7 The Conformality of Elements of Möb

In this section, we describe the last major property of Möb that we will make use of. We begin with a definition.

Definition 2.22

Given two smooth curves C_1 and C_2 in $\mathbb{C}$ that intersect at a point z_0, define the *angle* angle(C_1, C_2) *between* C_1 *and* C_2 at z_0 to be the angle between the tangent lines to C_1 and C_2 at z_0, measured from C_1 to C_2.

In our measurement of angle, we adopt the convention that counterclockwise angles are positive and clockwise angles are negative. By this definition of angle, we have that

$$\text{angle}(C_2, C_1) = -\text{angle}(C_1, C_2).$$

Note that angle as we have defined it is not a well defined notion, but instead it is defined only up to additive multiples of π. If we were to be formal, we would really need to define angle to take its values in $\mathbb{R}/\pi\mathbb{Z}$. However, this ambiguity in the definition of angle causes us no difficulty in this section.

A homeomorphism of $\overline{\mathbb{C}}$ that preserves the absolute value of the angle between curves is said to be *conformal*. We note that this usage is slightly nonstandard, as many authors use conformal to mean that the actual angles, and not merely the absolute values of the angles, are preserved.

The last major fact about Möb we need to establish is that the elements of Möb are conformal. The proof we give here is analytic. Although we do not give it here, it is possible to give a geometric proof using stereographic projection. See, for example, Jones and Singerman [23].

Theorem 2.23

The elements of Möb are conformal homeomorphisms of $\overline{\mathbb{C}}$.

Proof

The proof of Theorem 2.23 contains several calculations left for the interested reader.

As the angle between two curves is by definition the angle between their tangent lines, it suffices to check whether the angle $\text{angle}(X_1, X_2)$ between X_1 and X_2 is equal to the angle $\text{angle}(m(X_1), m(X_2))$ between $m(X_1)$ and $m(X_2)$, where X_1 and X_2 are Euclidean lines in $\mathbb{C}$.

So, let X_1 and X_2 be two Euclidean lines in $\mathbb{C}$ that intersect at a point z_0, let z_k be a point on X_k not equal to z_0, and let s_k be the slope of X_k. These quantities are connected by the equation

$$s_k = \frac{\text{Im}(z_k - z_0)}{\text{Re}(z_k - z_0)}.$$

Let θ_k be the angle that X_k makes with the real axis $\mathbb{R}$, and note that

$$s_k = \tan(\theta_k).$$

In particular, the angle $\text{angle}(X_1, X_2)$ between X_1 and X_2 is given by

$$\text{angle}(X_1, X_2) = \theta_2 - \theta_1 = \arctan(s_2) - \arctan(s_1).$$

We know from Section 2.6 that Möb is generated by the transformations of the form $f(z) = az + b$ for a, $b \in \mathbb{C}$ and $a \neq 0$, as well as the two transformations $J(z) = \frac{1}{z}$ and $C(z) = \overline{z}$. We take each in turn.

Consider $f(z) = az + b$, where a, $b \in \mathbb{C}$ and $a \neq 0$. Write $a = \rho e^{i\beta}$. As $f(\infty) = \infty$, both $f(X_1)$ and $f(X_1)$ are again Euclidean lines in $\mathbb{C}$. As $f(X_k)$ passes through the points $f(z_0)$ and $f(z_k)$, the slope t_k of the Euclidean line $f(X_k)$ is

$$\begin{aligned} t_k = \frac{\mathrm{Im}(f(z_k) - f(z_0))}{\mathrm{Re}(f(z_k) - f(z_0))} &= \frac{\mathrm{Im}(a(z_k - z_0))}{\mathrm{Re}(a(z_k - z_0))} \\ &= \frac{\mathrm{Im}(e^{i\beta}(z_k - z_0))}{\mathrm{Re}(e^{i\beta}(z_k - z_0))} = \tan(\beta + \theta_k). \end{aligned}$$

In particular, we see that

$$\begin{aligned} \mathrm{angle}(f(X_1), f(X_2)) &= \arctan(t_2) - \arctan(t_1) \\ &= (\beta + \theta_2) - (\beta + \theta_1) \\ &= \theta_2 - \theta_1 = \mathrm{angle}(X_1, X_2), \end{aligned}$$

and so m is conformal.

Consider now $J(z) = \frac{1}{z}$. Here, we need to take a slightly different approach, because $J(X_1)$ and $J(X_2)$ no longer need be Euclidean lines in $\mathbb{C}$, but instead they may both be Euclidean circles in $\mathbb{C}$ that intersect at 0, or one might be a Euclidean line and the other a Euclidean circle. We work through the case in which both are Euclidean circles and leave the other cases for the interested reader.

So, we may suppose that X_k is given as the set of solutions to the equation

$$\beta_k z + \overline{\beta_k}\overline{z} + 1 = 0,$$

where $\beta_k \in \mathbb{C}$. The slope of X_k is then given by

$$s_k = \frac{\mathrm{Re}(\beta_k)}{\mathrm{Im}(\beta_k)}.$$

Given the form of the equation for X_k, we also know that $J(X_k)$ is the set of solutions to the equation

$$z\overline{z} + \overline{\beta_k} z + \beta_k \overline{z} = 0,$$

which we can rewrite as

$$|z + \beta_k|^2 = |\beta_k|^2,$$

so that $J(X_k)$ is the Euclidean circle with Euclidean centre $-\beta_k$ and Euclidean radius $|\beta_k|$.

The slope of the tangent line to $J(X_k)$ at 0 is then

$$-\frac{\mathrm{Re}(\beta_k)}{\mathrm{Im}(\beta_k)} = -\tan(\theta_k) = \tan(-\theta_k),$$

and so $J(X_k)$ makes angle $-\theta_k$ with $\mathbb{R}$.

The angle between $J(X_1)$ and $J(X_2)$ is then given by

$$\text{angle}(J(X_1), J(X_2)) = -\theta_2 - (-\theta_1) = -\text{angle}(X_1, X_2),$$

and so J is conformal.

Exercise 2.35

Show that $C(z) = \overline{z}$ is conformal.

This completes the proof of Theorem 2.23. **QED**

Examining the proof carefully, we see that each $f(z) = az + b$ preserves the sign of the angle between X_1 and X_2 as well, whereas $C(z) = \overline{z}$ reverses the sign of the angle.

There is a subtlety with regard to $J(z) = \frac{1}{z}$. The angle between $J(X_1)$ and $J(X_2)$ at 0 is the angle between X_1 and X_2 at ∞, which is the negative of the angle between X_1 and X_2 at z_0. Hence, J also preserves the sign of the angle between X_1 and X_2.

Hence, every element of Möb^+ preserves the sign of the angle between X_1 and X_2, because Möb^+ is generated by $J(z) = \frac{1}{z}$ and the $f(z) = az + b$ for a, $b \in \mathbb{C}$ with $a \neq 0$, whereas each element of $\text{Möb} \setminus \text{Möb}^+$ reverses the sign of the angle between X_1 and X_2.

2.8 Preserving $\mathbb{H}$

Recall that our foray into Möbius transformations and the general Möbius group was undertaken as part of the attempt to determine those transformations of the upper half-plane $\mathbb{H}$ that take hyperbolic lines to hyperbolic lines.

One place to look for such transformations is the subgroup of Möb preserving $\mathbb{H}$. So, consider the group

$$\text{Möb}(\mathbb{H}) = \{m \in \text{Möb} \mid m(\mathbb{H}) = \mathbb{H}\}.$$

Theorem 2.24

Every element of $\text{Möb}(\mathbb{H})$ takes hyperbolic lines in $\mathbb{H}$ to hyperbolic lines in $\mathbb{H}$.

Proof

The proof of Theorem 2.24 is an immediate consequence of Theorem 2.23, which states that the elements of $\text{Möb}(\mathbb{H})$ preserve angles between circles in $\overline{\mathbb{C}}$, together with the facts that every hyperbolic line in $\mathbb{H}$ is the intersection of $\mathbb{H}$ with a circle in $\overline{\mathbb{C}}$ perpendicular to $\overline{\mathbb{R}}$ and that every element of Möb takes circles in $\overline{\mathbb{C}}$ to circles in $\overline{\mathbb{C}}$. **QED**

Let

$$\text{Möb}^+(\mathbb{H}) = \{m \in \text{Möb}^+ \mid m(\mathbb{H}) = \mathbb{H}\}$$

be the subgroup of $\text{Möb}(\mathbb{H})$ consisting of the Möbius transformations preserving the upper half-plane $\mathbb{H}$.

These definitions are somehow unsatisfying, as we do not have an explicit expression for an element of either $\text{Möb}(\mathbb{H})$ or $\text{Möb}^+(\mathbb{H})$. We spend the remainder of this section deriving these desired explicit expressions, as we will make use of them later.

As $\mathbb{H}$ is a disc in $\overline{\mathbb{C}}$ determined by the circle in $\overline{\mathbb{C}}$ $\overline{\mathbb{R}}$, we first determine the explicit form of an element of the subgroup

$$\text{Möb}(\overline{\mathbb{R}}) = \{m \in \text{Möb} \mid m(\overline{\mathbb{R}}) = \overline{\mathbb{R}}\}.$$

We know from Exercise 2.32 that every element of Möb can be written either as $m(z) = \frac{az+b}{cz+d}$ or as $m(z) = \frac{a\overline{z}+b}{c\overline{z}+d}$, where a, b, c, $d \in \mathbb{C}$ and $ad - bc = 1$. We are interested in determining the conditions imposed on a, b, c, and d by requiring that $m(\overline{\mathbb{R}}) = \overline{\mathbb{R}}$.

Note that in the latter case, because $C(\overline{\mathbb{R}}) = \overline{\mathbb{R}}$, we may instead consider the composition

$$m \circ C(z) = m(\overline{z}) = \frac{az+b}{cz+d},$$

and so reduce ourselves to considering just the former case in which

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

where a, b, c, $d \in \mathbb{C}$ and $ad - bc = 1$.

As m takes $\overline{\mathbb{R}}$ to $\overline{\mathbb{R}}$, we have that the three points

$$m^{-1}(\infty) = -\frac{d}{c},\ m(\infty) = \frac{a}{c},\ \text{and } m^{-1}(0) = -\frac{b}{a}$$

all lie in $\overline{\mathbb{R}}$.

Suppose for the moment that $a \neq 0$ and $c \neq 0$, so that these three points all lie in $\mathbb{R}$. We can then express each coefficient of m as a multiple of c. Specifically, we have that $a = m(\infty)c$, $b = -m^{-1}(0)a = -m^{-1}(0)m(\infty)c$, and $d = -m^{-1}(\infty)c$. In particular, we can rewrite m as

$$m(z) = \frac{az+b}{cz+d} = \frac{m(\infty)cz - m^{-1}(0)m(\infty)c}{cz - m^{-1}(\infty)c}.$$

However, normalizing so that the determinant of m is 1 imposes a condition on c, namely, that

$$\begin{aligned} 1 = ad - bc &= c^2\left[-m(\infty)m^{-1}(\infty) + m(\infty)m^{-1}(0)\right] \\ &= c^2\left[m(\infty)(m^{-1}(0) - m^{-1}(\infty))\right]. \end{aligned}$$

As $m(\infty)$, $m^{-1}(0)$, and $m^{-1}(\infty)$ are all real, this implies that c is either real or purely imaginary, and hence that the coefficients of m are either all real or all purely imaginary.

Exercise 2.36

Complete this analysis of the coefficients of m by considering the two remaining cases, namely, that $a = 0$ and that $c = 0$.

Conversely, if m has either the form $m(z) = \frac{az+b}{cz+d}$ with $ad - bc = 1$ or the form $m(z) = \frac{a\overline{z}+b}{c\overline{z}+d}$ with $ad - bc = 1$, where the coefficients of m are either all real or all purely imaginary, then the three points $m(0)$, $m(\infty)$, and $m^{-1}(\infty)$ all lie in $\overline{\mathbb{R}}$, and so m takes $\overline{\mathbb{R}}$ to $\overline{\mathbb{R}}$.

We summarize this analysis in the following theorem.

Theorem 2.25

Every element of $\text{Möb}(\overline{\mathbb{R}})$ has one of the following four forms:

1. $m(z) = \frac{az+b}{cz+d}$ with a, b, c, $d \in \mathbb{R}$ and $ad - bc = 1$;

2. $m(z) = \frac{a\overline{z}+b}{c\overline{z}+d}$ with a, b, c, $d \in \mathbb{R}$ and $ad - bc = 1$;

3. $m(z) = \frac{az+b}{cz+d}$ with a, b, c, d purely imaginary and $ad - bc = 1$;

4. $m(z) = \frac{a\overline{z}+b}{c\overline{z}+d}$ with a, b, c, d purely imaginary and $ad - bc = 1$.

Note that we now also have an explicit form for the subgroup

$$\text{Möb}(A) = \{m \in \text{Möb} \mid m(A) = A\}$$

of Möb for any circle in $\overline{\mathbb{C}}$ A. Choose some element p of Möb satisfying $p(\overline{\mathbb{R}}) = A$. If n is any element of Möb satisfying $n(A) = A$, then $p^{-1} \circ n \circ p(\overline{\mathbb{R}}) = \overline{\mathbb{R}}$. Hence, we can write $p^{-1} \circ n \circ p = m$ for some element m of Möb$(\overline{\mathbb{R}})$, and so $n = p \circ m \circ p^{-1}$, which gives that

$$\text{Möb}(A) = \{p \circ m \circ p^{-1} \mid m \in \text{Möb}(\overline{\mathbb{R}})\}.$$

This subgroup Möb(A) is independent of the choice of p. To see this, suppose that q is another element of Möb satisfying $q(\overline{\mathbb{R}}) = A$. Then, $p^{-1} \circ q$ takes $\overline{\mathbb{R}}$ to $\overline{\mathbb{R}}$, and so we can write $q = p \circ t$ for some element t of Möb$(\overline{\mathbb{R}})$. Hence, for any m in Möb$(\overline{\mathbb{R}})$, we have that $q \circ m \circ q^{-1} = p \circ (t \circ m \circ t^{-1}) \circ p^{-1}$, and so

$$\{p \circ m \circ p^{-1} \mid m \in \text{Möb}(\overline{\mathbb{R}})\} = \{q \circ m \circ q^{-1} \mid m \in \text{Möb}(\overline{\mathbb{R}})\}.$$

Exercise 2.37

Determine the general form of an element of Möb$(\mathbb{S}^1)$.

We are now ready to determine Möb$(\mathbb{H})$. Each element of Möb$(\overline{\mathbb{R}})$ either preserves each of the two discs in $\overline{\mathbb{C}}$ determined by $\overline{\mathbb{R}}$, namely, the upper and lower half-planes, or interchanges them. To determine which, we consider the image of a single point in one of the discs.

Specifically, an element m of Möb$(\overline{\mathbb{R}})$ is an element of Möb$(\mathbb{H})$ if and only if the imaginary part of $m(i)$ is positive. So, we need to check the value of Im$(m(i))$ for each of the four possible forms of an element of Möb$(\overline{\mathbb{R}})$.

If m has the form $m(z) = \frac{az+b}{cz+d}$, where a, b, c, and d are real and $ad - bc = 1$, then the imaginary part of $m(i)$ is given by

$$\begin{aligned} \text{Im}(m(i)) &= \text{Im}\left(\frac{ai+b}{ci+d}\right) \\ &= \text{Im}\left(\frac{(ai+b)(-ci+d)}{(ci+d)(-ci+d)}\right) = \frac{ad-bc}{c^2+d^2} = \frac{1}{c^2+d^2} > 0, \end{aligned}$$

and so m lies in Möb($\mathbb{H}$).

If m has the form $m(z) = \frac{a\overline{z}+b}{c\overline{z}+d}$, where a, b, c, and d are real and $ad - bc = 1$, then the imaginary part of $m(i)$ is given by

$$\begin{aligned}
\text{Im}(m(i)) &= \text{Im}\left(\frac{-ai+b}{-ci+d}\right) \\
&= \text{Im}\left(\frac{(-ai+b)(ci+d)}{(-ci+d)(ci+d)}\right) = \frac{-ad+bc}{c^2+d^2} = \frac{-1}{c^2+d^2} < 0,
\end{aligned}$$

and so m does not lie in Möb($\mathbb{H}$).

If m has the form $m(z) = \frac{az+b}{cz+d}$, where a, b, c, and d are purely imaginary and $ad - bc = 1$, write $a = \alpha i$, $b = \beta i$, $c = \gamma i$, and $d = \delta i$, so that $\alpha\delta - \beta\gamma = -1$. Then, the imaginary part of $m(i)$ is given by

$$\begin{aligned}
\text{Im}(m(i)) &= \text{Im}\left(\frac{ai+b}{ci+d}\right) = \text{Im}\left(\frac{-\alpha+\beta i}{-\gamma+\delta i}\right) \\
&= \text{Im}\left(\frac{(-\alpha+\beta i)(-\gamma-\delta i)}{(-\gamma+\delta i)(-\gamma-\delta i)}\right) = \frac{\alpha\delta-\beta\gamma}{\gamma^2+\delta^2} = \frac{-1}{\gamma^2+\delta^2} < 0,
\end{aligned}$$

and so m does not lie in Möb($\mathbb{H}$).

If m has the form $m(z) = \frac{a\overline{z}+b}{c\overline{z}+d}$, where a, b, c, and d are purely imaginary and $ad - bc = 1$, write $a = \alpha i$, $b = \beta i$, $c = \gamma i$, and $d = \delta i$, so that $\alpha\delta - \beta\gamma = -1$. Then, the imaginary part of $m(i)$ is given by

$$\begin{aligned}
\text{Im}(m(i)) &= \text{Im}\left(\frac{-ai+b}{-ci+d}\right) = \text{Im}\left(\frac{\alpha+\beta i}{\gamma+\delta i}\right) \\
&= \text{Im}\left(\frac{(\alpha+\beta i)(\gamma-\delta i)}{(\gamma+\delta i)(\gamma-\delta i)}\right) = \frac{-\alpha\delta+\beta\gamma}{\gamma^2+\delta^2} = \frac{1}{\gamma^2+\delta^2} > 0,
\end{aligned}$$

and so m lies in Möb($\mathbb{H}$).

We summarize this analysis in the following theorem.

Theorem 2.26

Every element of Möb($\mathbb{H}$) either has the form

$$m(z) = \frac{az+b}{cz+d}, \text{ where } a,\ b,\ c,\ d \in \mathbb{R} \text{ and } ad - bc = 1,$$

or has the form

$$n(z) = \frac{a\overline{z}+b}{c\overline{z}+d}, \text{ where } a,\ b,\ c,\ d \text{ are purely imaginary and } ad - bc = 1.$$

One consequence of Theorem 2.26 is that every element of $\text{Möb}^+(\mathbb{H})$ has the form

$$m(z) = \frac{az+b}{cz+d}, \text{ where } a,\ b,\ c,\ d \in \mathbb{R} \text{ and } ad - bc = 1,$$

because no element of $\text{Möb}(\mathbb{H})$ of the form

$$n(z) = \frac{a\overline{z}+b}{c\overline{z}+d}, \text{ where } a,\ b,\ c,\ d \text{ are purely imaginary and } ad - bc = 1$$

can be an element of $\text{Möb}^+(\mathbb{H})$.

Exercise 2.38

Show that $\text{Möb}(\mathbb{H})$ is generated by elements of the form $m(z) = az + b$ for $a > 0$ and $b \in \mathbb{R}$, $K(z) = \frac{-1}{z}$, and $B(z) = -\overline{z}$.

Exercise 2.39

Write the general form of an element of $\text{Möb}(\mathbb{D})$, where $\mathbb{D} = \{z \in \mathbb{C} | |z| < 1\}$ is the unit disc in $\mathbb{C}$.

Note that we have not actually addressed the question of whether $\text{Möb}(\mathbb{H})$ contains all transformations of $\mathbb{H}$ that take hyperbolic lines to hyperbolic lines. We have merely shown that every element of $\text{Möb}(\mathbb{H})$ has this property, which will suffice for the time being.

We close this section by showing that this characterization of the general form of an element of $\text{Möb}(\overline{\mathbb{R}})$ is exactly what we need to show that the definition of reflection in a circle in $\overline{\mathbb{C}}$ given in Section 2.6 is well defined.

Proposition 2.27

Reflection in a circle in $\overline{\mathbb{C}}$, as defined in Section 2.6, is well defined.

Proof

For any element m of $\text{Möb}(\overline{\mathbb{R}})$, a direct calculation based on the two possible forms for m shows that $C \circ m = m \circ C$, where $C(z) = \overline{z}$ is complex conjugation: If $m(z) = \frac{az+b}{cz+d}$, where a, b, c, $d \in \mathbb{R}$ and $ad - bc = 1$, then

$$C \circ m(z) = \frac{a\overline{z}+b}{c\overline{z}+d} = m \circ C(z);$$

if $m(z) = \frac{a\overline{z}+b}{c\overline{z}+d}$, where a, b, c, d are purely imaginary and $ad - bc = 1$, then

$$C \circ m(z) = \frac{-az-b}{-cz-d} = \frac{az+b}{cz+d} = m \circ C(z).$$

Let A be a circle in $\overline{\mathbb{C}}$, and let m and n be two elements of Möb($\overline{\mathbb{R}}$) taking $\overline{\mathbb{R}}$ to A. Then, $n^{-1} \circ m$ takes $\overline{\mathbb{R}}$ to $\overline{\mathbb{R}}$, and so $n^{-1} \circ m = p$ for some element p of Möb($\overline{\mathbb{R}}$). In particular, $p \circ C = C \circ p$. Write $m = n \circ p$, and calculate that

$$m \circ C \circ m^{-1} = n \circ p \circ C \circ p^{-1} \circ n^{-1} = n \circ p \circ p^{-1} \circ C \circ n^{-1} = n \circ C \circ n^{-1}.$$

Hence, reflection in a circle in $\overline{\mathbb{C}}$ is well defined. This completes the proof of Proposition 2.27. **QED**

2.9 Transitivity Properties of Möb($\mathbb{H}$)

In Section 2.2, we described some sets on which Möb$^+$ acts transitively, and we have also seen some ways in which knowing the transitivity of the action of Möb$^+$ on these sets is useful. In this section, we restrict our attention to the action of Möb($\mathbb{H}$) on $\mathbb{H}$, and show that we can obtain similar sorts of results.

We first observe that Möb($\mathbb{H}$) acts transitively on $\mathbb{H}$. That is, for each pair w_1 and w_2 of distinct points of $\mathbb{H}$, there exists an element m in Möb($\mathbb{H}$) taking w_1 to w_2. Even though we know that Möb acts transitively on triples of distinct points of $\overline{\mathbb{C}}$, it is not *a priori* obvious that there exists an element of Möb that both takes $\mathbb{H}$ to itself and takes w_1 to w_2.

Proposition 2.28

Möb($\mathbb{H}$) acts transitively on $\mathbb{H}$.

Proof

Using Lemma 2.8, it suffices to show that for any point w of $\mathbb{H}$, there exists an element m of Möb($\mathbb{H}$) satisfying $m(w) = i$.

Write $w = a+bi$, where $a, b \in \mathbb{R}$ and $b > 0$. We construct an element of Möb($\mathbb{H}$) taking w to i as a composition. We first move w to the positive imaginary axis using $p(z) = z - a$, so that $p(w) = p(a + bi) = bi$.

We next apply $q(z) = \frac{1}{b}z$ to $p(w)$, so that $q(p(w)) = q(bi) = i$. Note that as $-a \in \mathbb{R}$ and $\frac{1}{b} > 0$, we have by Theorem 2.26 that both $p(z)$ and $q(z)$, and hence $q \circ p(z)$, lie in Möb($\mathbb{H}$). This completes the proof of Proposition 2.28.

QED

Exercise 2.40

Show that Möb($\mathbb{H}$) acts transitively on the set $\mathcal{L}$ of hyperbolic lines in $\mathbb{H}$.

Exercise 2.41

Give an explicit expression for an element of Möb($\mathbb{H}$) taking the hyperbolic line ℓ determined by 1 and -2 to the positive imaginary axis I.

Exercise 2.42

Let X be the set of triples (ℓ, r, z), where ℓ is a hyperbolic line, z is a point on ℓ, and r is one of the two closed hyperbolic rays in ℓ determined by z. Show that Möb$^+$($\mathbb{H}$) acts transitively on X.

Even though Möb($\mathbb{H}$) acts transitively on the set $\mathcal{L}$ of hyperbolic lines in $\mathbb{H}$ and even though a hyperbolic line is determined by a pair of distinct points in $\mathbb{H}$, it does not follow that Möb($\mathbb{H}$) acts transitively on the set $\mathcal{P}$ of pairs of distinct points of $\mathbb{H}$, much less on the set $\mathcal{T}_{\mathbb{H}}$ of triples of distinct points of $\mathbb{H}$.

We can see this result directly by considering the positive imaginary axis I. As the endpoints at infinity of I are 0 and ∞, every element of Möb($\mathbb{H}$) taking I to itself either fixes both 0 and ∞, or else interchanges them. Recall from Theorem 2.26 that we know the general form of an element of Möb($\mathbb{H}$). Using this, we see that an element m of Möb($\mathbb{H}$) fixing both 0 and ∞ either has the form $m(z) = az$, where $a \in \mathbb{R}$ and $a > 0$, or has the form $m(z) = -a\overline{z}$, where again $a \in \mathbb{R}$ and $a > 0$.

An element m of Möb($\mathbb{H}$) interchanging 0 and ∞ either has the form $m(z) = -\frac{b}{z}$, where $b \in \mathbb{R}$ and $b > 0$, or has the form $m(z) = \frac{b}{\overline{z}}$, where again $b \in \mathbb{R}$ and $b > 0$.

In any case, we can see that no element of $\text{Möb}(\mathbb{H})$ takes the positive imaginary axis I to itself, takes i to i, and takes $2i$ to $3i$. In fact, the only elements of $\text{Möb}(\mathbb{H})$, other than the identity, that take I to itself and fix i are $B(z) = -\overline{z}$, which is reflection in I and so it necessarily fixes every point of I, and $K(z) = -\frac{1}{z}$, which interchanges the two hyperbolic rays in I radiating from i.

We will return to this failure of $\text{Möb}(\mathbb{H})$ to act transitively on the set $\mathcal{P}$ of pairs of distinct points of $\mathbb{H}$ after we have developed a means of measuring hyperbolic distance in $\mathbb{H}$.

We also need to make use of the analog in $\mathbb{H}$ of a disc in $\overline{\mathbb{C}}$.

Definition 2.29

An *open half-plane in* $\mathbb{H}$ is a component of the complement of a hyperbolic line in $\mathbb{H}$.

A *closed half-plane* is the union of a hyperbolic line ℓ with one of the open half-planes determined by ℓ. A *half-plane* is either an open half-plane or a closed half-plane. In particular, each half-plane, either open or closed, is determined by a unique hyperbolic line, and each hyperbolic line determines a pair of half-planes, either open or closed. The hyperbolic line determining a half-plane is the *bounding line* for that half-plane.

In much the same way that we extended the transitivity of Möb on the set $\mathcal{C}$ of circles in $\overline{\mathbb{C}}$ to transitivity on the set $\mathcal{D}$ of discs in $\overline{\mathbb{C}}$, we can extend the transitivity of $\text{Möb}(\mathbb{H})$ on the set $\mathcal{L}$ of hyperbolic lines in $\mathbb{H}$ to transitivity on the set of half-planes in $\mathbb{H}$.

Exercise 2.43

Show that $\text{Möb}(\mathbb{H})$ acts transitively on the set $\mathcal{H}$ of open half-planes in $\mathbb{H}$.

We can also consider the action of $\text{Möb}(\mathbb{H})$ on the boundary at infinity $\overline{\mathbb{R}}$ of $\mathbb{H}$.

Proposition 2.30

$\text{Möb}(\mathbb{H})$ acts triply transitively on the set $\mathcal{T}_{\overline{\mathbb{R}}}$ of triples of distinct points of $\overline{\mathbb{R}}$.

Proof

Again using Lemma 2.8, given a triple (z_1, z_2, z_3) of distinct points of $\overline{\mathbb{R}}$, it suffices to show that there exists an element of Möb($\mathbb{H}$) taking (z_1, z_2, z_3) to $(0, 1, \infty)$.

Let ℓ be the hyperbolic line whose endpoints at infinity are z_1 and z_3, and let m be an element of Möb($\mathbb{H}$) taking ℓ to the positive imaginary axis I. By composing m with $K(z) = -\frac{1}{z}$ if necessary, we can assume that $m(z_1) = 0$ and $m(z_3) = \infty$. Set $b = m(z_2)$.

If $b > 0$, then the composition of m with $p(z) = \frac{1}{b}z$ takes (z_1, z_2, z_3) to $(0, 1, \infty)$.

If $b < 0$, then $p(z) = \frac{1}{b}z$ no longer lies in Möb($\mathbb{H}$), but the composition of m with $q(z) = \frac{1}{b}\overline{z}$, which does lie in Möb($\mathbb{H}$), takes (z_1, z_2, z_3) to $(0, 1, \infty)$. This completes the proof of Proposition 2.30. **QED**

We close this section by noting that Möb$^+$($\mathbb{H}$) does not act triply transitively on $\mathcal{T}_{\overline{\mathbb{R}}}$, because no element of Möb$^+$($\mathbb{H}$) takes $(0, 1, \infty)$ to $(0, -1, \infty)$. To see this, note that if the element $m(z) = \frac{az+b}{cz+d}$ of Möb$^+$($\mathbb{H}$) fixes 0 and ∞, then $b = c = 0$; as $ad = 1$, we have that $m(z) = a^2 z$. In particular, $m(1) = a^2 > 0$, and so $m(1)$ cannot equal -1.

2.10 The Geometry of the Action of Möb($\mathbb{H}$)

The purpose of this section is consider how individual elements of Möb($\mathbb{H}$) act on $\mathbb{H}$. This section is perhaps best viewed as a catalogue of possibilities.

In Section 2.8, we saw that every nontrivial element of Möb($\mathbb{H}$) can be written either as

$$m(z) = \frac{az+b}{cz+d}, \text{ where } a, b, c, d \text{ are real with } ad - bc = 1,$$

or as

$$n(z) = \frac{\alpha\overline{z}+\beta}{\gamma\overline{z}+\delta}, \text{ where } \alpha, \beta, \gamma, \delta \text{ are purely imaginary with } \alpha\delta - \beta\gamma = 1.$$

Using these explicit formulae, we can determine the sets of fixed points.

We first consider the case in which $m(z) = \frac{az+b}{cz+d}$, where a, b, c, and d are real with $ad - bc = 1$. What follows is similar in spirit to the discussion in Section

2.4. In Section 2.1, we saw that the fixed points of m are the solutions to the equation

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

which are the roots in $\overline{\mathbb{C}}$ of the polynomial $p(z) = cz^2 + (d-a)z - b = 0$.

In the case in which $c = 0$, there is one fixed point at ∞. There is a second fixed point, namely, $\frac{b}{d-a}$, if and only if $a \neq d$, and such a fixed point is necessarily a real number. So, if $c = 0$, either there is a single fixed point at ∞ or there are two fixed points, one at ∞ and the other in $\mathbb{R}$.

In the case in which $c \neq 0$, there are two roots of $p(z)$ in $\mathbb{C}$, namely,

$$\frac{1}{2}[a - d \pm \sqrt{(d-a)^2 - 4bc}].$$

As the coefficients of $p(z)$ are real, the roots of $p(z)$ are invariant under complex conjugation, and so either both roots are real or one lies in $\mathbb{H}$ and the other in the lower half-plane.

Note that $p(z)$ has exactly one root, which is then necessarily real, if and only if

$$(a-d)^2 - 4bc = (a+d)^2 - 4 = 0;$$

has two real roots if and only if

$$(a-d)^2 - 4bc = (a+d)^2 - 4 > 0;$$

and has two complex roots, symmetric under complex conjugation, if and only if

$$(a-d)^2 - 4bc = (a+d)^2 - 4 < 0.$$

Combining this analysis with the classification of elements of Möb^+ as described in Section 2.4, we see that m has one fixed point inside $\mathbb{H}$ if and only if m is elliptic; that m has one fixed point on $\overline{\mathbb{R}}$ if and only if m is parabolic; that m has two fixed points on $\overline{\mathbb{R}}$ if and only if m is loxodromic; and that these are the only possibilities.

In the case in which m is elliptic and so has one fixed point inside $\mathbb{H}$, the action of m on $\mathbb{H}$ is rotation about the fixed point. In fact, if we take the fixed point of m in $\mathbb{H}$ to be i, so that the other fixed point of m is at $-i$, we may use Exercise 2.28 to see that m has the form

$$m(z) = \frac{\cos(\theta)z + \sin(\theta)}{-\sin(\theta)z + \cos(\theta)}$$

for some real number θ. As $\text{Möb}(\mathbb{H})$ acts transitively on $\mathbb{H}$, every elliptic element is conjugate to a Möbius transformation of this form: If $m \in \text{Möb}^+(\mathbb{H})$ is elliptic fixing $x_0 \in \mathbb{H}$, then let p be an element of $\text{Möb}(\mathbb{H})$ taking x_0 to i, so that $p \circ m \circ p^{-1}$ is elliptic fixing i.

Note, however, that m is not the standard Euclidean rotation about i. For instance, take $\theta = \frac{1}{2}\pi$ and note that $m(1+i) = -\frac{1}{2} + \frac{1}{2}i$. Indeed, the hyperbolic line passing through i and $1+i$ is not the horizontal Euclidean line $L = \{z \in \mathbb{H} \mid \text{Im}(z) = 1\}$ through i, which is not a hyperbolic line at all, but it is instead the hyperbolic line contained in the Euclidean circle with Euclidean centre $\frac{1}{2}$ and Euclidean radius $\frac{\sqrt{5}}{2}$, which passes through $-\frac{1}{2} + \frac{1}{2}i$.

In the case in which m is parabolic and so has one fixed point x on $\overline{\mathbb{R}}$, we may use the triple transitivity of the action of $\text{Möb}(\mathbb{H})$ on $\overline{\mathbb{R}}$ to conjugate m by an element of $\text{Möb}(\mathbb{H})$ to have the form $m(z) = z + 1$. However, note that a parabolic element of $\text{Möb}^+(\mathbb{H})$ is not necessarily conjugate by an element of $\text{Möb}^+(\mathbb{H})$ to $m(z) = z + 1$, because of the failure of $\text{Möb}^+(\mathbb{H})$ to act triply transitively on $\overline{\mathbb{R}}$.

In particular, a parabolic transformation m in $\text{Möb}(\mathbb{H})$ with fixed point x preserves every circle in $\overline{\mathbb{C}}$ that is contained in $\mathbb{H} \cup \overline{\mathbb{R}}$ and that is tangent to $\overline{\mathbb{R}}$ at x. These circles are the *horocircles* taken to themselves by m. The components of the complement of a horocircle in $\mathbb{H}$ are the two *horodiscs* in $\mathbb{H}$ determined by the horocircle. This is most easily seen in the case for the fixed point $x = \infty$; in which case, these circles in $\overline{\mathbb{C}}$ are precisely the circles in $\overline{\mathbb{C}}$ that are the union of a horizontal Euclidean line in $\mathbb{H}$ with $\{\infty\}$.

In the case in which m is loxodromic and so has two fixed points x and y in $\overline{\mathbb{R}}$, we may use the transitivity of $\text{Möb}(\mathbb{H})$ on pairs of distinct points of $\overline{\mathbb{R}}$ to conjugate m to have the form $m(z) = \lambda z$ for some positive real number λ. In this case, the positive imaginary axis is taken to itself by m, as are both of the half-planes determined by the positive imaginary axis.

In general, we define the *axis* of a loxodromic m, denoted $\text{axis}(m)$, to be the hyperbolic line in $\mathbb{H}$ determined by the fixed points of m. Exactly as in the previous paragraph, we have that m takes its axis to itself, and takes each of the half-planes determined by $\text{axis}(m)$ to itself.

We summarize this analysis in the following theorem.

Theorem 2.31

Let $m(z) = \frac{az+b}{cz+d}$ be an element of $\text{Möb}^+(\mathbb{H})$, so that $a, b, c, d \in \mathbb{R}$ and $ad - bc = 1$. Then, exactly one of the following holds:

1. m is the identity;

2. m has exactly two fixed points in $\overline{\mathbb{R}}$; in which case, m is loxodromic and is conjugate in $\text{Möb}^+(\mathbb{H})$ to $q(z) = \lambda z$ for some positive real number λ;

3. m has one fixed point in $\overline{\mathbb{R}}$; in which case, m is parabolic and is conjugate in $\text{Möb}(\mathbb{H})$ to $q(z) = z + 1$; or

4. m has one fixed point in $\mathbb{H}$; in which case, m is elliptic and is conjugate in $\text{Möb}^+(\mathbb{H})$ to $q(z) = \frac{\cos(\theta)z+\sin(\theta)}{-\sin(\theta)z+\cos(\theta)}$ for some real number θ.

Let m be a loxodromic transformation in $\text{Möb}^+(\mathbb{H})$, let x and y be the fixed points of m in $\overline{\mathbb{R}}$, and let A be any circle in $\overline{\mathbb{C}}$ that passes through x and y, not necessarily perpendicular to $\overline{\mathbb{R}}$. As m takes circles in $\overline{\mathbb{C}}$ to circles in $\overline{\mathbb{C}}$, preserves angles, and preserves the half-planes determined by its axis, we see that m takes $A \cap \mathbb{H}$ to itself. Moreover, m acts as translation along $A \cap \mathbb{H}$.

This completes our brief tour of the action of the elements of $\text{Möb}^+(\mathbb{H})$ on $\mathbb{H}$. There are also the elements of $\text{Möb}(\mathbb{H}) \setminus \text{Möb}^+(\mathbb{H})$ to consider, where

$$\text{Möb}(\mathbb{H}) \setminus \text{Möb}^+(\mathbb{H}) = \{m \in \text{Möb}(\mathbb{H}) \mid m \notin \text{Möb}^+(\mathbb{H})\}.$$

As we saw in Section 2.8, every element n of $\text{Möb}(\mathbb{H}) \setminus \text{Möb}^+(\mathbb{H})$ has the form

$$n(z) = \frac{\alpha\overline{z} + \beta}{\gamma\overline{z} + \delta},$$

where α, β, γ, and δ are purely imaginary with $\alpha\delta - \beta\gamma = 1$.

As above, we begin our description of the action of n on $\mathbb{H}$ by determining the fixed points of n, which are the points z of $\mathbb{H}$ satisfying

$$\frac{\alpha\overline{z} + \beta}{\gamma\overline{z} + \delta} = z.$$

We begin our analysis by considering a particular example, namely, the transformation

$$q(z) = \frac{i\overline{z} + 2i}{i\overline{z} + i}.$$

The fixed points in $\overline{\mathbb{C}}$ of q are the solutions in $\overline{\mathbb{C}}$ of the equation $q(z) = z$. As $q(\infty) = 1 \neq \infty$, these are those points z in $\mathbb{C}$ satisfying

$$i\overline{z} + 2i = z(i\overline{z} + i),$$

which we may rewrite as

$$-2\,\mathrm{Im}(z) + i[|z|^2 - 2] = 0.$$

Taking real and imaginary parts, we see that $\mathrm{Im}(z) = 0$ for every fixed point z of q, and so there are no fixed points of q in $\mathbb{H}$. As $|z|^2 = 2$ as well, there are two fixed points of q in $\overline{\mathbb{R}}$, namely, at $\pm\sqrt{2}$.

In this case, we see that q takes the hyperbolic line ℓ determined by $\pm\sqrt{2}$ to itself, but it does not fix any point on ℓ. Instead, q acts as reflection in ℓ followed by translation along ℓ. In particular, the action of q interchanges the two half-planes in $\mathbb{H}$ determined by ℓ. We refer to q as a *glide reflection* along ℓ.

Exercise 2.44

Express q as the composition of reflection in ℓ and a loxodromic with axis ℓ.

To attack the general case, write $\alpha = ai$, $\beta = bi$, $\gamma = ci$, and $\delta = di$, where a, b, c, and d are real with $ad - bc = -1$. Also, write $x = \mathrm{Re}(z)$ and $y = \mathrm{Im}(z)$. The fixed points of n then satisfy the equation

$$c|z|^2 + dz - a\overline{z} - b = cx^2 + cy^2 + (d-a)x - b + i(d+a)y = 0.$$

Assume that n has a fixed point $z = x + iy$ in $\mathbb{H}$. Taking the imaginary parts of the terms in this equation and noting that $y > 0$, we see that $a + d = 0$, and so $d = -a$. In particular, we see that $ad - bc = -d^2 - bc = -1$. The fixed points of n therefore satisfy the equation

$$cx^2 + cy^2 + 2dx - b = 0.$$

In the case in which $c = 0$, we have no restriction on the imaginary part of the fixed point z. Also, we have that $d \neq 0$ because $ad - bc = -1$, and so the fixed points of n are exactly the points in $\mathbb{H}$ that lie on the Euclidean line $\{z \in \mathbb{H} \mid \mathrm{Re}(z) = \frac{b}{2d}\}$, which is the hyperbolic line determined by ∞ and $\frac{b}{2d}$.

That is, every point on the the hyperbolic line ℓ determined by ∞ and $\frac{b}{2d}$ is fixed by n. Let r_ℓ be reflection in ℓ, and consider the composition $r_\ell \circ n$. As $n \in \mathrm{Möb}(\mathbb{H}) \setminus \mathrm{Möb}^+(\mathbb{H})$, we have that $r_\ell \circ n \in \mathrm{Möb}^+(\mathbb{H})$. As $r_\ell \circ n$ fixes more than two points of $\mathbb{H}$, namely, every point of ℓ, we see that $r_\ell \circ n$ is the identity, and so $n = r_\ell$ is reflection in ℓ.

Exercise 2.45

Determine the fixed points of $q(z) = -\overline{z} + 1$.

In the case in which $c \neq 0$, divide through by c and complete the square to see that the fixed points of n in $\mathbb{H}$ are given by the equation

$$x^2 + y^2 + \frac{2d}{c}x - \frac{b}{c} = \left(x + \frac{d}{c}\right)^2 + y^2 - \frac{d^2 + bc}{c^2} = \left(x + \frac{d}{c}\right)^2 + y^2 - \frac{1}{c^2} = 0,$$

which is the Euclidean circle A in $\mathbb{C}$ with Euclidean centre $-\frac{d}{c}$ and Euclidean radius $\frac{1}{|c|}$.

In particular, this equation gives that the fixed points of n are exactly the points on the hyperbolic line $A \cap \mathbb{H}$. As in the case in which $c = 0$, in this case, n is equal to reflection in $A \cap \mathbb{H}$.

Exercise 2.46

Determine the fixed points of

$$q(z) = \frac{2i\overline{z} - i}{3i\overline{z} - 2i}.$$

We need to exercise a bit of caution, however, as there are elements of Möb($\mathbb{H}$), such as the transformation $q(z) = \frac{i\overline{z}+2i}{i\overline{z}+i}$ considered earlier in the section, that do not act as reflection in a hyperbolic line.

The difficulty lies in that we began the analysis of the elements of Möb($\mathbb{H}$) $\setminus$ Möb$^+$($\mathbb{H}$) by assuming that the element in question had a fixed point in $\mathbb{H}$.

So, to complete our analysis of the elements of Möb($\mathbb{H}$) $\setminus$ Möb$^+$($\mathbb{H}$), we consider the case in which no fixed points of n are in $\mathbb{H}$.

In this case, the solutions of $n(z) = z$ are the points z in $\overline{\mathbb{C}}$ that satisfy the equation

$$cx^2 + cy^2 + (d - a)x - b + i(d + a)y = 0.$$

As we are interested in the case in which no solutions are in $\mathbb{H}$, we set $y = 0$ and consider those solutions that lie in $\overline{\mathbb{R}}$.

In the case in which $c = 0$, we have two solutions, namely, ∞ and $\frac{b}{2d}$. In this case, n takes the hyperbolic line ℓ determined by ∞ and $\frac{b}{2d}$ to itself and interchanges the two half-planes determined by ℓ, but no point on ℓ is fixed by n, because n has no fixed points in $\mathbb{H}$ by assumption. That is, n acts as a glide reflection along ℓ.

In this case, we can express n as the composition of reflection in ℓ and a loxodromic with axis ℓ. The easiest way to see this is to note that the composition

$$n \circ B(z) = n(-\overline{z}) = \frac{-\alpha z + \beta}{-\gamma z + \delta} = \frac{-az + b}{-cz + d}$$

is loxodromic, where $\alpha = ai$, $\beta = bi$, $\gamma = ci$, and $\delta = di$ are purely imaginary with $\alpha\delta - \beta\gamma = 1$.

In the case in which $c \neq 0$, the fixed points of n can be found by applying the quadratic equation to $cx^2 + (d-a)x - b = 0$, to get

$$x = \frac{1}{2c}\left[a - d \pm \sqrt{(d-a)^2 + 4bc}\right] = \frac{1}{2c}\left[a - d \pm \sqrt{(a+d)^2 + 4}\right],$$

using that $ad - bc = -1$.

In particular, if n has no fixed points in $\mathbb{H}$, then n necessarily has exactly two fixed points on $\overline{\mathbb{R}}$, and so n acts as a glide reflection along the hyperbolic line determined by these two points. Exactly as in the cases above, such an n is the composition of reflection in this hyperbolic line and a loxodromic with this hyperbolic line as its axis.

We summarize the analysis of elements of $\text{Möb}(\mathbb{H}) \setminus \text{Möb}^+(\mathbb{H})$ in the following theorem.

Theorem 2.32

Let $n(z) = \frac{\alpha\overline{z}+\beta}{\gamma\overline{z}+\delta}$ be an element of $\text{Möb}(\mathbb{H}) \setminus \text{Möb}^+(\mathbb{H})$, so that α, β, γ, and δ are purely imaginary with $\alpha\delta - \beta\gamma = 1$. Then, exactly one of the following holds:

1. n fixes a point of $\mathbb{H}$; in which case, there is a hyperbolic line ℓ in $\mathbb{H}$ so that n acts as reflection in ℓ; or

2. n fixes no point of $\mathbb{H}$; in which case, n fixes exactly two points of $\overline{\mathbb{R}}$ and acts as a glide reflection along the hyperbolic line ℓ determined by these two points.

Exercise 2.47

Let $p(z) = z + 1$ be parabolic, and let n be reflection in a hyperbolic line ℓ. Compute the composition $p \circ n$, and determine the fixed points of $p \circ n$.

3
Length and Distance in $\mathbb{H}$

We now have a reasonable group of transformations of $\mathbb{H}$, namely, Möb($\mathbb{H}$). This group is reasonable in the sense that its elements take hyperbolic lines to hyperbolic lines and preserve angles. In this chapter, we derive a means of measuring lengths of paths in $\mathbb{H}$ that is invariant under the action of this group, expressed as an *invariant element of arc-length.* From this invariant element of arc-length, we construct an *invariant notion of distance* on $\mathbb{H}$ and explore some of its basic properties.

3.1 Paths and Elements of Arc-length

Now that we have a group of transformations of $\mathbb{H}$ taking hyperbolic lines to hyperbolic lines, namely, Möb($\mathbb{H}$), we are in a position to attempt to derive the element of arc-length for the hyperbolic metric on $\mathbb{H}$. However, we first need to recall from calculus the definition of an element of arc-length.

A C^1 *path* in the plane $\mathbb{R}^2$ is a function $f : [a, b] \to \mathbb{R}^2$ that is continuous on $[a, b]$ and differentiable on (a, b) with continuous derivative. In coordinates, we can write $f(t) = (x(t), y(t))$, where $x(t)$ and $y(t)$ are continuous on $[a, b]$ and differentiable on (a, b) with continuous derivative. The image of an interval (either open, half-open, or closed) under a path f is a *curve* in $\mathbb{R}^2$.

The *Euclidean length* of f is given by the integral

$$\text{length}(f) = \int_a^b \sqrt{(x'(t))^2 + (y'(t))^2}\, \mathrm{d}t,$$

where $\sqrt{(x'(t))^2 + (y'(t))^2}\, \mathrm{d}t$ is the *element of arc-length* in $\mathbb{R}^2$.

Note that the length of a graph of a C^1 path $g : [a, b] \to \mathbb{R}$ is a special case of the length of a C^1 path as described above. In this case, given $g : [a, b] \to \mathbb{R}$, we construct a path $f : [a, b] \to \mathbb{R}^2$ by setting $f(t) = (t, g(t))$.

As an example, consider the C^1 path $f : [0, 2] \to \mathbb{R}^2$ given by $f(t) = (1+t, \frac{1}{2}t^2)$. The length of f is

$$\begin{aligned} \text{length}(f) = \int_0^2 \sqrt{1+t^2}\, \mathrm{d}t &= \frac{1}{2}[t\sqrt{1+t^2} + \ln|t + \sqrt{1+t^2}|]\,\Big|_0^2 \\ &= \sqrt{5} + \frac{1}{2}\ln(2+\sqrt{5}). \end{aligned}$$

We now engage in a bit of notational massage. If we view f as a path into $\mathbb{C}$ instead of $\mathbb{R}^2$ and write $f(t) = x(t) + iy(t)$, we then have that $f'(t) = x'(t) + iy'(t)$ and $|f'(t)| = \sqrt{(x'(t))^2 + (y'(t))^2}$. In particular, the integral for the length of f becomes

$$\text{length}(f) = \int_a^b \sqrt{(x'(t))^2 + (y'(t))^2}\, \mathrm{d}t = \int_a^b |f'(t)|\, \mathrm{d}t.$$

At this point, we introduce a new piece of notation and abbreviate this integral as

$$\int_a^b |f'(t)|\, \mathrm{d}t = \int_f |\mathrm{d}z|,$$

where we write the standard Euclidean element of arc-length in $\mathbb{C}$ as

$$|\mathrm{d}z| = |f'(t)|\, \mathrm{d}t.$$

One advantage of this notation is that it is extremely flexible and easily extendable. For instance, we may easily write any path integral in this notation. That is, let ρ be a continuous function $\rho : \mathbb{C} \to \mathbb{R}$. The *path integral* of ρ along a C^1 path $f : [a, b] \to \mathbb{C}$ is the integral

$$\int_f \rho(z)\, |\mathrm{d}z| = \int_a^b \rho(f(t))\, |f'(t)|\, \mathrm{d}t.$$

We can interpret this path integral as giving rise to a new element of arc-length, denoted $\rho(z)\, |\mathrm{d}z|$, given by scaling the Euclidean element of arc-length $|\mathrm{d}z|$ at every point $z \in \mathbb{C}$, where the amount of scaling is described by the function ρ. This gives rise to the following definition.

Definition 3.1

For a C^1 path $f : [a, b] \to \mathbb{C}$, we define the *length of f with respect to the element of arc-length $\rho(z)|\mathrm{d}z|$* to be the integral

$$\operatorname{length}_\rho(f) = \int_f \rho(z)\, |\mathrm{d}z| = \int_a^b \rho(f(t))\, |f'(t)|\, \mathrm{d}t.$$

Innumerable variations on this theme exist, and we spend the remainder of this section exploring some of them. In the next section, we restrict consideration to such elements of arc-length on $\mathbb{H}$.

As a specific example, set $\rho(z) = \frac{1}{1+|z|^2}$ and consider the element of arc-length

$$\rho(z)\, |\mathrm{d}z| = \frac{1}{1+|z|^2} |\mathrm{d}z|$$

on $\mathbb{C}$.

For $r > 0$, consider the C^1 path $f : [0, 2\pi] \to \mathbb{C}$ given by $f(t) = re^{it}$, which parametrizes the Euclidean circle with Euclidean centre 0 and Euclidean radius r. As $|f(t)| = r$ and $|f'(t)| = |ire^{it}| = r$, the length of f with respect to the element of arc-length $\frac{1}{1+|z|^2}\,|\mathrm{d}z|$ is

$$\operatorname{length}_\rho(f) = \int_f \frac{1}{1+|z|^2} |\mathrm{d}z| = \int_0^{2\pi} \frac{1}{1+|f(t)|^2} |f'(t)| \mathrm{d}t = \frac{2\pi r}{1+r^2}.$$

Exercise 3.1

Consider the function δ on $\mathbb{D} = \{z \in \mathbb{C} \mid |z| < 1\}$, defined to be the reciprocal of the Euclidean distance from z to $\mathbb{S}^1 = \partial\mathbb{D}$. Give an explicit formula for $\delta(z)$ in terms of z. For each $0 < r < 1$, let C_r be the Euclidean circle in $\mathbb{D}$ with Euclidean centre 0 and Euclidean radius r, and calculate the length of C_r with respect to the element of arc-length $\delta(z)|\mathrm{d}z|$.

We refer to an element of arc-length of the form $\rho(z)\, |\mathrm{d}z|$ as a *conformal distortion* of the standard element of arc-length $|\mathrm{d}z|$ on $\mathbb{C}$.

Conformal distortions of $|\mathrm{d}z|$ are not the most general form that an element of arc-length on an open subset of $\mathbb{C}$ can take, but they are the most general we will consider in this book. We will not work with more general elements of arc-length, largely because we do not have to: As we will see, to derive an element of arc-length on $\mathbb{H}$ invariant under the action of Möb($\mathbb{H}$), it suffices to work with conformal distortions of $|\mathrm{d}z|$.

Up to this point, we have been considering only C^1 paths. It is both easy and convenient to enlarge the set of paths considered. A path $f : [a, b] \to \mathbb{C}$ is *piecewise* C^1 if f is continuous and if there is a partition of the interval $[a, b]$ into subintervals $[a = a_0, a_1]$, $[a_1, a_2], \ldots, [a_n, a_{n+1} = b]$ so that f is a C^1 path when restricted to each subinterval $[a_k, a_{k+1}]$.

A natural example of a piecewise C^1 path that is not a C^1 path comes from considering absolute value. Specifically, consider $f : [-1, 1] \to \mathbb{C}$ defined by $f(t) = t + i|t|$. As $|t|$ is not differentiable at $t = 0$, this is not a C^1 path. However, on $[-1, 0]$, we have that $|t| = -t$ and hence that $f(t) = t - it$, which is a C^1 path. Similarly, on $[0, 1]$, we have that $|t| = t$ and hence that $f(t) = t+it$, which again is a C^1 path. So, f is a piecewise C^1 path on $[-1, 1]$.

Any calculation or operation that we can perform on a C^1 path, we can also perform on a piecewise C^1 path, by expressing it as the concatenation of the appropriate number of C^1 paths. Unless otherwise stated, we assume that all paths are piecewise C^1.

Exercise 3.2

Calculate the length of the piecewise C^1 path $f : [-1, 1] \to \mathbb{C}$ given by $f(t) = t + i|t|$ with respect to the element of arc-length $\frac{1}{1+|z|^2}|\mathrm{d}z|$.

One question to consider is what happens to the length of a piecewise C^1 path $f : [a, b] \to \mathbb{C}$ with respect to the element of arc-length $\rho(z)|\mathrm{d}z|$ when the domain of f is changed. That is, suppose that $h : [\alpha, \beta] \to [a, b]$ is a surjective piecewise C^1 function (so that $[a, b] = h([\alpha, \beta])$), and construct a new piecewise C^1 path by taking the composition $g = f \circ h$. How are $\mathrm{length}_\rho(f)$ and $\mathrm{length}_\rho(g)$ related?

The length of f with respect to $\rho(z)|\mathrm{d}z|$ is the integral

$$\mathrm{length}_\rho(f) = \int_a^b \rho(f(t))\, |f'(t)|\, \mathrm{d}t,$$

whereas the length of g with respect to $\rho(z)|\mathrm{d}z|$ is the integral

$$\begin{aligned}
\mathrm{length}_\rho(g) &= \int_\alpha^\beta \rho(g(t))\,|g'(t)|\,\mathrm{d}t \\
&= \int_\alpha^\beta \rho((f\circ h)(t))\,|(f\circ h)'(t)|\,\mathrm{d}t \\
&= \int_\alpha^\beta \rho(f(h(t)))\,|f'(h(t))|\,|h'(t)|\,\mathrm{d}t.
\end{aligned}$$

If $h'(t)\geq 0$ for all t in $[\alpha,\beta]$, then $h(\alpha)=a$ and $h(\beta)=b$, and $|h'(t)|=h'(t)$, and so after making the substitution $s=h(t)$, the length of g with respect to $\rho(z)|\mathrm{d}z|$ becomes

$$\begin{aligned}
\mathrm{length}_\rho(g) &= \int_\alpha^\beta \rho(f(h(t)))\,|f'(h(t))|\,|h'(t)|\,\mathrm{d}t \\
&= \int_a^b \rho(f(s))\,|f'(s)|\,\mathrm{d}s = \mathrm{length}_\rho(f).
\end{aligned}$$

Similarly, if $h'(t)\leq 0$ for all t in $[\alpha,\beta]$, then $h(\alpha)=b$ and $h(\beta)=a$, and $|h'(t)|=-h'(t)$, and so after making the substitution $s=h(t)$, the length of g with respect to $\rho(z)|\mathrm{d}z|$ becomes

$$\begin{aligned}
\mathrm{length}_\rho(g) &= \int_\alpha^\beta \rho(f(h(t)))\,|f'(h(t))|\,|h'(t)|\,\mathrm{d}t \\
&= -\int_b^a \rho(f(s))\,|f'(s)|\,\mathrm{d}s = \mathrm{length}_\rho(f).
\end{aligned}$$

So, we have shown that if $h'(t)$ does not change sign (so that either $h'(t)\geq 0$ for all t in $[\alpha,\beta]$, or $h'(t)\leq 0$ for all t in $[\alpha,\beta]$), then

$$\mathrm{length}_\rho(f)=\mathrm{length}_\rho(f\circ h),$$

where $f:[a,b]\to\mathbb{C}$ is a piecewise C^1 path and $h:[\alpha,\beta]\to[a,b]$ is piecewise C^1. In this case, we refer to $f\circ h$ as a *reparametrization* of f. Note that reparametrization allows us to choose the domain of definition for a piecewise C^1 path at will, because we can always find such an h between two intervals.

Although we do not prove it here, the converse of this argument holds as well, namely, that $\mathrm{length}_\rho(f)=\mathrm{length}_\rho(f\circ h)$ implies that $h'(t)\geq 0$ for all t or $h'(t)\leq 0$ for all t. This discussion of reparametrization is summarized in the following proposition.

Proposition 3.2

Let $f:[a,b]\to\mathbb{C}$ be a piecewise C^1 path, let $[\alpha,\beta]$ be an interval in $\mathbb{R}$, and let $h:[\alpha,\beta]\to[a,b]$ be a surjective piecewise C^1 function. Let $\rho(z)|\mathrm{d}z|$ be an element of arc-length on $\mathbb{C}$. Then

$$\text{length}_\rho(f \circ h) \geq \text{length}_\rho(f)$$

with equality if and only if $h \circ f$ is a reparametrization of f, that is, with equality if and only if $h'(t)$ does not change sign (so that either $h'(t) \geq 0$ for all t in $[\alpha, \beta]$, or $h'(t) \leq 0$ for all t in $[\alpha, \beta]$).

We close this section with some terminology.

Definition 3.3

A *parametrization* of a subset X of $\mathbb{C}$ is a piecewise C^1 path $f : [a, b] \to \mathbb{C}$ such that $X = f([a, b])$.

As an example, the piecewise C^1 path $g : [0, 4\pi] \to \mathbb{C}$ given by $g(t) = \cos(t) + i\sin(t)$ is a parametrization of the unit circle $\mathbb{S}^1$ in $\mathbb{C}$.

We can think of a parametrization $f : [a, b] \to X$ of a set X as describing a way of walking along X: as t walks along the interval $[a, b]$, the point $f(t)$ walks along X.

Definition 3.4

A piecewise C^1 path $f : [a, b] \to \mathbb{C}$ is a *simple path* if f is injective.

An example of a simple path is the piecewise C^1 path $f : [-1, 1] \to \mathbb{C}$ given by $f(t) = t + |t|i$ considered earlier. An example of a piecewise C^1 path that is not simple is the path $g : [0, 4\pi] \to \mathbb{C}$ given by $g(t) = \cos(t) + i\sin(t)$, because $g(t) = g(2\pi + t)$ for all $0 \leq t \leq 2\pi$.

Definition 3.5

Let f be a parametrization of the set X in $\mathbb{C}$. If f is a simple path, we say that f is a *simple parametrization* of X.

For much of what we do, this definition of simple parametrization is too restrictive, in the same way that working only with C^1 paths, rather than piecewise C^1 paths, is too restrictive. This leads us to the following definitions.

Definition 3.6

A piecewise C^1 path $f : [c, d] \to \mathbb{C}$ is an *almost simple* path if f can be expressed as a composition $f = h \circ g$, where $h : [a, b] \to \mathbb{C}$ is a simple path and $g : [c, d] \to [a, b]$ is a piecewise C^1 function with the property that $g'(t)$ does not change sign (so that either $g'(t) \geq 0$ for all t in $[c, d]$ or $g'(t) \leq 0$ for all t in $[c, d]$).

Definition 3.7

Let f be a parametrization of the set X in $\mathbb{C}$. If f is an almost simple path, we say that f is an *almost simple parametrization* of X.

The advantage of these definitions of almost simple path and almost simple parametrization is that, if we think of a parametrization of a set X as a way of walking along X, then an almost simple parametrization of X is a way of walking along X with pauses, as long as we do not backtrack. As we have already seen, in Proposition 3.2, this has the implication that if f is an almost simple path that is the composition $f = h \circ g$, where h is a simple path and $g'(t)$ never changes sign, then the length of f and the length of h are equal.

Definition 3.8

A set X in $\mathbb{C}$ is a *simple closed curve* if there exists a parametrization f of X so that f is injective on $[a, b)$ and $f(a) = f(b)$.

As a specific example, the unit circle $\mathbb{S}^1$ is a simple closed curve in $\mathbb{C}$, with the parametrization $g : [0, 2\pi] \to \mathbb{C}$ given by $g(t) = \cos(t) + i \sin(t)$.

3.2 The Element of Arc-length on $\mathbb{H}$

Our goal is to develop a means of measuring hyperbolic length and hyperbolic distance in $\mathbb{H}$, starting with $\text{Möb}(\mathbb{H})$ and showing that the information contained in knowing the hyperbolic lines in $\mathbb{H}$ and the transformations of $\mathbb{H}$ taking hyperbolic lines to hyperbolic lines is sufficient to determine hyperbolic length. To measure hyperbolic length, we need to find an appropriate hyperbolic element of arc-length. Thus, it seems reasonable to consider those elements of arc-length on $\mathbb{H}$ that are invariant under the action of $\text{Möb}(\mathbb{H})$.

Let ρ be a continuous nonzero function on $\mathbb{H}$. The element of arc-length $\rho(z)|\mathrm{d}z|$ on $\mathbb{H}$ is a conformal distortion of the standard Euclidean element of arc-length on $\mathbb{H}$, for which the length of a piecewise C^1 path $f : [a, b] \to \mathbb{H}$ is given by the integral

$$\text{length}_\rho(f) = \int_f \rho(z)|\mathrm{d}z| = \int_a^b \rho(f(t))\, |f'(t)|\, \mathrm{d}t.$$

Although it seems evident that this integral is finite for every piecewise C^1 path f in $\mathbb{H}$, we show in Proposition 3.13 that this is actually the case.

By the phrase *length is invariant under the action of* $\text{Möb}(\mathbb{H})$, we mean that for every piecewise C^1 path $f : [a, b] \to \mathbb{H}$ and every element γ of $\text{Möb}(\mathbb{H})$, we have

$$\text{length}_\rho(f) = \text{length}_\rho(\gamma \circ f).$$

Let us see what conditions this assumption imposes on ρ. We start by taking γ to be an element of $\text{Möb}^+(\mathbb{H})$. Expanding out $\text{length}_\rho(f)$ and $\text{length}_\rho(\gamma \circ f)$, we have

$$\text{length}_\rho(f) = \int_a^b \rho(f(t))\, |f'(t)|\, \mathrm{d}t$$

and

$$\text{length}_\rho(\gamma \circ f) = \int_a^b \rho((\gamma \circ f)(t))\, |(\gamma \circ f)'(t)|\, \mathrm{d}t,$$

and so we have that

$$\int_a^b \rho(f(t))\, |f'(t)|\, \mathrm{d}t = \int_a^b \rho((\gamma \circ f)(t))\, |(\gamma \circ f)'(t)|\, \mathrm{d}t$$

for every piecewise C^1 path $f : [a, b] \to \mathbb{H}$ and every element γ of $\text{Möb}^+(\mathbb{H})$.

Using the chain rule to expand $(\gamma \circ f)'(t)$ as $(\gamma \circ f)'(t) = \gamma'(f(t))\, f'(t)$ and substituting this into the integral for $\text{length}_\rho(\gamma \circ f)$, the condition on ρ becomes that

$$\int_a^b \rho(f(t))|f'(t)|\, \mathrm{d}t = \int_a^b \rho((\gamma \circ f)(t))\, |\gamma'(f(t))|\, |f'(t)|\, \mathrm{d}t$$

for every piecewise C^1 path $f : [a, b] \to \mathbb{H}$ and every element γ of $\text{Möb}^+(\mathbb{H})$.

Note 3.9

At this point, we need to insert a note about *differentiation of elements of* Möb. Unlike the case of functions of a single real variable, such as piecewise C^1 paths, there are two different ways in which to talk about the derivative of an element of Möb.

One way is to use complex analysis. That is, we view an element m of Möb as a function from $\overline{\mathbb{C}}$ to $\overline{\mathbb{C}}$, and define its derivative $m'(z)$ (using the usual definition) to be

$$m'(z) = \lim_{w \to z} \frac{m(w) - m(z)}{w - z},$$

if this limit exists. Using this definition, all usual formulae for derivatives hold, such as the product, quotient, and chain rules, and the derivative of an element $m(z) = \frac{az+b}{cz+d}$ of Möb^+ (normalized so that $ad - bc = 1$) is

$$m'(z) = \frac{1}{(cz + d)^2}.$$

This is the definition of differentiable we usually use, for instance, for Möbius transformations. These functions are often referred to as *holomorphic* or *analytic*; we will use *holomorphic*.

However, one disadvantage of this definition is that the derivative of an element of $\text{Möb} \setminus \text{Möb}^+$ is not defined. In particular, the derivative of $C(z) = \overline{z}$ does not exist, and so $C(z)$ is not holomorphic.

The second way of defining the derivative of an element of Möb is to use multivariable calculus. That is, we forget that an element m of Möb is a function of a complex variable and instead view it as a function taking an open subset $X \subset \mathbb{R}^2$ to an open subset $Y \subset \mathbb{R}^2$. In this case, the derivative is no longer a single function, but instead it is the 2×2 matrix of partial derivatives. That is, if we write z in terms of its real and imaginary parts as $z = x + iy$ and m in terms of its real and imaginary parts as $m(x, y) = (f(x, y), g(x, y))$, where f and g are real-valued functions, then the derivative of m is

$$Dm = \begin{pmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} \\ \frac{\partial g}{\partial x} & \frac{\partial g}{\partial y} \end{pmatrix}.$$

This definition of differentiable is used in the definition of hyperbolic area in Section 5.3. We refer to this notion of differentiability by saying that m is *differentiable as a function of x and y*.

It is true that holomorphic implies differentiable as a function of x and y, but not conversely. The distinction between these two definitions is one of the topics covered in complex analysis. This concludes Note 3.9.

Getting back to the argument in progress, the condition on $\rho(z)$ can be written as

$$\int_a^b (\rho(f(t)) - \rho((\gamma \circ f)(t)) \, |\gamma'(f(t))|) \; |f'(t)| \, \mathrm{d}t = 0$$

for every piecewise C^1 path $f : [a,b] \to \mathbb{H}$ and every element γ of $\text{Möb}^+(\mathbb{H})$.

For an element γ of $\text{Möb}^+(\mathbb{H})$, set

$$\mu_\gamma(z) = \rho(z) - \rho(\gamma(z))|\gamma'(z)|,$$

so that the condition on $\rho(z)$ becomes a condition on $\mu_\gamma(z)$, namely, that

$$\int_f \mu_\gamma(z)|\mathrm{d}z| = \int_a^b \mu_\gamma(f(t)) \, |f'(t)| \, \mathrm{d}t = 0$$

for every piecewise C^1 path $f : [a,b] \to \mathbb{H}$ and every element γ of $\text{Möb}^+(\mathbb{H})$. Note that, as $\rho(z)$ is continuous and γ is holomorphic, we have that $\mu_\gamma(z)$ is continuous for every element γ of $\text{Möb}^+(\mathbb{H})$.

This derived condition on $\mu_\gamma(z)$ is more apparently tractable than is the original condition on $\rho(z)$, as it is easier to subject to analysis. In particular, making use of this derived condition allows us to remove the requirement that we consider all piecewise C^1 paths in $\mathbb{H}$. This is the content of the following lemma.

Lemma 3.10

Let D be an open subset of $\mathbb{C}$, let $\mu : D \to \mathbb{R}$ be a continuous function, and suppose that $\int_f \mu(z)|\mathrm{d}z| = 0$ for every piecewise C^1 path $f : [a,b] \to D$. Then, $\mu \equiv 0$.

Proof

The proof of Lemma 3.10 is by contradiction, so suppose there exists a point $z \in D$ at which $\mu(z) \neq 0$. Replacing μ by $-\mu$ if necessary, we may assume that $\mu(z) > 0$.

The hypothesis that μ is continuous yields that for each $\varepsilon > 0$, there exists $\delta > 0$ so that $U_\delta(z) \subset D$, and that $w \in U_\delta(z)$ implies that $\mu(w) \in U_\varepsilon(\mu(z))$, where

$$U_\delta(z) = \{u \in \mathbb{C} \mid |u - z| < \delta\} \text{ and } U_\varepsilon(t) = \{s \in \mathbb{R} \mid |s - t| < \varepsilon\}.$$

Taking $\varepsilon = \frac{1}{3}|\mu(z)|$, we see that there exists $\delta > 0$ so that $w \in U_\delta(z)$ implies that $\mu(w) \in U_\varepsilon(\mu(z))$. Using the triangle inequality and that $\mu(z) > 0$, this implies that $\mu(w) > 0$ for all $w \in U_\delta(z)$.

We now choose a specific nonconstant C^1 path, namely, $f : [0, 1] \to U_\delta(z)$ given by

$$f(t) = z + \frac{1}{3}\delta t.$$

Observe that $\mu(f(t)) > 0$ for all t in $[0, 1]$, because $f(t) \in U_\delta(z)$ for all t in $[0, 1]$. In particular, we have that $\int_f \mu(z)|\mathrm{d}z| > 0$, which gives the desired contradiction. This completes the proof of Lemma 3.10. **QED**

Recall that we are assuming that length is invariant under the action of $\text{Möb}^+(\mathbb{H})$, which implies that $\int_f \mu_\gamma(z)|\mathrm{d}z| = 0$ for every piecewise C^1 path $f : [a, b] \to \mathbb{H}$ and every element γ of $\text{Möb}^+(\mathbb{H})$. Applying Lemma 3.10 to $\mu_\gamma(z)$, this leads us to the conclusion that

$$\mu_\gamma(z) = \rho(z) - \rho(\gamma(z))|\gamma'(z)| = 0$$

for every $z \in \mathbb{H}$ and every element γ of $\text{Möb}^+(\mathbb{H})$.

To simplify our analysis, we consider how μ_γ behaves under composition of elements of $\text{Möb}^+(\mathbb{H})$. Let γ and φ be two elements in $\text{Möb}^+(\mathbb{H})$. Calculating, we see that

$$\begin{aligned}
\mu_{\gamma\circ\varphi}(z) &= \rho(z) - \rho((\gamma\circ\varphi)(z))|(\gamma\circ\varphi)'(z)| \\
&= \rho(z) - \rho((\gamma\circ\varphi)(z))|\gamma'(\varphi(z))||\varphi'(z)| \\
&= \rho(z) - \rho(\varphi(z))|\varphi'(z)| + \rho(\varphi(z))|\varphi'(z)| \\
&\quad - \rho((\gamma\circ\varphi)(z))|\gamma'(\varphi(z))||\varphi'(z)| \\
&= \mu_\varphi(z) + \mu_\gamma(\varphi(z))|\varphi'(z)|.
\end{aligned}$$

In particular, if $\mu_\gamma \equiv 0$ for every γ in a generating set for $\text{Möb}^+(\mathbb{H})$, then $\mu_\gamma \equiv 0$ for every element γ of $\text{Möb}^+(\mathbb{H})$. We saw in Exercise 2.38 that there exists a generating set for $\text{Möb}^+(\mathbb{H})$ consisting of the transformations $m(z) = az + b$ for a, $b \in \mathbb{R}$ and $a > 0$, together with the transformation $K(z) = -\frac{1}{z}$.

Again, we are putting off consideration of $B(z) = -\overline{z}$ until later, as it is not an element of $\text{Möb}^+(\mathbb{H})$.

So, it suffices to analyze our condition on μ_γ, and hence on ρ, for the elements of this generating set. We consider these generators one at a time.

We first consider a generator of the form $\gamma(z) = z + b$ for $b \in \mathbb{R}$ (so that $a = 1$). As $\gamma'(z) = 1$ for every $z \in \mathbb{H}$, the condition imposed on $\rho(z)$ is that

$$0 = \mu_\gamma(z) = \rho(z) - \rho(\gamma(z))|\gamma'(z)| = \rho(z) - \rho(z + b)$$

for every $z \in \mathbb{H}$ and every $b \in \mathbb{R}$. That is,

$$\rho(z) = \rho(z + b)$$

for every $z \in \mathbb{H}$ and every $b \in \mathbb{R}$. In particular, $\rho(z)$ depends only on the imaginary part $y = \text{Im}(z)$ of $z = x + iy$. To see this explicitly, suppose that $z_1 = x_1 + iy$ and $z_2 = x_2 + iy$ have the same imaginary part, and write $z_2 = z_1 + (x_2 - x_1)$. As $x_2 - x_1$ is real, we have that $\rho(z_2) = \rho(z_1)$.

Hence, we may view ρ as a real-valued function of the single real variable $y = \text{Im}(z)$. Explicitly, consider the real-valued function $r : (0, \infty) \to (0, \infty)$ given by $r(y) = \rho(iy)$, and note that $\rho(z) = r(\text{Im}(z))$ for every $z \in \mathbb{H}$.

We now consider a generator of the form $\gamma(z) = az$ for $a > 0$ (so that $b = 0$). As $\gamma'(z) = a$ for every $z \in \mathbb{H}$, the condition imposed on $\rho(z)$ is that

$$0 = \mu_\gamma(z) = \rho(z) - \rho(\gamma(z))|\gamma'(z)| = \rho(z) - a\rho(az)$$

for every $z \in \mathbb{H}$ and every $a > 0$. That is,

$$\rho(z) = a\rho(az)$$

for every $z \in \mathbb{H}$ and every $a > 0$. In particular, we have

$$r(y) = ar(ay)$$

for every $y > 0$ and every $a > 0$. Dividing through by a, we have

$$r(ay) = \frac{1}{a}\, r(y).$$

Taking $y = 1$, this yields that

$$r(a) = \frac{1}{a}\, r(1),$$

and so r is completely determined by its value at 1.

Recalling the definition of r, we have that the assumption of invariance of length under $\text{Möb}^+(\mathbb{H})$ implies that $\rho(z)$ has the form

$$\rho(z) = r(\text{Im}(z)) = \frac{c}{\text{Im}(z)},$$

where c is an arbitrary positive constant.

Exercise 3.3

For a real number $\lambda > 0$, let A_λ be the Euclidean line segment joining $-1 + i\lambda$ to $1 + i\lambda$, and let B_λ be the hyperbolic line segment joining $-1 + i\lambda$ to $1 + i\lambda$. Calculate the lengths of A_λ and B_λ with respect to the element of arc-length $\frac{c}{\mathrm{Im}(z)}\,|\mathrm{d}z|$.

Note that the derivation of $\rho(z)$ we have just performed does not use alll generators of Möb($\mathbb{H}$). One question to be addressed is whether this form for $\rho(z)$ is consistent with lengths of piecewise C^1 paths being assumed to be invariant under both $K(z) = -\frac{1}{z}$ and $B(z) = -\overline{z}$.

Exercise 3.4

Check that the length of a piecewise C^1 path $f : [a, b] \to \mathbb{H}$ calculated with respect to the element of arc-length $\frac{c}{\mathrm{Im}(z)}\,|\mathrm{d}z|$ is invariant under both $K(z) = -\frac{1}{z}$ and $B(z) = -\overline{z}$. (Note that for $B(z)$, we cannot use the argument just given, as $B'(z)$ is not defined; instead, proceed directly by first evaluating the composition $B \circ f$ and then differentiating it as a path.)

Assuming the result of Exercise 3.4, we have proven the following theorem.

Theorem 3.11

For every positive constant c, the element of arc-length

$$\frac{c}{\mathrm{Im}(z)}\,|\mathrm{d}z|$$

on $\mathbb{H}$ is invariant under the action of Möb($\mathbb{H}$).

That is, for every piecewise C^1 path $f : [a, b] \to \mathbb{H}$ and every element γ of Möb($\mathbb{H}$), we have that

$$\mathrm{length}_\rho(f) = \mathrm{length}_\rho(\gamma \circ f).$$

However, nothing we have done to this point has given us a way of determining a specific value of c. In fact, it is not possible to specify the value of c using solely the action of Möb($\mathbb{H}$). To avoid carrying c through all our calculations, we set $c = 1$.

Definition 3.12

For a piecewise C^1 path $f : [a, b] \to \mathbb{H}$, we define the *hyperbolic length* of f to be

$$\text{length}_{\mathbb{H}}(f) = \int_f \frac{1}{\text{Im}(z)}\, |\mathrm{d}z| = \int_a^b \frac{1}{\text{Im}(f(t))}\, |f'(t)|\, \mathrm{d}t.$$

There are some piecewise C^1 paths whose hyperbolic length is straightforward to calculate. As an example, take $0 < a < b$ and consider the piecewise C^1 path $f : [a, b] \to \mathbb{H}$ given by $f(t) = it$. The image $f([a, b])$ of $[a, b]$ under f is the segment of the positive imaginary axis between ia and ib. As $\text{Im}(f(t)) = t$ and $|f'(t)| = 1$, we see that

$$\text{length}_{\mathbb{H}}(f) = \int_f \frac{1}{\text{Im}(z)} |\mathrm{d}z| = \int_a^b \frac{1}{t} \mathrm{d}t = \ln\left[\frac{b}{a}\right].$$

There are also piecewise C^1 paths whose hyperbolic length is more difficult to calculate.

Exercise 3.5

For each natural number n, write the integral for the hyperbolic length of the piecewise C^1 path $f_n : [0, 1] \to \mathbb{H}$ given by

$$f_n(t) = t + i(t^n + 1).$$

Exercise 3.6

For each piecewise C^1 path f_n defined in Exercise 3.5, make a conjecture about the behaviour of the hyperbolic length of $\gamma_n = f_n([0, 1])$ as $n \to \infty$, and calculate the putative limit of the hyperbolic length of γ_n as $n \to \infty$.

Exercise 3.7

Let G be the subgroup of Möb^+ generated by all *parabolic* Möbius transformations fixing ∞. Show that if $\lambda(z)|\mathrm{d}z|$ is an element of arc-length on $\mathbb{C}$ invariant under G, then $\lambda(z)$ is constant.

Exercise 3.8

Let H be the subgroup of Möb^+ consisting of all Möbius transformations fixing ∞. Show that if $\lambda(z)|\mathrm{d}z|$ is an element of arc-length on $\mathbb{C}$ invariant under H, then $\lambda(z) = 0$ for all $z \in \mathbb{C}$.

There is one subtlety regarding hyperbolic length that we mentioned at the beginning of this section that we need to address before going on, namely, that piecewise C^1 paths in $\mathbb{H}$ have finite hyperbolic length.

Proposition 3.13

Let $f : [a, b] \to \mathbb{H}$ be a piecewise C^1 path. Then, the hyperbolic length $\text{length}_{\mathbb{H}}(f)$ of f is finite.

Proof

The proof of Proposition 3.13 is an immediate consequence of the fact that there exists a constant $B > 0$ so that the image $f([a, b])$ of $[a, b]$ under f is contained in the subset

$$K_B = \{z \in \mathbb{H} \mid \text{Im}(z) \geq B\}$$

of $\mathbb{H}$. This fact follows as $[a, b]$, and hence $f([a, b])$, are *compact*, a concept discussed in more detail in Section 3.7.

Given that $f([a, b])$ is contained in K_B, we can estimate the integral giving the hyperbolic length of f. We first note that by the definition of a piecewise C^1 path, there is a partition P of $[a, b]$ into subintervals

$$P = \{[a = a_0, a_1], [a_1, a_2], \ldots, [a_n, a_{n+1} = b]\}$$

so that f is C^1 on each subinterval $[a_k, a_{k+1}]$.

In particular, its derivative $f'(t)$ is continuous on each subinterval. By the extreme value theorem for a continuous function on a closed interval, there then exists for each k a number A_k so that

$$|f'(t)| \leq A_k \text{ for all } t \in [a_k, a_{k+1}].$$

Let A be the maximum of $A_0, \ldots, A_n$. Then, we have that

$$\text{length}_{\mathbb{H}}(f) = \int_a^b \frac{1}{\text{Im}(f(t))} \, |f'(t)| \, \mathrm{d}t \leq \int_a^b \frac{1}{B} \, A \, \mathrm{d}t = \frac{A}{B} \, (b-a),$$

which is finite. This completes the proof of Proposition 3.13. **QED**

We close this section by noting that the proof of Proposition 3.13 gives a crude way of estimating an upper bound for the hyperbolic length of a piecewise C^1 path in $\mathbb{H}$.

3.3 Path Metric Spaces

We now know how to calculate the hyperbolic length of every piecewise C^1 path in $\mathbb{H}$, namely, by integrating the hyperbolic element of arc-length $\frac{1}{\text{Im}(z)} \, |\mathrm{d}z|$ along the path. We can now apply a general construction to pass from calculating hyperbolic lengths of paths in $\mathbb{H}$ to getting a hyperbolic metric on $\mathbb{H}$.

We begin by recalling the definition of a *metric*. Roughly, a metric on a set X is a means of consistently assigning a distance between pairs of points of X. We give only a brief and noncomprehensive description of metrics in this section. For a more detailed discussion of metrics, the interested reader should consult a textbook on point-set topology, such as Munkres [26].

Definition 3.14

A *metric* on a set X is a function

$$\mathrm{d} : X \times X \to \mathbb{R}$$

satisfying three conditions:

1. $\mathrm{d}(x, y) \geq 0$ for all $x,\ y \in X$, and $\mathrm{d}(x, y) = 0$ if and only if $x = y$.

2. $\mathrm{d}(x,y) = \mathrm{d}(y,x)$ for all x, $y \in X$.

3. $\mathrm{d}(x,z) \le \mathrm{d}(x,y) + \mathrm{d}(y,z)$ for all x, y, $z \in X$ (the triangle inequality).

If d is a metric on X, we often refer to the *metric space* (X, d). The notion of a metric is general, but it is good to keep in mind that we have already encountered several examples of metrics.

One example is the standard metric on $\mathbb{R}$ and $\mathbb{C}$ given by absolute value. On $\mathbb{C}$, this metric is given explicitly by the function

$$\mathrm{n} : \mathbb{C} \times \mathbb{C} \to \mathbb{R}, \text{ where } \mathrm{n}(z,w) = |z - w|.$$

The three conditions defining a metric on a general set can be thought of as an abstraction of the familiar properties of this function n.

A more complicated example is the metric on the Riemann sphere $\overline{\mathbb{C}}$ given by the function

$$\mathrm{s} : \overline{\mathbb{C}} \times \overline{\mathbb{C}} \to \mathbb{R},$$

where

$$\mathrm{s}(z,w) = \frac{2|z-w|}{\sqrt{(1+|z|^2)(1+|w|^2)}}$$

for z, $w \in \mathbb{C}$, and

$$\mathrm{s}(z,\infty) = \mathrm{s}(\infty,z) = \frac{2}{\sqrt{1+|z|^2}}$$

for $z \in \mathbb{C}$.

The proof that s is a metric on $\overline{\mathbb{C}}$ makes use of stereographic projection. These formulae are the expressions, in terms of the coordinate on $\overline{\mathbb{C}}$, of the Eulidean distances in $\mathbb{R}^3$ between the corresponding points on $\mathbb{S}^2$.

Note 3.15

Note that whenever we have a metric d on a space X, we can mimic in X the definitions of open and closed sets that we have in $\mathbb{C}$ and in $\overline{\mathbb{C}}$, and so we have notions of convergence of sequences in (X, d), and continuity of functions whose domain or range is the metric space (X, d).

Specifically, in the metric space (X, d), we define the *open disc* $U_\varepsilon(x)$ of radius $\varepsilon > 0$ centred at a point x to be

$$U_\varepsilon(x) = \{y \in X \mid \mathrm{d}(x,y) < \varepsilon\}.$$

Then, a subset A of X is *open* if for every $x \in A$, there exists some $\varepsilon > 0$ so that $U_\varepsilon(x) \subset A$; a subset B of X is *closed* if its complement $X - B$ is open.

A sequence $\{x_n\}$ of points of X *converges* to a point x of X if for every $\varepsilon > 0$, there exists some $N > 0$ so that $x_n \in U_\varepsilon(x)$ for all $n > N$.

We can also define *continuity* of functions between metric spaces. If (X, d_X) and (Y, d_Y) are two metric spaces and if $f : X \to Y$ is a function, then f is *continuous at a point x of X* if given $\varepsilon > 0$, there exists $\delta > 0$ so that $f(U_\delta(x)) \subset U_\varepsilon(f(x))$. We say that f is *continuous* if it is continuous at every point of X.

One example of a continuous function comes from the metric. Fix a point $z \in X$, and consider the function $f : X \to \mathbb{R}$ given by $f(x) = \mathrm{d}(z, x)$. Then, this function f is continuous. We actually make use of the continuity of this function in Section 5.1. This concludes Note 3.15.

One other example of a metric space will be important to us in our study of the hyperbolic plane. Let X be a set in which we know how to measure lengths of paths. Specifically, for each pair x and y of points in X, suppose there exists a nonempty collection $\mathit{\Gamma}[x, y]$ of paths $f : [a, b] \to X$ satisfying $f(a) = x$ and $f(b) = y$, and assume that to each path f in $\mathit{\Gamma}[x, y]$ we can associate in a reasonable way a nonnegative real number $\mathrm{length}(f)$, which we refer to as the *length of f*.

As an example to keep in mind, take X to be the upper half-plane $\mathbb{H}$, and take $\mathit{\Gamma}[x, y]$ to be the set of all piecewise C^1 paths $f : [a, b] \to \mathbb{H}$ with $f(a) = x$ and $f(b) = y$, where the length of each path f in $\mathit{\Gamma}[x, y]$ is the hyperbolic length $\mathrm{length}_{\mathbb{H}}(f)$ of f. Note that in this case, each $\mathit{\Gamma}[x, y]$ is nonempty, as we can parametrize the hyperbolic line segment joining x and y.

Consider the function $\mathrm{d} : X \times X \to \mathbb{R}$ defined by taking the infimum

$$\mathrm{d}(x, y) = \inf\{\mathrm{length}(f) \mid f \in \mathit{\Gamma}[x, y]\}.$$

There are several questions to ask about the construction of this function d. One question is as follows: What conditions on the definition of length are needed to determine whether d defines a metric on X? To avoid technical difficulties, we do not consider this question in general, as we are most interested in the case of the metric on $\mathbb{H}$ coming from hyperbolic lengths of paths, which we consider in detail in Section 3.4.

A second question is as follows: Assuming that d does indeed define a metric on X, do there necessarily exist *distance-realizing paths* in X? That is, given a pair x and y of points in X, does there necessarily exist a path f in $\Gamma[x,y]$ for which $\text{length}(f) = \text{d}(x,y)$?

As mentioned above, we consider both questions in detail for the upper half-plane $\mathbb{H}$ in Section 3.4. As an illustrative case, though, we consider some general properties of this construction for the case $X = \mathbb{C}$. We do not give any specific details, because they are similar to the details given below for $\mathbb{H}$.

For each pair x and y of points of $\mathbb{C}$, let $\Gamma[x,y]$ be the set of all piecewise C^1 paths $f : [a,b] \to \mathbb{C}$ with $f(a) = x$ and $f(b) = y$, and let $\text{length}(f)$ be the usual Euclidean length of f. In this case, because the shortest Euclidean distance between two points is along a Euclidean line, which can be parametrized by a C^1 path, we see that

$$\inf\{\text{length}(f) \mid f \in \Gamma[x,y]\} = \text{n}(x,y).$$

Note that in this case, this construction of a function on $\mathbb{C} \times \mathbb{C}$ by taking the infimum of the lengths of paths gives rise to the standard metric $\text{n}(\cdot,\cdot)$ on $\mathbb{C}$.

There is a related example that illustrates one of the difficulties that can arise. Let $X = \mathbb{C} - \{0\}$ be the punctured plane, and for each pair of points x and y of X, let $\Gamma[x,y]$ be the set of all piecewise C^1 paths $f : [a,b] \to X$ with $f(a) = x$ and $f(b) = y$.

In this case, we can bring what we know about the behaviour of $(\mathbb{C}, \text{n})$ to bear in our analysis of X. Again, this construction of a function on $X \times X$ by taking the infimum of the lengths of paths gives rise to the metric $\text{n}(x,y) = |x-y|$ on X.

However, we no longer have that there necessarily exists a path in $\Gamma[x,y]$ realizing the Euclidean distance between x and y. Specifically, consider the two points 1 and -1: The Euclidean line segment in $\mathbb{C}$ joining 1 to -1 passes through 0, and so it is not a path in X. Every other path joining 1 to -1 has length strictly greater than $\text{n}(1,-1) = 2$.

So, recall that we are working in a set X in which we know how to measure lengths of paths. For each pair x and y of points of X, there exists a nonempty collection $\Gamma[x,y]$ of paths $f : [a,b] \to X$ satisfying $f(a) = x$ and $f(b) = y$, and for each path f in $\Gamma[x,y]$, we denote the length of f by $\text{length}(f)$.

Suppose that, in addition, X is a metric space with metric d. We say that (X, d) is a *path metric space* if $\text{d}(x,y) = \inf\{\text{length}(f) \mid f \in \Gamma[x,y]\}$ for each pair of points x and y of X, and if there exists, for each pair of points x and y of X, a distance-realizing path in $\Gamma[x,y]$, which is a path f in $\Gamma[x,y]$ satisfying

$$\text{d}(x,y) = \text{length}(f).$$

We note that this definition of path metric space is stronger than the standard definition, as we require the existence of a distance realizing path between any pair of points of X.

Of the metric spaces mentioned in this section, we have with this definition that $(\mathbb{C}, \mathrm{n})$ and $(\overline{\mathbb{C}}, \mathrm{s})$ are path metric spaces, whereas $(\mathbb{C} - \{0\}, \mathrm{n})$ is not.

3.4 From Arc-length to Metric

We are now ready to prove that $\mathbb{H}$ is a path metric space. The proof of this fact takes up the bulk of this section.

For each pair of points x and y of $\mathbb{H}$, let $\mathit{\Gamma}[x, y]$ denote the set of all piecewise C^1 paths $f : [a, b] \to \mathbb{H}$ with $f(a) = x$ and $f(b) = y$.

As we can parametrize the hyperbolic line segment joining x to y by a piecewise C^1 path, we see that $\mathit{\Gamma}[x, y]$ is nonempty. Also, by Proposition 3.13, we know that every path f in $\mathit{\Gamma}[x, y]$ has finite hyperbolic length $\mathrm{length}_{\mathbb{H}}(f)$.

Consider the function

$$\mathrm{d}_{\mathbb{H}} : \mathbb{H} \times \mathbb{H} \to \mathbb{R}$$

defined by

$$\mathrm{d}_{\mathbb{H}}(x, y) = \inf\{\mathrm{length}_{\mathbb{H}}(f) \mid f \in \mathit{\Gamma}[x, y]\}.$$

In anticipation of the proof of Theorem 3.16, we refer to $\mathrm{d}_{\mathbb{H}}(x, y)$ as the *hyperbolic distance* between x and y.

Theorem 3.16

$(\mathbb{H}, \mathrm{d}_{\mathbb{H}})$ is a path metric space. Moreover, the distance-realizing paths in $\mathit{\Gamma}[x, y]$ are the almost simple parametrizations of the hyperbolic line segment joining x to y.

Proof

As the hyperbolic length of a path is invariant under the action of $\mathrm{M\ddot{o}b}(\mathbb{H})$, we have the following useful observation.

Proposition 3.17

For every element γ of Möb($\mathbb{H}$) and for every pair x and y of points of $\mathbb{H}$, we have that

$$\mathrm{d}_{\mathbb{H}}(x,y) = \mathrm{d}_{\mathbb{H}}(\gamma(x),\gamma(y)).$$

Proof

We begin by observing that $\{\gamma \circ f \mid f \in \mathit{\Gamma}[x,y]\} \subset \mathit{\Gamma}[\gamma(x),\gamma(y)]$. To see this, take a path $f : [a,b] \to \mathbb{H}$ in $\mathit{\Gamma}[x,y]$, so that $f(a) = x$ and $f(b) = y$. As $\gamma \circ f(a) = \gamma(x)$ and $\gamma \circ f(b) = \gamma(y)$, we have that $\gamma \circ f$ lies in $\mathit{\Gamma}[\gamma(x),\gamma(y)]$.

As $\mathrm{length}_{\mathbb{H}}(f)$ is invariant under the action of Möb($\mathbb{H}$), we have that

$$\mathrm{length}_{\mathbb{H}}(\gamma \circ f) = \mathrm{length}_{\mathbb{H}}(f)$$

for every path f in $\mathit{\Gamma}[x,y]$, and so

$$\begin{aligned} \mathrm{d}_{\mathbb{H}}(\gamma(x),\gamma(y)) &= \inf\{\mathrm{length}_{\mathbb{H}}(g) \mid g \in \mathit{\Gamma}[\gamma(x),\gamma(y)]\} \\ &\leq \inf\{\mathrm{length}_{\mathbb{H}}(\gamma \circ f) \mid f \in \mathit{\Gamma}[x,y]\} \\ &\leq \inf\{\mathrm{length}_{\mathbb{H}}(f) \mid f \in \mathit{\Gamma}[x,y]\} = \mathrm{d}_{\mathbb{H}}(x,y). \end{aligned}$$

As γ is invertible and γ^{-1} is an element of Möb($\mathbb{H}$), we may repeat the argument just given to see that

$$\{\gamma^{-1} \circ g \mid g \in \mathit{\Gamma}[\gamma(x),\gamma(y)]\} \subset \mathit{\Gamma}[x,y],$$

and hence that

$$\begin{aligned} \mathrm{d}_{\mathbb{H}}(x,y) &= \inf\{\mathrm{length}_{\mathbb{H}}(f) \mid f \in \mathit{\Gamma}[x,y]\} \\ &\leq \inf\{\mathrm{length}_{\mathbb{H}}(\gamma^{-1} \circ g) \mid g \in \mathit{\Gamma}[\gamma(x),\gamma(y)]\} \\ &\leq \inf\{\mathrm{length}_{\mathbb{H}}(g) \mid g \in \mathit{\Gamma}[\gamma(x),\gamma(y)]\} = \mathrm{d}_{\mathbb{H}}(\gamma(x),\gamma(y)). \end{aligned}$$

In particular, we have that $\mathrm{d}_{\mathbb{H}}(x,y) = \mathrm{d}_{\mathbb{H}}(\gamma(x),\gamma(y))$. This completes the proof of Proposition 3.17. **QED**

To show that $\mathrm{d}_{\mathbb{H}}$ does indeed define a metric, we need to show that $\mathrm{d}_{\mathbb{H}}$ satisfies the three conditions given in Definition 3.14.

Let $f : [a,b] \to \mathbb{H}$ be a path in $\mathit{\Gamma}[x,y]$, and recall the definition of $\mathrm{length}_{\mathbb{H}}(f)$:

$$\mathrm{length}_{\mathbb{H}}(f) = \int_f \frac{1}{\mathrm{Im}(z)}\,|\mathrm{d}z| = \int_a^b \frac{1}{\mathrm{Im}(f(t))}\,|f'(t)|\mathrm{d}t.$$

As the integrand is always nonnegative, it is immediate that the integral is nonnegative. As $\text{length}_{\mathbb{H}}(f)$ is nonnegative for every path f in $\Gamma[x,y]$, the infimum $\mathrm{d}_{\mathbb{H}}(x,y)$ of these integrals is nonnegative, which shows that the first part of Condition 1 of the definition of a metric is satisfied by $\mathrm{d}_{\mathbb{H}}$. For reasons that will become clear at the time, the proof that $\mathrm{d}_{\mathbb{H}}$ satisfies the second part of Condition 1 is postponed to later in the section.

We now consider Condition 2 of the definition of a metric. We need to compare the lengths of paths in $\Gamma[x,y]$ and $\Gamma[y,x]$. Let $f : [a,b] \to \mathbb{H}$ be a path in $\Gamma[x,y]$, and consider the composition of f with the function $h : [a,b] \to [a,b]$ given by $h(t) = a + b - t$. Note that $h'(t) = -1$. This is a specific example of reparametrization.

It is evident that $f \circ h$ lies in $\Gamma[y,x]$, because $(f \circ h)(a) = f(b) = y$ and $(f \circ h)(b) = f(a) = x$. Moreover, direct calculation with the substitution $s = h(t)$ yields that

$$\begin{aligned}
\text{length}_{\mathbb{H}}(f \circ h) &= \int_{f\circ h} \frac{1}{\mathrm{Im}(z)}\,|\mathrm{d}z| \\
&= \int_a^b \frac{1}{\mathrm{Im}((f\circ h)(t))}\,|(f\circ h)'(t)|\,\mathrm{d}t \\
&= \int_a^b \frac{1}{\mathrm{Im}(f(h(t)))}\,|f'(h(t))|\,|h'(t)|\,\mathrm{d}t \\
&= -\int_b^a \frac{1}{\mathrm{Im}(f(s))}\,|f'(s)|\,\mathrm{d}s \\
&= \int_a^b \frac{1}{\mathrm{Im}(f(s))}\,|f'(s)|\,\mathrm{d}s = \text{length}_{\mathbb{H}}(f).
\end{aligned}$$

So, every path in $\Gamma[x,y]$ gives rise to a path in $\Gamma[y,x]$ of equal length, by composing with the appropriate h. Using the same argument, every path in $\Gamma[y,x]$ gives rise to a path in $\Gamma[x,y]$ of equal length.

In particular, we see that the two sets of hyperbolic lengths

$$\{\text{length}_{\mathbb{H}}(f) \mid f \in \Gamma[x,y]\} \text{ and } \{\text{length}_{\mathbb{H}}(g) \mid g \in \Gamma[y,x]\}$$

are equal. Hence, they have the same infimum, and so $\mathrm{d}_{\mathbb{H}}(x,y) = \mathrm{d}_{\mathbb{H}}(y,x)$. This completes the proof that Condition 2 of the definition of a metric is satisfied by $\mathrm{d}_{\mathbb{H}}$.

We now consider Condition 3 of the definition of a metric, the triangle inequality. To this end, let x, y, and z be points in $\mathbb{H}$.

Conceptually, the simplest proof would be for us to choose a path $f : [a,b] \to \mathbb{H}$ in $\Gamma[x,y]$ with $\text{length}_{\mathbb{H}}(f) = \mathrm{d}_{\mathbb{H}}(x,y)$ and a path $g : [b,c] \to \mathbb{H}$ in $\Gamma[y,z]$ with

$\text{length}_{\mathbb{H}}(g) = \text{d}_{\mathbb{H}}(y, z)$. The concatenation $h : [a, c] \to \mathbb{H}$ of f and g would then lie in $\Gamma[x, z]$. Moreover, we would have the desired inequality

$$\text{d}_{\mathbb{H}}(x, z) \leq \text{length}_{\mathbb{H}}(h) = \text{length}_{\mathbb{H}}(f) + \text{length}_{\mathbb{H}}(g) = \text{d}_{\mathbb{H}}(x, y) + \text{d}_{\mathbb{H}}(y, z).$$

We note here that the concatenation of piecewise C^1 paths is again a piecewise C^1 path, whereas the concatenation of C^1 paths is not necessarily a C^1 path. This observation is the main reason to consider piecewise C^1 paths instead of C^1 paths.

Unfortunately, we do not yet know that there always exists a path realizing the hyperbolic distance between a pair of points. We consider this question later in this section. For now, we take a route that is slightly more roundabout. We use proof by contradiction.

Suppose that Condition 3, the triangle inequality, does not hold for $\text{d}_{\mathbb{H}}$. That is, suppose that there exist distinct points x, y, and z in $\mathbb{H}$ so that

$$\text{d}_{\mathbb{H}}(x, z) > \text{d}_{\mathbb{H}}(x, y) + \text{d}_{\mathbb{H}}(y, z).$$

Set

$$\varepsilon = \text{d}_{\mathbb{H}}(x, z) - (\text{d}_{\mathbb{H}}(x, y) + \text{d}_{\mathbb{H}}(y, z)).$$

As $\text{d}_{\mathbb{H}}(x, y) = \inf\{\text{length}_{\mathbb{H}}(f) \mid f \in \Gamma[x, y]\}$, there exists a path $f : [a, b] \to \mathbb{H}$ in $\Gamma[x, y]$ with

$$\text{length}_{\mathbb{H}}(f) - \text{d}_{\mathbb{H}}(x, y) < \frac{1}{2}\varepsilon.$$

Similarly, there exists a path $g : [b, c] \to \mathbb{H}$ in $\Gamma[y, z]$ with

$$\text{length}_{\mathbb{H}}(g) - \text{d}_{\mathbb{H}}(y, z) < \frac{1}{2}\varepsilon.$$

Recall that we can choose the domains of definition of f and g at will, using our discussion of reparametrization in Section 3.1.

Let $h : [a, c] \to \mathbb{H}$ be the concatenation of f and g. As the concatenation of two piecewise C^1 paths is again a piecewise C^1 path, we have that h lies in $\Gamma[x, z]$. Calculating, we see that

$$\text{length}_{\mathbb{H}}(h) = \text{length}_{\mathbb{H}}(f) + \text{length}_{\mathbb{H}}(g) < \text{d}_{\mathbb{H}}(x, y) + \text{d}_{\mathbb{H}}(y, z) + \varepsilon.$$

As $\text{d}_{\mathbb{H}}(x, z) \leq \text{length}_{\mathbb{H}}(h)$ by definition of $\text{d}_{\mathbb{H}}$, this gives that

$$\text{d}_{\mathbb{H}}(x, z) < \text{d}_{\mathbb{H}}(x, y) + \text{d}_{\mathbb{H}}(y, z) + \varepsilon,$$

which contradicts the construction of ε. This completes the proof that Condition 3 of the definition of a metric is satisfied by $\text{d}_{\mathbb{H}}$.

Two things remain to be checked before we can conclude that $(\mathbb{H}, \mathrm{d}_{\mathbb{H}})$ is a path metric space. We need to show that $\mathrm{d}_{\mathbb{H}}$ satisfies the second part of Condition 1 of the definition of a metric, and we need to show that there exists a distance-realizing path $\mathrm{d}_{\mathbb{H}}(x, y)$ between any pair of points x and y of $\mathbb{H}$.

The approach we take comes from the observation that if there exists a path in $\mathbb{H}$ realizing the hyperbolic distance between any pair of points of $\mathbb{H}$, then this implies that $\mathrm{d}_{\mathbb{H}}(x, y) > 0$ for $x \neq y$, because the lengths of nonconstant paths are positive. Thus, we would have the second part of Condition 1 for free.

So, let x and y be a pair of distinct points of $\mathbb{H}$, and let ℓ be the hyperbolic line passing through x and y. We begin by simplifying the situation. From our work in Section 2.9, specifically Exercise 2.40, we know that there exists an element γ of $\text{Möb}(\mathbb{H})$ so that $\gamma(\ell)$ is the positive imaginary axis in $\mathbb{H}$.

Write $\gamma(x) = \mu i$ and $\gamma(y) = \lambda i$ with $\mu < \lambda$. If $\lambda < \mu$, then use $K \circ \gamma$ instead of γ, where $K(z) = -\frac{1}{z}$, so that $\mu < \lambda$. By Proposition 3.17, we have that $\mathrm{d}_{\mathbb{H}}(x, y) = \mathrm{d}_{\mathbb{H}}(\gamma(x), \gamma(y))$. So, we have reduced ourselves to showing that there exists a distance-realizing path between μi and λi for $\mu < \lambda$.

We begin by calculating the hyperbolic length of a specific path, namely, the path $f_0 : [\mu, \lambda] \to \mathbb{H}$ defined by $f_0(t) = it$. The image of f_0 is the hyperbolic line segment joining μi and λi. As we expect the shortest hyperbolic distance between two points to be along a hyperbolic line, this path seems to be a reasonable choice to be the shortest path in $\Gamma[\mu i, \lambda i]$.

To calculate the length of f_0, we observe that $\mathrm{Im}(f_0(t)) = t$ and $|f_0'(t)| = 1$, and so

$$\text{length}_{\mathbb{H}}(f_0) = \int_{\mu}^{\lambda} \frac{1}{t}\, \mathrm{d}t = \ln\left[\frac{\lambda}{\mu}\right].$$

Now, let $f : [a, b] \to \mathbb{H}$ be any path in $\Gamma[\mu i, \lambda i]$. We complete the proof that $\text{length}_{\mathbb{H}}(f_0) = \mathrm{d}_{\mathbb{H}}(\mu i, \lambda i)$ by showing that $\text{length}_{\mathbb{H}}(f_0) \leq \text{length}_{\mathbb{H}}(f)$. We prove this in several stages, at each stage modifying f to decrease its hyperbolic length, and arguing that it becomes no shorter than f_0 through these modifications.

Write $f(t) = x(t) + y(t)i$. The first modification of f is to ignore the real part. That is, consider the path $g : [a, b] \to \mathbb{H}$ defined by setting

$$g(t) = \mathrm{Im}(f(t))i = y(t)i.$$

As $g(a) = f(a) = \mu i$ and $g(b) = f(b) = \lambda i$, we see that g lies in $\Gamma[\mu i, \lambda i]$.

Using that $(x'(t))^2 \geq 0$ and that $\mathrm{Im}(g(t)) = \mathrm{Im}(f(t)) = y(t)$ for all t, we have that

$$\begin{aligned}
\mathrm{length}_{\mathbb{H}}(g) &= \int_a^b \frac{1}{\mathrm{Im}(g(t))}\, |g'(t)|\, \mathrm{d}t \\
&= \int_a^b \frac{1}{y(t)} \sqrt{(y'(t))^2}\, \mathrm{d}t \\
&\leq \int_a^b \frac{1}{y(t)} \sqrt{(x'(t))^2 + (y'(t))^2}\, \mathrm{d}t \\
&\leq \int_a^b \frac{1}{\mathrm{Im}(f(t))}\, |f'(t)|\, \mathrm{d}t = \mathrm{length}_{\mathbb{H}}(f).
\end{aligned}$$

So, given any path f in $\Gamma[\mu i, \lambda i]$, we can construct a path g in $\Gamma[\mu i, \lambda i]$ whose length is at most the length of f, by setting $g(t) = \mathrm{Im}(f(t))\, i$. To complete the proof, we need only to show that if $g : [a,b] \to \mathbb{H}$ is any path in $\Gamma[\mu i, \lambda i]$ of the form $g(t) = y(t)i$, then

$$\mathrm{length}_{\mathbb{H}}(f_0) \leq \mathrm{length}_{\mathbb{H}}(g).$$

This fact follows from Proposition 3.2. The image $g([a,b])$ of g is the hyperbolic line segment joining αi and βi, where $\alpha \leq \mu < \lambda \leq \beta$. Define $f_1 : [\alpha, \beta] \to \mathbb{H}$ by $f_1(t) = it$. By restricting the range of f_1, we can consider f_1 as a homeomorphism $f_1 : [\alpha, \beta] \to \ell_{\alpha i, \beta i}$, where $\ell_{\alpha i, \beta i}$ is the hyperbolic line segment joining αi and βi. Calculating, we see that

$$\mathrm{length}_{\mathbb{H}}(f_0) = \ln\left[\frac{\lambda}{\mu}\right] \leq \ln\left[\frac{\beta}{\alpha}\right] = \mathrm{length}_{\mathbb{H}}(f_1).$$

Then, we can write $g = f_1 \circ (f_1^{-1} \circ g)$, where $f_1^{-1} \circ g : [a,b] \to [\alpha, \beta]$ is by construction a surjective function. By Proposition 3.2,

$$\mathrm{length}_{\mathbb{H}}(f_1) \leq \mathrm{length}_{\mathbb{H}}(g).$$

This completes the argument that

$$\mathrm{length}_{\mathbb{H}}(f_0) \leq \mathrm{length}_{\mathbb{H}}(f)$$

for every path f in $\Gamma[\mu i, \lambda i]$. That is, we have shown that

$$\mathrm{d}_{\mathbb{H}}(\mu i, \lambda i) = \mathrm{length}_{\mathbb{H}}(f_0) = \ln\left[\frac{\lambda}{\mu}\right].$$

Note that because we have written $g(t) = y(t)i$ and $f_1(t) = it$, we have that $f_1^{-1} \circ g(t) = y(t)$, and so by Proposition 3.2, we have

$$\mathrm{length}_{\mathbb{H}}(g) = \mathrm{length}_{\mathbb{H}}(f_1)$$

if and only if either $y'(t) \geq 0$ for all t in $[a, b]$ or $y'(t) \leq 0$ for all t in $[a, b]$. That is, the only distance realizing paths in $\Gamma[\mu i, \lambda i]$ are those that are almost simple parametrizations of the hyperbolic line segment joining μi and λi.

Exercise 3.9

Consider the path $g : [-1, 1] \to \mathbb{H}$ given by

$$g(t) = (t^2 + 1)i.$$

Determine the image of g in $\mathbb{H}$, and calculate $\text{length}_{\mathbb{H}}(g)$.

The transitivity of Möb($\mathbb{H}$) on the set of hyperbolic lines in $\mathbb{H}$ and the invariance of both hyperbolic lengths of paths in $\mathbb{H}$ and hyperbolic distances between pairs of points of $\mathbb{H}$ under the action of Möb($\mathbb{H}$) combine to yield that for any pair of points x and y in $\mathbb{H}$, there exists a distance-realizing path in $\Gamma[x, y]$, namely, any almost simple parametrization of the hyperbolic line segment joining x to y.

Explicitly, let ℓ be the hyperbolic line passing through x and y, and let γ be an element of Möb($\mathbb{H}$) taking ℓ to the positive imaginary axis I. Write $\gamma(x) = \mu i$ and $\gamma(y) = \lambda i$. Note that as before we can choose γ so that $\mu < \lambda$: if $\mu > \lambda$, then replace γ with $K \circ \gamma$, where $K(z) = -\frac{1}{z}$.

We have just seen that the simple path $f_0 : [\mu, \lambda] \to \mathbb{H}$ given by $f_0(t) = ti$ is a distance-realizing path in $\Gamma[\mu i, \lambda i]$. As Möb($\mathbb{H}$) preserves hyperbolic lengths of paths, we have that

$$\text{length}_{\mathbb{H}}(\gamma^{-1} \circ f_0) = \text{length}_{\mathbb{H}}(f_0).$$

As Möb($\mathbb{H}$) preserves hyperbolic distance, we have that

$$\mathrm{d}_{\mathbb{H}}(x, y) = \mathrm{d}_{\mathbb{H}}(\gamma^{-1}(\mu i), \gamma^{-1}(\lambda i)) = \mathrm{d}_{\mathbb{H}}(\mu i, \lambda i) = \text{length}_{\mathbb{H}}(f_0).$$

Combining these equations yields that

$$\text{length}_{\mathbb{H}}(\gamma^{-1} \circ f_0) = \mathrm{d}_{\mathbb{H}}(x, y),$$

and so $\gamma^{-1} \circ f_0$ is a distance-realizing path in $\Gamma[x, y]$.

As mentioned at the beginning of this section, this also completes the proof that the second part of Condition 1 of the definition of a metric is satisfied by $\mathrm{d}_{\mathbb{H}}$. So, $(\mathbb{H}, \mathrm{d}_{\mathbb{H}})$ is a path metric space. This completes the proof of Theorem 3.16. **QED**

Exercise 3.10

Let S be the hyperbolic line segment between $2i$ and $10i$. For each $n \geq 2$, find the points that divide S into n segments of equal length.

As we have that $\mathrm{d}_{\mathbb{H}}(x, y)$ is a metric on $\mathbb{H}$, the discussion in Note 3.15 yields that we now have notions of open and closed sets in $\mathbb{H}$, of convergent sequences of points of $\mathbb{H}$, and of continuous functions with domain and range $\mathbb{H}$.

We close this section by justifying why the boundary at infinity $\overline{\mathbb{R}} = \partial\mathbb{H}$ of $\mathbb{H}$ is called the boundary at infinity. Choose a point z on the boundary at infinity $\overline{\mathbb{R}}$ of $\mathbb{H}$, say $z = \infty$, and consider the hyperbolic ray ℓ determined by i and ∞.

As ℓ can be expressed as the image of the path $f : [1, \infty) \to \mathbb{H}$ given by $f(t) = it$, the hyperbolic distance between i and ∞ is equal to the length of f, namely, the improper integral

$$\mathrm{length}_{\mathbb{H}}(f) = \int_1^\infty \frac{1}{t}\,\mathrm{d}t,$$

which is infinite.

In particular, even though the points of $\overline{\mathbb{R}}$ form the topological boundary $\partial\mathbb{H}$ of $\mathbb{H}$ when we view $\mathbb{H}$ as a disc in $\overline{\mathbb{C}}$, the points of $\overline{\mathbb{R}}$ are infinitely far away from the points of $\mathbb{H}$ in terms of the hyperbolic metric on $\mathbb{H}$.

3.5 Formulae for Hyperbolic Distance in $\mathbb{H}$

The proof of Theorem 3.16 gives a method for calculating the hyperbolic distance between a pair of points in $\mathbb{H}$.

Given a pair of points x and y in $\mathbb{H}$, find or construct an element γ of Möb($\mathbb{H}$) so that $\gamma(x) = i\mu$ and $\gamma(y) = i\lambda$ both lie on the positive imaginary axis. Then, determine the values of μ and λ to find the hyperbolic distance

$$\mathrm{d}_{\mathbb{H}}(x, y) = \mathrm{d}_{\mathbb{H}}(\mu i, \lambda i) = \left| \ln\left[\frac{\lambda}{\mu}\right] \right|.$$

Note that here we use the absolute value, as we have made no assumption about whether $\lambda < \mu$ or $\mu < \lambda$.

For example, consider the two points $x = 2 + i$ and $y = -3 + i$. By Exercise 1.3, the hyperbolic line ℓ passing through x and y lies in the Euclidean circle with

Euclidean centre $-\frac{1}{2}$ and Euclidean radius $\frac{\sqrt{29}}{2}$. In particular, the endpoints at infinity of ℓ are

$$p = \frac{-1+\sqrt{29}}{2} \text{ and } q = \frac{-1-\sqrt{29}}{2}.$$

Set $\gamma(z) = \frac{z-p}{z-q}$. The determinant of γ is $p - q > 0$, and so γ lies in $\text{Möb}^+(\mathbb{H})$. As by construction γ takes the endpoints at infinity of ℓ to the endpoints at infinity of the positive imaginary axis, namely, 0 and ∞, we see that γ takes ℓ to the positive imaginary axis.

Calculating, we see that

$$\gamma(2+i) = \frac{2+i-p}{2+i-q} = \frac{p-q}{(2-q)^2+1}\, i$$

and

$$\gamma(-3+i) = \frac{-3+i-p}{-3+i-q} = \frac{p-q}{(3+q)^2+1}\, i.$$

In particular, we have that

$$\begin{aligned} d_{\mathbb{H}}(2+i,-3+i) = d_{\mathbb{H}}(\gamma(2+i),\gamma(-3+i)) &= \left| \ln\left[\frac{(2-q)^2+1}{(3+q)^2+1}\right]\right| \\ &= \ln\left[\frac{58+10\sqrt{29}}{58-10\sqrt{29}}\right]. \end{aligned}$$

As is demonstrated by this example, going through this procedure can be extremely tedious. It would be preferable to have an explicit and general formula of calculating hyperbolic distance. One way would be to repeat the procedure carried out in this example for a general pair of points z_1 and z_2.

Exercise 3.11

Let $z_1 = x_1 + iy_1$ and $z_2 = x_2 + iy_2$ be two points in $\mathbb{H}$ with $x_1 \neq x_2$. Derive a formula for $d_{\mathbb{H}}(z_1, z_2)$ in terms of x_1, y_1, x_2, and y_2 by constructing an element γ of $\text{Möb}(\mathbb{H})$ so that $\gamma(z_1)$ and $\gamma(z_2)$ both lie on the positive imaginary axis.

Exercise 3.12

Calculate the hyperbolic distance between each pair of the four points $A = i$, $B = 1 + 2i$, $C = -1 + 2i$, and $D = 7i$.

A related formula for the hyperbolic distance $\mathrm{d}_{\mathbb{H}}(z_1, z_2)$ between z_1 and z_2 in terms of their real and imaginary parts can be derived by making use of the fact that hyperbolic lines lie in Euclidean circles and Euclidean lines perpendicular to $\overline{\mathbb{R}}$. As above, write $z_1 = x_1 + y_1 i$ and $z_2 = x_2 + y_2 i$.

We can assume that $x_1 \neq x_2$, because in the case in which $x_1 = x_2$, we have already seen that

$$\mathrm{d}_{\mathbb{H}}(z_1, z_2) = \left| \ln \left[\frac{y_2}{y_1} \right] \right|.$$

Let c be the Euclidean centre and r the Euclidean radius of the Euclidean circle containing the hyperbolic line passing through z_1 and z_2. Suppose that $x_1 > x_2$, and let θ_k be the argument of z_k, taken in the range $[0, \pi)$ and as usual measured counterclockwise from the positive real axis.

Consider the path $f : [\theta_1, \theta_2] \to \mathbb{H}$ given by $f(t) = c + re^{it}$. The image of f is the hyperbolic line segment between z_1 and z_2, and so $\mathrm{d}_{\mathbb{H}}(z_1, z_2) = \mathrm{length}_{\mathbb{H}}(f)$.

As $\mathrm{Im}(f(t)) = r\sin(t)$ and $|f'(t)| = |rie^{it}| = r$, we have that

$$\mathrm{d}_{\mathbb{H}}(z_1, z_2) = \mathrm{length}_{\mathbb{H}}(f) = \int_{\theta_1}^{\theta_2} \frac{1}{\sin(t)} \mathrm{d}t = \ln \left| \frac{\csc(\theta_2) - \cot(\theta_2)}{\csc(\theta_1) - \cot(\theta_1)} \right|.$$

To rewrite this expression in terms of x_1, x_2, y_1, and y_2, it is possible but not necessary to express the θ_k in terms of the x_k and y_k. We might also express $\csc(\theta_k)$ and $\cot(\theta_k)$ in terms of the x_k and y_k, and c and r.

Note that θ_k is the angle of the right triangle with opposite side y_k, adjacent side $x_k - c$, and hypotenuse r. So, we have that

$$\csc(\theta_k) = \frac{r}{y_k} \text{ and } \cot(\theta_k) = \frac{x_k - c}{y_k}.$$

These equations give that

$$|\csc(\theta_k) - \cot(\theta_k)| = \left| \frac{r + c - x_k}{y_k} \right|,$$

and so

$$\mathrm{d}_{\mathbb{H}}(z_1, z_2) = \mathrm{length}_{\mathbb{H}}(f) = \ln \left| \frac{\csc(\theta_2) - \cot(\theta_2)}{\csc(\theta_1) - \cot(\theta_1)} \right| = \ln \left| \frac{(x_1 - c - r)y_2}{y_1(x_2 - c - r)} \right|.$$

Note that if instead we have that $x_2 < x_1$ and we go through this calculation, we get that

$$\mathrm{d}_{\mathbb{H}}(z_1, z_2) = \mathrm{length}_{\mathbb{H}}(f) = \ln \left| \frac{\csc(\theta_1) - \cot(\theta_1)}{\csc(\theta_2) - \cot(\theta_2)} \right| = \ln \left| \frac{y_1(x_2 - c - r)}{(x_1 - c - r)y_2} \right|,$$

which differs from $\ln\left|\frac{(x_1-c-r)y_2}{y_1(x_2-c-r)}\right|$ by a factor of -1.

So, if we make no assumption of the relationship between x_1 and x_2, we obtain the formula

$$\mathrm{d}_{\mathbb{H}}(z_1, z_2) = \left|\ln\left|\frac{(x_1-c-r)y_2}{y_1(x_2-c-r)}\right|\right|$$

for the hyperbolic distance between z_1 and z_2.

If we wish to express this formula solely in terms of the x_k and y_k, we may recall the result of Exercise 1.3, in which we gave expressions for c and r in terms of the x_k and y_k. Unfortunately, the resulting expression does not simplify much, and so we do not give it explicitly here.

Although it can be unwieldy, we can sometimes make explicit use of this formula. For example, we can determine whether or not there exists a positive real number s so that

$$\mathrm{d}_{\mathbb{H}}(-s+i, i) = \mathrm{d}_{\mathbb{H}}(i, s+i) = \mathrm{d}_{\mathbb{H}}(-s+i, s+i).$$

As $-s+i$ and $s+i$ lie on the Euclidean circle with Euclidean centre $c=0$ and Euclidean radius $r=\sqrt{1+s^2}$, we have that

$$\mathrm{d}_{\mathbb{H}}(-s+i, s+i) = \ln\left[\frac{\sqrt{s^2+1}+s}{\sqrt{s^2+1}-s}\right].$$

As $s+i$ and i lie on the Euclidean circle with Euclidean centre $c=\frac{s}{2}$ and Euclidean radius $r=\frac{1}{2}\sqrt{4+s^2}$, we have that

$$\mathrm{d}_{\mathbb{H}}(s+i, i) = \ln\left[\frac{\sqrt{s^2+4}+s}{\sqrt{s^2+4}-s}\right].$$

As there is no solution to

$$\ln\left[\frac{\sqrt{s^2+1}+s}{\sqrt{s^2+1}-s}\right] = \ln\left[\frac{\sqrt{s^2+4}+s}{\sqrt{s^2+4}-s}\right],$$

no such value of s exists, and so there does not exist an equilateral hyperbolic triangle with vertices on the horocircle $\{z \in \mathbb{H} \mid \mathrm{Im}(z) = 1\}$.

Now that we understand hyperbolic distance, and specifically now that we have a notion of hyperbolic distance that is invariant under the action of Möb($\mathbb{H}$), we can see the obstruction to Möb($\mathbb{H}$) acting transitively on pairs of distinct points of $\mathbb{H}$.

Exercise 3.13

Given two pairs (z_1, z_2) and (w_1, w_2) of distinct points of $\mathbb{H}$, prove that there exists an element q of $\text{Möb}(\mathbb{H})$ satisfying $q(z_1) = w_1$ and $q(z_2) = w_2$ if and only if $\text{d}_{\mathbb{H}}(z_1, z_2) = \text{d}_{\mathbb{H}}(w_1, w_2)$.

Recall from the discussion in Section 2.10 that if m is a loxodromic transformation in $\text{Möb}^+(\mathbb{H})$ fixing points x and y in $\overline{\mathbb{R}}$, and if A is any circle in $\overline{\mathbb{C}}$ that passes through x and y, then m takes $A \cap \mathbb{H}$ to itself. Now that we understand and can calculate hyperbolic distance in $\mathbb{H}$, we can see how m acts on $A \cap \mathbb{H}$.

Definition 3.18

The *translation distance* of m along $A \cap \mathbb{H}$ is $\text{d}_{\mathbb{H}}(a, m(a))$, where a is a point of $A \cap \mathbb{H}$.

In the case in which $A \cap \mathbb{H}$ is equal to the axis of m, which occurs in the case in which A is perpendicular to $\overline{\mathbb{R}}$, we have already calculated the translation distance of m along $A \cap \mathbb{H}$ to be

$$\text{d}_{\mathbb{H}}(\mu i, m(\mu i)) = \text{d}_{\mathbb{H}}(\mu i, \lambda\mu i) = \ln\left[\frac{\lambda\mu}{\mu}\right] = \ln(\lambda),$$

where $m(z)$ is conjugate to $q(z) = \lambda z$.

3.6 Isometries

In general, an *isometry* of a metric space (X, d) is a homeomorphism f of X that preserves distance. That is, an isometry of (X, d) is a homeomorphism f of X for which

$$\text{d}(x, y) = \text{d}(f(x), f(y))$$

for every pair x and y of points of X. In fact, as is demonstrated in the following exercise, this definition of an isometry is partially redundant.

Exercise 3.14

Let $f : X \to X$ be any function that preserves distance. Prove that f is injective and continuous.

In general, we cannot conclude that a distance preserving function $f : X \to X$ is a homeomorphism. To illustrate one thing that can go wrong, consider the metric e on $\mathbb{Z}$ defined by setting

$$\mathrm{e}(n, m) = \begin{cases} 0 & \text{if } m = n, \text{ and} \\ 1 & \text{if } m \neq n. \end{cases}$$

This function e gives a metric on $\mathbb{Z}$ that is different from the usual metric on $\mathbb{Z}$. The function $f : \mathbb{Z} \to \mathbb{Z}$ defined by $f(m) = 2m$ is distance preserving but is not surjective, and hence it is not a homeomorphism.

It is true, though, that a distance preserving function $f : X \to X$ is a homeomorphism onto its image $f(X)$, because f is a bijection when considered as a function $f : X \to f(X)$. More generally, any distance preserving function $f : (X, d_X) \to (Y, d_Y)$ gives a homeomorphism between X and $f(X) \subset Y$: For each pair of points z and w of $f(X)$, we have that

$$\mathrm{d}(z, w) = \mathrm{d}(f(f^{-1}(z)), f(f^{-1}(w))) = \mathrm{d}(f^{-1}(z), f^{-1}(w)).$$

Hence, $f^{-1} : f(X) \to X$ is also a distance preserving function and so is continuous by Exercise 3.14.

Exercise 3.15

Prove that the function $f : \mathbb{C} \to \mathbb{C}$ given by $f(z) = az$ is an isometry of the metric space $(\mathbb{C}, \mathrm{n})$ if and only if $|a| = 1$. Here, $\mathrm{n}(z, w) = |z - w|$, as in Section 3.3.

As the identity function of any metric space is a distance preserving homeomorphism, as the inverse of a distance preserving homeomorphism is necessarily a distance preserving homeomorphism, and as the composition of two distance preserving homeomorphisms is again a distance preserving homeomorphism, the set of all isometries of a metric space is a group. Define a *hyperbolic isometry* to be an isometry of $(\mathbb{H}, \mathrm{d}_{\mathbb{H}})$, and let $\mathrm{Isom}(\mathbb{H}, \mathrm{d}_{\mathbb{H}})$ denote the *group of isometries* of $(\mathbb{H}, \mathrm{d}_{\mathbb{H}})$. In fact, we have already encountered the group $\mathrm{Isom}(\mathbb{H}, \mathrm{d}_{\mathbb{H}})$ of isometries of the metric space $(\mathbb{H}, \mathrm{d}_{\mathbb{H}})$.

Theorem 3.19

$\mathrm{Isom}(\mathbb{H}, \mathrm{d}_{\mathbb{H}}) = \mathrm{M\ddot{o}b}(\mathbb{H})$.

Proof

By our construction of the hyperbolic metric $\mathrm{d}_{\mathbb{H}}$ on $\mathbb{H}$, specifically Proposition 3.17, we have that every element of Möb($\mathbb{H}$) is a hyperbolic isometry, and so Möb($\mathbb{H}$) $\subset$ Isom($\mathbb{H}, \mathrm{d}_{\mathbb{H}}$).

We begin the proof of the opposite inclusion with the observation that hyperbolic line segments can be characterized purely in terms of hyperbolic distance.

Proposition 3.20

Let x, y, and z be distinct points in $\mathbb{H}$. Then,

$$\mathrm{d}_{\mathbb{H}}(x, y) + \mathrm{d}_{\mathbb{H}}(y, z) = \mathrm{d}_{\mathbb{H}}(x, z)$$

if and only if y is contained in the hyperbolic line segment joining x to z.

Proof

Using Exercise 2.42, there exists an element m of Möb($\mathbb{H}$) for which $m(x) = i$ and $m(z) = \alpha\, i$, where $\alpha = \mathrm{d}_{\mathbb{H}}(i, \alpha i) = \mathrm{d}_{\mathbb{H}}(x, z)$. Write $m(y) = a + bi$. There are several cases to consider.

Suppose that y lies on the hyperbolic line segment joining x to z. Then, $m(y)$ lies on the hyperbolic line segment joining $m(x) = i$ to $m(z) = \alpha i$. In particular, $a = 0$ and $1 \leq b \leq \alpha$, and so

$$\mathrm{d}_{\mathbb{H}}(x, y) = \mathrm{d}_{\mathbb{H}}(i, bi) = \ln(b)$$

and

$$\mathrm{d}_{\mathbb{H}}(y, z) = \mathrm{d}_{\mathbb{H}}(bi, \alpha i) = \ln\left[\frac{\alpha}{b}\right] = \mathrm{d}_{\mathbb{H}}(x, z) - \ln(b).$$

Hence, $\mathrm{d}_{\mathbb{H}}(x, z) = \mathrm{d}_{\mathbb{H}}(x, y) + \mathrm{d}_{\mathbb{H}}(y, z)$.

Suppose now that y does not lie on the hyperbolic line segment joining x to z. There are two cases, namely, that $m(y)$ lies on the positive imaginary axis, so that $a = 0$, or that $m(y)$ does not lie on the positive imaginary axis, so that $a \neq 0$.

If $a = 0$, then $m(y) = bi$, where either $0 < b < 1$ or $\alpha < b$.

If $0 < b < 1$, then

$$\mathrm{d}_{\mathbb{H}}(x, y) = -\ln(b) \text{ and } \mathrm{d}_{\mathbb{H}}(y, z) = \ln\left[\frac{\alpha}{b}\right] = \mathrm{d}_{\mathbb{H}}(x, z) - \ln(b).$$

As $\ln(b) < 0$, we have that

$$\mathrm{d}_{\mathbb{H}}(x,y) + \mathrm{d}_{\mathbb{H}}(y,z) = \mathrm{d}_{\mathbb{H}}(x,z) - 2\ln(b) > \mathrm{d}_{\mathbb{H}}(x,z).$$

If $b > \alpha$, then

$$\mathrm{d}_{\mathbb{H}}(x,y) = \ln(b) \text{ and } \mathrm{d}_{\mathbb{H}}(y,z) = \ln\left[\frac{b}{\alpha}\right] = \ln(b) - \mathrm{d}_{\mathbb{H}}(x,z).$$

As $\ln(b) > \mathrm{d}_{\mathbb{H}}(x,z)$, we have that

$$\mathrm{d}_{\mathbb{H}}(x,y) + \mathrm{d}_{\mathbb{H}}(y,z) = 2\ln(b) - \mathrm{d}_{\mathbb{H}}(x,z) > \mathrm{d}_{\mathbb{H}}(x,z).$$

If $a \neq 0$, we begin with the observation that

$$\mathrm{d}_{\mathbb{H}}(i,bi) < \mathrm{d}_{\mathbb{H}}(i,a+bi) = \mathrm{d}_{\mathbb{H}}(x,y).$$

This observation follows from the argument given in Section 3.4. Specifically, let $f : [\alpha,\beta] \to \mathbb{H}$ be a distance realizing path between $i = f(\alpha)$ and $a+bi = f(\beta)$. Note that the path $g : [\alpha,\beta] \to \mathbb{H}$ given by $g(t) = \mathrm{Im}(f(t))\, i$ satisfies $g(\alpha) = i$, $g(\beta) = bi$, and $\mathrm{length}_{\mathbb{H}}(g) < \mathrm{length}_{\mathbb{H}}(f)$ because $a \neq 0$.

Similarly, we have that

$$\mathrm{d}_{\mathbb{H}}(bi,\alpha i) < \mathrm{d}_{\mathbb{H}}(a+bi,\alpha i) = \mathrm{d}_{\mathbb{H}}(y,z).$$

If $1 \leq b \leq \alpha$, then

$$\mathrm{d}_{\mathbb{H}}(x,z) = \mathrm{d}_{\mathbb{H}}(i,\alpha i) = \mathrm{d}_{\mathbb{H}}(i,bi) + \mathrm{d}_{\mathbb{H}}(bi,\alpha i) < \mathrm{d}_{\mathbb{H}}(x,y) + \mathrm{d}_{\mathbb{H}}(y,z).$$

If b does not lie in $[1,\alpha]$, then again we have two cases, namely, that $0 < b < 1$ and that $\alpha < b$.

Making use of the calculations of the previous few paragraphs, in the case in which $0 < b < 1$, we have

$$\mathrm{d}_{\mathbb{H}}(x,z) < \mathrm{d}_{\mathbb{H}}(x,z) - 2\ln(b) = \mathrm{d}_{\mathbb{H}}(i,bi) + \mathrm{d}_{\mathbb{H}}(bi,\alpha i) < \mathrm{d}_{\mathbb{H}}(x,y) + \mathrm{d}_{\mathbb{H}}(y,z),$$

whereas in the case in which $b > \alpha$, we have

$$\mathrm{d}_{\mathbb{H}}(x,z) < 2\ln(b) - \mathrm{d}_{\mathbb{H}}(x,z) = \mathrm{d}_{\mathbb{H}}(i,bi) + \mathrm{d}_{\mathbb{H}}(bi,\alpha i) < \mathrm{d}_{\mathbb{H}}(x,y) + \mathrm{d}_{\mathbb{H}}(y,z).$$

Hence, the only case in which $\mathrm{d}_{\mathbb{H}}(x,z) = \mathrm{d}_{\mathbb{H}}(x,y) + \mathrm{d}_{\mathbb{H}}(y,z)$ is the case in which y is contained in the hyperbolic line segment joining x and z. This completes the proof of Proposition 3.20. **QED**

Exercise 3.16

Prove that every hyperbolic isometry of $\mathbb{H}$ takes hyperbolic lines to hyperbolic lines.

Let f be a hyperbolic isometry, and recall that we are in the process of proving that f is an element of Möb($\mathbb{H}$). For each pair of points p and q of $\mathbb{H}$, let ℓ_{pq} denote the hyperbolic line segment joining p to q. With this notation, Proposition 3.20 implies that $\ell_{f(p)f(q)} = f(\ell_{pq})$.

Let ℓ be the perpendicular bisector of the hyperbolic line segment ℓ_{pq}, which is the hyperbolic line

$$\ell = \{z \in \mathbb{H} \mid \mathrm{d}_{\mathbb{H}}(p,z) = \mathrm{d}_{\mathbb{H}}(q,z)\}.$$

As ℓ is defined in terms of hyperbolic distance, we have that $f(\ell)$ is the perpendicular bisector of $f(\ell_{pq}) = \ell_{f(p)f(q)}$.

We now normalize the hyperbolic isometry f. Pick a pair of points x and y on the positive imaginary axis I in $\mathbb{H}$, and let H be one of the half-planes in $\mathbb{H}$ determined by I.

By the solution to Exercise 3.13, there exists an element γ of Möb($\mathbb{H}$) that satisfies $\gamma(f(x)) = x$ and $\gamma(f(y)) = y$, because $\mathrm{d}_{\mathbb{H}}(x,y) = \mathrm{d}_{\mathbb{H}}(f(x), f(y))$. In particular, we see that $\gamma \circ f$ fixes both x and y, and so $\gamma \circ f$ takes I to I. If necessary, replace γ by the composition $B \circ \gamma$ of γ with the reflection $B(z) = -\overline{z}$ in I to obtain an element γ of Möb($\mathbb{H}$) so that $\gamma \circ f$ takes I to I and takes H to H.

Let z be any point on I. As z is uniquely determined by the two hyperbolic distances $\mathrm{d}_{\mathbb{H}}(x,z)$ and $\mathrm{d}_{\mathbb{H}}(y,z)$ and as both hyperbolic distances are preserved by $\gamma \circ f$, we have that $\gamma \circ f$ fixes every point z of I.

Exercise 3.17

Let $x = \lambda i$ and $z = \mu i$ be two distinct points on the positive imaginary axis I. Let y be any point on I. Show that y is uniquely determined by the two hyperbolic distances $\mathrm{d}_{\mathbb{H}}(x,y)$ and $\mathrm{d}_{\mathbb{H}}(y,z)$.

Now, let w be any point in $\mathbb{H}$ that does not lie on I, and let ℓ be the hyperbolic line through w that is perpendicular to I. Explicitly, we can describe ℓ as the hyperbolic line contained in the Euclidean circle with Euclidean centre 0 and Euclidean radius $|w|$. Let z be the point of intersection of ℓ and I.

At this point, we know several facts about ℓ. As ℓ is the perpendicular bisector of some hyperbolic line segment in I and as $\gamma \circ f$ fixes every point of I, we have that $\gamma \circ f(\ell) = \ell$.

As $\gamma \circ f$ fixes z, as $\mathrm{d}_{\mathbb{H}}(z,w) = \mathrm{d}_{\mathbb{H}}(\gamma \circ f(z), \gamma \circ f(w)) = \mathrm{d}_{\mathbb{H}}(z, \gamma \circ f(w))$, and as $\gamma \circ f$ preserves the two half-planes determined by I, we have that $\gamma \circ f$ fixes w.

As $\gamma \circ f$ fixes every point of $\mathbb{H}$, we have that $\gamma \circ f$ is the identity. In particular, we have that $f = \gamma^{-1}$, and so f is an element of $\text{Möb}(\mathbb{H})$. This completes the proof of Theorem 3.19. **QED**

3.7 Metric Properties of $(\mathbb{H}, \mathrm{d}_{\mathbb{H}})$

In this section, we investigate some properties of the hyperbolic metric on $\mathbb{H}$.

In much the same way that we can define the hyperbolic distance between a pair of points, there is a notion of the hyperbolic distance between a pair X and Y of subsets of $\mathbb{H}$, namely,

$$\mathrm{d}_{\mathbb{H}}(X, Y) = \inf\{\mathrm{d}_{\mathbb{H}}(x, y) \mid x \in X,\ y \in Y\}.$$

As we will see later in this section, there exist distinct sets X and Y in $\mathbb{H}$ for which $\mathrm{d}_{\mathbb{H}}(X, Y) = 0$, and so this does not define a metric on the set of subsets of $\mathbb{H}$.

In general, calculating this infimum can be difficult. We spend some of this section exploring in some detail the case in which one or both of X and Y is a hyperbolic line. There is one general fact about this distance between sets that will prove to be useful.

We first need to make a definition.

Definition 3.21

A subset X of $\mathbb{H}$ is *bounded* if there exists some $C > 0$ so that X is contained in the open hyperbolic disc

$$U_C(i) = \{z \in \mathbb{H} \mid \mathrm{d}_{\mathbb{H}}(z, i) < C\}.$$

A subset X of $\mathbb{H}$ is *compact* if X is closed and bounded.

One easy example of a compact subset of $\mathbb{H}$ is any set containing a finite number of points $X = \{x_1, \ldots, x_n\}$. For any z in $\mathbb{H} - X$, set

$$\varepsilon = \inf\{\mathrm{d}_{\mathbb{H}}(z, x_1), \ldots, \mathrm{d}_{\mathbb{H}}(z, x_n)\}.$$

Then, $\varepsilon > 0$ and $U_\varepsilon(z)$ is contained in $\mathbb{H} - X$, so that $\mathbb{H} - X$ is open and hence X is closed. Also, if we set

$$C = \sup\{\mathrm{d}_{\mathbb{H}}(i, x_1), \ldots, \mathrm{d}_{\mathbb{H}}(i, x_n)\},$$

then X is contained in $U_{2C}(i)$ and so X is bounded.

Although we do not prove it here, a basic property of compact sets is that if X is a compact subset of $\mathbb{H}$ and if $\{x_n\}$ is a sequence of points of X, then there is a subsequence $\{x_{n_k}\}$ of $\{x_n\}$ so that $\{x_{n_k}\}$ converges to a point x of X. In words, a sequence of points in a compact set X contains a convergent subsequence.

Exercise 3.18

Let X be a compact subset of $\mathbb{H}$, and let Y be any subset of $\mathbb{H}$. Prove that $\mathrm{d}_{\mathbb{H}}(X, Y) > 0$ if and only if X and Y have disjoint closures.

Alhough this notion of hyperbolic distance between sets does not give a metric on the set of subsets of $\mathbb{H}$, it does give one way of measuring when two subsets of $\mathbb{H}$ are close. A particularly interesting application of this notion is to pairs of hyperbolic lines and hyperbolic rays.

Recall that there are two different types of parallelism for pairs of hyperbolic lines. Some pairs of hyperbolic lines are disjoint in $\mathbb{H}$ but the circles in $\overline{\mathbb{C}}$ containing them are not disjoint, and some pairs of hyperbolic lines are disjoint in $\mathbb{H}$ but the circles in $\overline{\mathbb{C}}$ containing them are also disjoint. We refer to the latter hyperbolic lines as *ultraparallel*.

We saw in Section 1.3 that we can distinguish these two cases by examining the endpoints at infinity of the two hyperbolic lines. Now that we have a means of measuring hyperbolic distance, we can distinguish these two cases intrinsically as well.

Let ℓ_0 and ℓ_1 be parallel hyperbolic lines that share an endpoint at infinity at the point x of $\overline{\mathbb{R}}$. Let y_k be the other endpoint at infinity of ℓ_k. As by Proposition 2.30 we have that Möb($\mathbb{H}$) acts triply transitively on $\overline{\mathbb{R}}$, we may assume that $x = \infty$, that $y_0 = 0$, and that $y_1 = 1$.

We now calculate. Each point of ℓ_0 has the form λi for some $\lambda > 0$, and each point of ℓ_1 has the form $1 + \lambda i$ for some $\lambda > 0$.

The path $f : [0, 1] \to \mathbb{H}$ given by $f(t) = t + \lambda i$ parametrizes the horizontal Euclidean line segment joining λi and $1 + \lambda i$, and so

$$\mathrm{d}_{\mathbb{H}}(\ell_0, \ell_1) \leq \mathrm{d}_{\mathbb{H}}(\lambda i, 1 + \lambda i) \leq \mathrm{length}_{\mathbb{H}}(f) = \int_0^1 \frac{1}{\lambda}\, \mathrm{d}t = \frac{1}{\lambda}$$

for every $\lambda > 0$. Letting λ tend to ∞, we see that

$$\mathrm{d}_{\mathbb{H}}(\ell_0, \ell_1) = 0$$

for two parallel hyperbolic lines ℓ_0 and ℓ_1 that share an endpoint at infinity.

Suppose, on the other hand, that ℓ_0 and ℓ_1 are ultraparallel hyperbolic lines.

Proposition 3.22

Let ℓ_0 and ℓ_1 be ultraparallel hyperbolic lines in $\mathbb{H}$. Then, $\mathrm{d}_{\mathbb{H}}(\ell_0, \ell_1) > 0$.

Proof

Again, by making use of the triple transitivity of $\mathrm{M\ddot{o}b}(\mathbb{H})$ on $\overline{\mathbb{R}}$, we may assume that the endpoints at infinity of ℓ_0 are 0 and ∞, and that the endpoints at infinity of ℓ_1 are 1 and $x > 1$. We wish to calculate the hyperbolic distance $\mathrm{d}_{\mathbb{H}}(\ell_0, \ell_1)$ between ℓ_0 and ℓ_1.

We make use of the following fact.

Exercise 3.19

Let ℓ be a hyperbolic line, and let p be a point of $\mathbb{H}$ not on ℓ. Prove that there exists a unique point z on ℓ so that the hyperbolic line segment through z and p is perpendicular to ℓ, and so that

$$\mathrm{d}_{\mathbb{H}}(p, \ell) = \mathrm{d}_{\mathbb{H}}(p, z).$$

For each $r > 0$, let c_r be the hyperbolic line contained in the Euclidean circle with Euclidean centre 0 and Euclidean radius r, so that c_r is perpendicular to ℓ_0 for every r. Note that c_r intersects ℓ_1 only for $1 < r < x$. Write the point of intersection of c_r and ℓ_1 as $re^{i\theta}$, where $0 < \theta < \frac{\pi}{2}$.

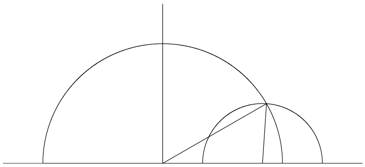

Figure 3.1: The Euclidean triangle in $\mathbb{H}$ with vertices 0, $\frac{1}{2}(x+1)$, and $re^{i\theta}$.

We can determine θ by considering the Euclidean triangle with vertices 0, $\frac{1}{2}(x+1)$ (the Euclidean centre of the Euclidean circle containing ℓ_1), and $re^{i\theta}$. See Figure 3.1.

The Euclidean lengths of the two sides of this Euclidean triangle adjacent to the vertex 0, which has angle θ, are r and $\frac{1}{2}(x+1)$, and the length of the opposite side is $\frac{1}{2}(x-1)$. Calculating, we see that

$$\left[\frac{1}{2}(x-1)\right]^2 = \left[\frac{1}{2}(x+1)\right]^2 + r^2 - 2r\left[\frac{1}{2}(x+1)\right]\cos(\theta)$$

by the law of cosines.

Simplifying, we see that this is equivalent to

$$x + r^2 = r(x+1)\cos(\theta),$$

and so

$$\cos(\theta) = \frac{x+r^2}{r(x+1)}$$

and

$$\sin(\theta) = \sqrt{1-\cos^2(\theta)} = \frac{\sqrt{(r^2-1)(x^2-r^2)}}{r(x+1)}.$$

The hyperbolic distance between ri and $re^{i\theta}$ for this value of θ is the length of the hyperbolic line segment joining ri and $re^{i\theta}$. Parametrizing this hyperbolic line segment by $f_r(t) = re^{it}$ for $\theta \le t \le \frac{\pi}{2}$, we calculate that

$$\text{length}_{\mathbb{H}}(f_r) = \int_\theta^{\frac{\pi}{2}} \frac{1}{\sin(t)}\,\mathrm{d}t = -\ln|\csc(\theta) - \cot(\theta)| = \frac{1}{2}\ln\left[\frac{(r+1)(x+r)}{(r-1)(x-r)}\right].$$

As c_r is perpendicular to ℓ_0, we know by Exercise 3.19 that

$$\mathrm{d}_{\mathbb{H}}(re^{i\theta}, ri) = \mathrm{d}_{\mathbb{H}}(re^{i\theta}, \ell_0).$$

In particular, the hyperbolic distance between the two hyperbolic lines ℓ_0 and ℓ_1 is the minimum hyperbolic distance between $re^{i\theta}$ and ri as r varies over the interval $(1, x)$.

The hyperbolic distance between $re^{i\theta}$ and ri is minimized when

$$\frac{\mathrm{d}}{\mathrm{d}r}\ln\left[\frac{(r+1)(x+r)}{(r-1)(x-r)}\right]=\frac{2(r^2-x)(x+1)}{(r+1)(x+r)(r-1)(x-r)}=0.$$

As $r>0$ and $x>1$, this can only occur when $r=\sqrt{x}$.

Hence, the hyperbolic distance between the two hyperbolic lines ℓ_0 and ℓ_1 is

$$\mathrm{d}_{\mathbb{H}}(\ell_0,\ell_1)=\frac{1}{2}\ln\left[\frac{(\sqrt{x}+1)(x+\sqrt{x})}{(\sqrt{x}-1)(x-\sqrt{x})}\right]=\ln\left[\frac{\sqrt{x}+1}{\sqrt{x}-1}\right],$$

which is positive because $x>1$. This completes the proof of Proposition 3.22.
QED

One consequence of the proof of Proposition 3.22 is that it also shows that there exists a unique common perpendicular for any pair of ultraparallel hyperbolic lines.

Proposition 3.23

Let ℓ_0 and ℓ_1 be two ultraparallel hyperbolic lines. Then, there exists a unique hyperbolic line ℓ that is perpendicular to both ℓ_0 and ℓ_1.

Proof

We use the same notation and normalizations as in the proof of Proposition 3.22. We know that c_r is perpendicular to ℓ_0 for all values of r by construction.

To determine the values of r for which c_r is perpendicular to ℓ_1, we apply the Pythagorean theorem to the Euclidean triangle with vertices 0, $\frac{1}{2}(x+1)$, and $re^{i\theta}$. The angle between c_r and ℓ_1 is $\frac{\pi}{2}$ if and only if

$$\left[\frac{1}{2}(x+1)\right]^2=\left[\frac{1}{2}(x-1)\right]^2+r^2,$$

which occurs if and only if $r=\sqrt{x}$. This completes the proof of Proposition 3.23.
QED

Exercise 3.20

Let I be the positive imaginary axis in $\mathbb{H}$. For a positive real number $\varepsilon>0$, let W_ε be the set of points in $\mathbb{H}$ whose hyperbolic distance from I is equal to ε. Prove that W_ε is the union of two Euclidean rays from 0 that make equal angle θ with I.

Exercise 3.21

Prove that if ℓ_0 and ℓ_1 are hyperbolic lines that share an endpoint at infinity, then there does not exist a hyperbolic line perpendicular to both ℓ_0 and ℓ_1.

Exercise 3.22

Let ℓ_0 and ℓ_1 be ultraparallel hyperbolic lines in $\mathbb{H}$. Label the endpoints at infinity of ℓ_0 as z_0 and z_1, and the endpoints at infinity of ℓ_1 as w_0 and w_1, so that they occur in the order z_0, w_0, w_1, z_1 moving counterclockwise around $\overline{\mathbb{R}}$. Prove that

$$\tanh^2\left[\frac{1}{2}\mathrm{d}_{\mathbb{H}}(\ell_0, \ell_1)\right] = \frac{1}{1 - [z_0, w_0; w_1, z_1]}.$$

Although we will not explore it in detail, we do note here that this notion of distance between sets can be used to give a description of the boundary at infinity of $\mathbb{H}$ that is intrinsic to $\mathbb{H}$ and that does not make use of how $\mathbb{H}$ sits as a subset of $\overline{\mathbb{C}}$.

Let $\mathcal{R}$ be the set of all hyperbolic rays in $\mathbb{H}$. For each ray R in $\mathcal{R}$, let $\mathrm{sub}(R)$ be the set of all subrays of R, which are the hyperbolic rays contained in R. Given any two rays R_1 and R_2 in $\mathcal{R}$, say that $R_1 \sim R_2$ if and only if

$$\sup\{\mathrm{d}_{\mathbb{H}}(R_1^0, R_2^0) \mid R_1^0 \in \mathrm{sub}(R_1), R_2^0 \in \mathrm{sub}(R_2)\} = 0.$$

Note that if two nonequal hyperbolic rays R_1 and R_2 have the same initial point in $\mathbb{H}$, then this supremum is infinite, and so $R_1 \not\sim R_2$. In fact, for any two hyperbolic rays, this supremum is either 0 or infinite, and is 0 if and only if the two rays have the same endpoint at infinity.

This equivalence gives a way of identifying the boundary at infinity $\overline{\mathbb{R}}$ of $\mathbb{H}$ with equivalence classes in $\mathcal{R}$. Morever, as elements of $\mathrm{M\ddot{o}b}(\mathbb{H})$ take hyperbolic rays to hyperbolic rays and preserve hyperbolic distance, we see that $\mathrm{M\ddot{o}b}(\mathbb{H})$ preserves this equivalence relation, and so we get an action of $\mathrm{M\ddot{o}b}(\mathbb{H})$ on $\mathcal{R}/\sim$.

So, fix a point x_0 in $\mathbb{H}$ and consider the collection of all hyperbolic rays emanating from x_0. As no two of these hyperbolic rays are equivalent (as noted above) and as every point on the boundary at infinity $\overline{\mathbb{R}}$ of $\mathbb{H}$ is the endpoint at infinity of a hyperbolic ray from x_0, we can identify $\overline{\mathbb{R}}$ with the collection of hyperbolic rays from x_0. The collection of hyperbolic rays from x_0 can be parametrized by the unit circle $\mathbb{S}^1$, as follows: Fix a hyperbolic ray r_0 from x_0. For any other hyperbolic ray r from x_0, let θ_r be the angle between r_0 and r, measured counterclockwise, and then identify the hyperbolic ray r with the point $e^{i\theta_r}$ on $\mathbb{S}^1$. We note here that if we choose a different base hyperbolic ray r_0' at x_0, then the parametrizations of the hyperbolic rays from x_0 with these two different base hyperbolic rays differ by a rotation of $\mathbb{S}^1$ by the angle between r_0 and r_0'.

This construction gives us a way of determining the size of a set X in $\mathbb{H} \cup \overline{\mathbb{R}}$ as viewed from x_0, by considering the proportion of hyperbolic rays from x_0 that either pass through X or that have their endpoint at infinity in X. We refer to this measure of size as the *visual measure* of X from x_0. The visual measures of a fixed set X when viewed from distinct points of $\mathbb{H}$ will usually be different.

As an example, consider the interval $X = [0, 1]$ in $\overline{\mathbb{R}}$. The visual measure of X when viewed from $x_0 = i$ is $\frac{1}{4}$: The hyperbolic rays from i to 0 and to 1 meet perpendicularly at i, and the hyperbolic rays from i that have an endpoint at infinity in X lie between these two hyperbolic rays. On the other hand, the visual measure of X when viewed from $10i$ is 0.0317.

We can see this as follows. The equation of the Euclidean circle A passing through 1 and $10i$ and perpendicular to the real axis $\mathbb{R}$ is

$$\left(x + \frac{99}{2}\right)^2 + y^2 = \left(\frac{101}{2}\right)^2.$$

By construction, the point of intersection of A and the positive imaginary axis I in $\mathbb{H}$ is $10i$. The slope of the tangent line to A at the point of intersection $10i$ is $-\frac{99}{20}$ (obtained by implicitly differentiating the equation of A), and so the angle between A and I (as discussed in Section 2.7) is

$$\arctan\left(-\frac{99}{20}\right) - \frac{\pi}{2} = -2.9422 = 0.1993 \ \text{(modulo } \pi).$$

Hence, the visual measure of the interval $[0, 1]$ from $10i$ is $\frac{0.1993}{2\pi} = 0.0317$. As we would expect from drawing the picture, the visual measure of $[0, 1]$ from $10i$ is small.

For a general point λi, the equation of the Euclidean circle A_λ passing through 1 and λi is

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

The slope of the tangent line to A_λ at λi is $\frac{1-\lambda^2}{2\lambda}$, and so the angle between A_λ and I is

$$\arctan\left(\frac{1 - \lambda^2}{2\lambda}\right) - \frac{\pi}{2} \ \text{(modulo } \pi).$$

(When $\lambda = 1$, this result agrees with our earlier calculation.)

Exercise 3.23

Let ℓ be a hyperbolic line in $\mathbb{H}$, and let p be a point in $\mathbb{H}$ not on ℓ. Determine the proportion of the hyperbolic rays from p that intersect ℓ, that is, the visual measure of ℓ when viewed from p.

4
Planar Models of the Hyperbolic Plane

Up to this point, we have focused our attention exclusively on the upper half-plane model $\mathbb{H}$ of the hyperbolic plane, but there are many other useful models. In this chapter, we explore a second particular model, the *Poincaré disc model* $\mathbb{D}$, of the hyperbolic plane, which we construct starting from the upper half-plane model. We go on to show that the construction used for the Poincaré disc model is but one instance of a *general construction* using techniques from complex analysis for producing planar models of the hyperbolic plane.

4.1 The Poincaré Disc Model

Up to this point, we have focused our attention on developing the upper half-plane model $\mathbb{H}$ of the hyperbolic plane and on studying its properties. There are a number of other models of the hyperbolic plane. One of the most useful of these other models, at least for our purposes, is the *Poincaré disc* model $\mathbb{D}$.

There are a number of ways we could develop this, and other, models of the hyperbolic plane. One way is to retrace the steps we undertook to develop the upper half-plane model. However, for developing the Poincaré disc model and related planar models of the hyperbolic plane, this process is inefficient. Another way is to make use of the work we have already done in developing the upper half-plane model, and find a way of transferring this work to the other model. We take this latter approach.

The underlying space of the Poincaré disc model of the hyperbolic plane is the open unit disc

$$\mathbb{D} = \{z \in \mathbb{C} \mid |z| < 1\}$$

in the complex plane $\mathbb{C}$. As $\mathbb{H}$ and $\mathbb{D}$ are both discs in the Riemann sphere $\overline{\mathbb{C}}$, we know from Theorem 2.11 that there exist many elements m of Möb taking $\mathbb{D}$ to $\mathbb{H}$. (In fact, in Exercise 2.10, you constructed an explicit element of Möb taking $\mathbb{D}$ to $\mathbb{H}$.) We now show how to use an element $m : \mathbb{D} \to \mathbb{H}$ of Möb to transport hyperbolic geometry from $\mathbb{H}$ to $\mathbb{D}$.

To start, define a *hyperbolic line in* $\mathbb{D}$ to be the image under m^{-1} of a hyperbolic line in $\mathbb{H}$. We know that every hyperbolic line in $\mathbb{H}$ is contained in a circle in $\overline{\mathbb{C}}$ perpendicular to $\overline{\mathbb{R}}$, that every element of Möb takes circles in $\overline{\mathbb{C}}$ to circles in $\overline{\mathbb{C}}$, and that every element of Möb preserves the angle between circles in $\overline{\mathbb{C}}$. Hence, every hyperbolic line in $\mathbb{D}$ is the intersection of $\mathbb{D}$ with a circle in $\overline{\mathbb{C}}$ perpendicular to the unit circle $\mathbb{S}^1$ bounding $\mathbb{D}$, and every such intersection is a hyperbolic line in $\mathbb{D}$. Note that this definition of a hyperbolic line in $\mathbb{D}$ is independent of the choice of element of Möb taking $\mathbb{D}$ to $\mathbb{H}$. (We leave this argument for the interested reader.)

We can make this observation concrete. Consider the element

$$\xi(z) = \frac{\frac{i}{\sqrt{2}}z + \frac{1}{\sqrt{2}}}{-\frac{1}{\sqrt{2}}z - \frac{i}{\sqrt{2}}}$$

of Möb$^+$ taking $\mathbb{D}$ to $\mathbb{H}$. A hyperbolic line in $\mathbb{D}$, defined as the image under ξ^{-1} of a hyperbolic line in $\mathbb{H}$, then has one of two possible equations. If the hyperbolic line ℓ in $\mathbb{H}$ has the equation $\ell = \{z \in \mathbb{H} \mid \mathrm{Re}(z) = c\}$ for $c \in \mathbb{R}$, then $\xi^{-1}(\ell)$ has the equation $\ell_c = \{z \in \mathbb{D} \mid \mathrm{Re}(\xi(z)) = c\}$, which we can calculate; namely,

$$\mathrm{Re}(\xi(z)) = \frac{-\mathrm{Re}(z)}{\frac{1}{2}z\overline{z} + \frac{1}{2} + \frac{i}{2}(\overline{z} - z)}.$$

For $c = 0$, the hyperbolic line ℓ_c in $\mathbb{D}$ is

$$\ell_c = \{z \in \mathbb{D} \mid \mathrm{Re}(z) = 0\},$$

whereas for $c \neq 0$, the hyperbolic line ℓ_c in $\mathbb{D}$ is

$$\ell_c = \left\{ z \in \mathbb{D} \,\middle|\, \left| z - \left(-\frac{1}{c} - i \right) \right| = \frac{1}{c} \right\}.$$

Using Exercise 1.2, it is easy to verify that the Euclidean circle containing ℓ_c intersects the unit circle $\mathbb{S}^1$ perpendicularly.

If the hyperbolic line ℓ in $\mathbb{H}$ has the equation $\ell = \{z \in \mathbb{H} \mid |z-c|^2 = r^2\}$ for $c \in \mathbb{R}$, $r > 0$, then $\xi^{-1}(\ell)$ has the equation $\ell_{c,r} = \{z \in \mathbb{D} \mid |\xi(z) - c|^2 = r^2\}$, which we can calculate; simplifying, we see that the equation $|\xi(z) - c|^2 = r^2$ becomes the equation

$$\left| z - \left(\frac{-2c + i(1 + r^2 - c^2)}{1 + c^2 - r^2} \right) \right|^2 = \frac{4r^2}{(1 + c^2 - r^2)^2}.$$

Again using Exercise 1.2, it is not difficult, although slightly messy, to verify that the Euclidean circle determined by the equation $|\xi(z) - c|^2 = r^2$ is perpendicular to $\mathbb{S}^1$. If we were to repeat this calculation with a different choice of $\xi(z)$, we would get a seemingly different set of equations, but as the parameters c and r range over all possible values, the two sets of equations would describe the same set of hyperbolic lines.

A picture of some hyperbolic lines in $\mathbb{D}$ is given in Figure 4.1. Note that this picture of the Poincaré disc model of the hyperbolic plane is (very) vaguely reminiscent of some of the drawings of M. C. Escher. The interested reader is directed to the books of Schattschneider [30] and Locher [25] for more information about the work of Escher.

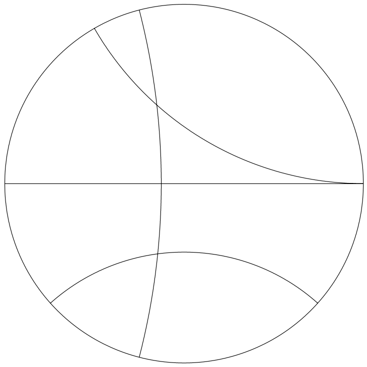

Figure 4.1: Some hyperbolic lines in $\mathbb{D}$

Let m be any element of Möb taking $\mathbb{D}$ to $\mathbb{H}$. Then, every element q of Möb$(\mathbb{D})$ has the form $q = m^{-1} \circ p \circ m$, where p is an element of Möb$(\mathbb{H})$. In particular, the action of Möb$(\mathbb{D})$ on $\mathbb{D}$ inherits all transitivity properties that Möb$(\mathbb{H})$ has for its action of $\mathbb{H}$.

In fact, in Exercise 2.39, we saw that every element of Möb$(\mathbb{D})$ has either the form

$$p(z) = \frac{\alpha z + \beta}{\overline{\beta} z + \overline{\alpha}}$$

or the form

$$p(z) = \frac{\alpha\overline{z} + \beta}{\overline{\beta}\overline{z} + \overline{\alpha}},$$

where α, $\beta \in \mathbb{C}$ and $|\alpha|^2 - |\beta|^2 = 1$. The Möbius transformations taking $\mathbb{D}$ to $\mathbb{D}$ are the elements of

$$\text{Möb}^+(\mathbb{D}) = \text{Möb}^+ \cap \text{Möb}(\mathbb{D}),$$

which are those elements of Möb($\mathbb{D}$) of the form

$$p(z) = \frac{\alpha z + \beta}{\overline{\beta} z + \overline{\alpha}}.$$

To transfer the hyperbolic element of arc-length from $\mathbb{H}$ to $\mathbb{D}$, we begin with an explicit element ξ of Möb$^+$ taking $\mathbb{D}$ to $\mathbb{H}$. The element we use here is the one we used above, namely,

$$\xi(z) = \frac{\frac{i}{\sqrt{2}} z + \frac{1}{\sqrt{2}}}{-\frac{1}{\sqrt{2}} z - \frac{i}{\sqrt{2}}}.$$

We transfer the hyperbolic element of arc-length from $\mathbb{H}$ to $\mathbb{D}$ by making the following observation: For any piecewise C^1 path $f : [a, b] \to \mathbb{D}$, the composition $\xi \circ f : [a, b] \to \mathbb{H}$ is a piecewise C^1 path into $\mathbb{H}$. We know how to calculate the hyperbolic length of $\xi \circ f$, namely, by integrating the hyperbolic element of arc-length $\frac{1}{\text{Im}(z)}|\text{d}z|$ on $\mathbb{H}$ along $\xi \circ f$. So, define the hyperbolic length of f in $\mathbb{D}$ by

$$\text{length}_{\mathbb{D}}(f) = \text{length}_{\mathbb{H}}(\xi \circ f).$$

Theorem 4.1

The hyperbolic length of a piecewise C^1 path $f : [a, b] \to \mathbb{D}$ is given by the integral

$$\text{length}_{\mathbb{D}}(f) = \int_f \frac{2}{1 - |z|^2}\, |\text{d}z|.$$

Proof

The proof of Theorem 4.1 consists of several parts. We begin by deriving the form of the hyperbolic element of arc-length on $\mathbb{D}$. We then show that this hyperbolic element of arc-length is independent of the choice of the element of Möb taking $\mathbb{D}$ to $\mathbb{H}$.

We are given that the hyperbolic length of a piecewise C^1 path $f : [a, b] \to \mathbb{D}$ is given by

$$\begin{aligned} \text{length}_{\mathbb{D}}(f) = \text{length}_{\mathbb{H}}(\xi \circ f) &= \int_{\xi \circ f} \frac{1}{\text{Im}(z)} |\mathrm{d}z| \\ &= \int_a^b \frac{1}{\text{Im}((\xi \circ f)(t))} \, |(\xi \circ f)'(t)| \, \mathrm{d}t \\ &= \int_a^b \frac{1}{\text{Im}(\xi(f(t)))} \, |\xi'(f(t))| \, |f'(t)| \, \mathrm{d}t \\ &= \int_f \frac{1}{\text{Im}(\xi(z))} |\xi'(z)| |\mathrm{d}z|. \end{aligned}$$

Calculating, we see that

$$\text{Im}(\xi(z)) = \text{Im}\left(\frac{\frac{i}{\sqrt{2}} z + \frac{1}{\sqrt{2}}}{-\frac{1}{\sqrt{2}} z - \frac{i}{\sqrt{2}}} \right) = \frac{1 - |z|^2}{| - z - i|^2}$$

and that

$$|\xi'(z)| = \frac{2}{|z + i|^2},$$

and so

$$\frac{1}{\text{Im}(\xi(z))} |\xi'(z)| = \frac{2}{1 - |z|^2}.$$

Hence,

$$\text{length}_{\mathbb{D}}(f) = \int_f \frac{2}{1 - |z|^2} |\mathrm{d}z|.$$

We now need to show that this hyperbolic element of arc-length $\frac{2}{1-|z|^2}|\mathrm{d}z|$ on $\mathbb{D}$ is independent of the choice of ξ. So, let $f : [a, b] \to \mathbb{D}$ be a piecewise C^1 path, and let p be any element of Möb taking $\mathbb{D}$ to $\mathbb{H}$. As $p \circ \xi^{-1}$ is an element of Möb and takes $\mathbb{H}$ to $\mathbb{H}$, we have that $q = p \circ \xi^{-1} \in$ Möb($\mathbb{H}$).

As $\xi \circ f$ is a piecewise C^1 path in $\mathbb{H}$, the invariance of hyperbolic length calculated with respect to the element of arc-length $\frac{1}{\text{Im}(z)}|\mathrm{d}z|$ on $\mathbb{H}$ under Möb($\mathbb{H}$) immediately implies that

$$\text{length}_{\mathbb{H}}(\xi \circ f) = \text{length}_{\mathbb{H}}(q \circ \xi \circ f) = \text{length}_{\mathbb{H}}(p \circ f).$$

This last equality follows from $q \circ \xi = p \circ \xi^{-1} \circ \xi = p$. Hence, $\text{length}_{\mathbb{D}}(f)$ is well defined. This completes the proof of Theorem 4.1. **QED**

As an example calculation, let $0 < r < 1$ and consider the piecewise C^1 path $f : [0, r] \to \mathbb{D}$ given by $f(t) = t$. Then,

$$\begin{aligned}
\text{length}_{\mathbb{D}}(f) &= \int_f \frac{2}{1-|z|^2}\,|\mathrm{d}z| \\
&= \int_0^r \frac{2}{1-t^2}\,\mathrm{d}t \\
&= \int_0^r \left[\frac{1}{1+t} + \frac{1}{1-t}\right]\mathrm{d}t \\
&= \ln\left[\frac{1+r}{1-r}\right].
\end{aligned}$$

Exercise 4.1

Let m be an element of Möb taking $\mathbb{H}$ to $\mathbb{D}$, and let $f : [a, b] \to \mathbb{H}$ be a piecewise C^1 path. Show that $\text{length}_{\mathbb{D}}(m \circ f) = \text{length}_{\mathbb{H}}(f)$.

We now use hyperbolic lengths of piecewise C^1 paths in $\mathbb{D}$ to define hyperbolic distance in $\mathbb{D}$. Given points x and y in $\mathbb{D}$, let $\Theta[x, y]$ be the set of all piecewise C^1 paths $f : [a, b] \to \mathbb{D}$ with $f(a) = x$ and $f(b) = y$, and define

$$\mathrm{d}_{\mathbb{D}}(x, y) = \inf\{\text{length}_{\mathbb{D}}(f) \mid f \in \Theta[x, y]\}.$$

Proposition 4.2

$(\mathbb{D}, \mathrm{d}_{\mathbb{D}})$ is a path metric space with $\text{Isom}(\mathbb{D}, \mathrm{d}_{\mathbb{D}}) = \text{Möb}(\mathbb{D})$. Moreover, a distance-realizing path between two points x and y of $\mathbb{D}$ is an almost simple parametrization of the hyperbolic line segment joining x to y.

Proof

Let m be any element of Möb taking $\mathbb{H}$ to $\mathbb{D}$. The first step of the proof of Proposition 4.2 is to show that m is distance-preserving.

As in Section 3.4, let $\Gamma[z, w]$ be the set of all piecewise C^1 paths $f : [a, b] \to \mathbb{H}$ with $f(a) = z$ and $f(b) = w$. For each pair of points z and w of $\mathbb{H}$, we have that

$$\begin{aligned}\mathrm{d}_{\mathbb{H}}(z,w) &= \inf\{\mathrm{length}_{\mathbb{H}}(f) \mid f \in \Gamma[z,w]\} \\ &= \inf\{\mathrm{length}_{\mathbb{D}}(m \circ f) \mid f \in \Gamma[z,w]\} \\ &\leq \inf\{\mathrm{length}_{\mathbb{D}}(g) \mid g \in \Theta[m(z),m(w)]\} \\ &\leq \mathrm{d}_{\mathbb{D}}(m(z),m(w)).\end{aligned}$$

Similarly, if x and y are points of $\mathbb{D}$, write $x = m(z)$ and $y = m(w)$ for points z and w of $\mathbb{H}$. Calculating, we see that

$$\begin{aligned}\mathrm{d}_{\mathbb{D}}(m(z),m(w)) = \mathrm{d}_{\mathbb{D}}(x,y) &= \inf\{\mathrm{length}_{\mathbb{D}}(f) \mid f \in \Theta[x,y]\} \\ &= \inf\{\mathrm{length}_{\mathbb{H}}(m^{-1} \circ f) \mid f \in \Theta[x,y]\} \\ &\leq \inf\{\mathrm{length}_{\mathbb{H}}(g) \mid g \in \Gamma[z,w]\} \\ &\leq \mathrm{d}_{\mathbb{H}}(z,w).\end{aligned}$$

As $\mathrm{d}_{\mathbb{H}}(z,w) = \mathrm{d}_{\mathbb{D}}(m(z),m(w))$ for all z, $w \in \mathbb{H}$ and all $m \in$ Möb taking $\mathbb{H}$ to $\mathbb{D}$, and as $\mathrm{d}_{\mathbb{H}}$ is a metric on $\mathbb{H}$, we have that $\mathrm{d}_{\mathbb{D}}$ is a metric on $\mathbb{D}$. Moreover, this argument shows that m is a distance-preserving homeomorphism between $(\mathbb{H}, \mathrm{d}_{\mathbb{H}})$ and $(\mathbb{D}, \mathrm{d}_{\mathbb{D}})$.

Let x and y be two points of $\mathbb{D}$, let $z = m^{-1}(x)$ and $w = m^{-1}(y)$ be the corresponding points of $\mathbb{H}$, and let $f : [a,b] \to \mathbb{H}$ be a piecewise C^1 path with $f(a) = z$, $f(b) = w$, and $\mathrm{length}_{\mathbb{H}}(f) = \mathrm{d}_{\mathbb{H}}(z,w)$. Note that by Theorem 3.16, f is an almost simple parametrization of the hyperbolic line segment joining z and w.

As m is a distance-preserving homeomorphism between $\mathbb{H}$ and $\mathbb{D}$, there necessarily exists a path in $\Theta[x,y]$ realizing the hyperbolic distance $\mathrm{d}_{\mathbb{D}}(x,y)$, namely, $m \circ f$. Moreover, as f is an almost simple parametrization of the hyperbolic line segment in $\mathbb{H}$ between z and w, and as m takes hyperbolic lines in $\mathbb{H}$ to hyperbolic lines in $\mathbb{D}$, we see that $m \circ f$ is an almost simple parametrization of the hyperbolic line segment in $\mathbb{D}$ between x and y.

Conversely, if $g : [c,d] \to \mathbb{D}$ is a distance-realizing path joining x and y, then $m^{-1} \circ g : [c,d] \to \mathbb{H}$ is a distance-realizing path joining z and w, and hence it is an almost simple parametrization of the hyperbolic line segment joining z and w. Therefore, $g = m \circ (m^{-1} \circ g)$ is an almost simple parametization of the hyperbolic line segment joining x and y. That is, in $\mathbb{D}$ as in $\mathbb{H}$, the distance-realizing paths between two points are exactly the almost simple parametrizations of the hyperbolic line segment joining the two points.

The fact that Möb($\mathbb{D}$) is exactly the group of isometries of $(\mathbb{D}, \mathrm{d}_{\mathbb{D}})$ follows from the fact that Möb($\mathbb{H}$) is exactly the group of isometries of $(\mathbb{H}, \mathrm{d}_{\mathbb{H}})$, by Theorem

3.19, and that any element m of Möb taking $\mathbb{H}$ to $\mathbb{D}$ is a distance-preserving homeomorphism and, hence, an isometry.

Specifically, if g is an isometry of $(\mathbb{D}, \mathrm{d}_{\mathbb{D}})$, then $m^{-1} \circ g \circ m$ is an isometry of $(\mathbb{H}, \mathrm{d}_{\mathbb{H}})$. By Theorem 3.19, we have that $m^{-1} \circ g \circ m$ is an element of Möb($\mathbb{H}$), and hence, g is an element of Möb($\mathbb{D}$). Conversely, if g is an element of Möb($\mathbb{D}$), then $m^{-1} \circ g \circ m$ is an element of Möb($\mathbb{H}$), and hence, it is an isometry of $(\mathbb{H}, \mathrm{d}_{\mathbb{H}})$. As m and m^{-1} are distance-preserving, we have that g is an isometry of $(\mathbb{D}, \mathrm{d}_{\mathbb{D}})$. This completes the proof of Proposition 4.2. **QED**

Exercise 4.2

For $0 < r < 1$, show that

$$\mathrm{d}_{\mathbb{D}}(0, r) = \ln \left[\frac{1+r}{1-r} \right]$$

and, hence, that

$$r = \tanh \left[\frac{1}{2} \mathrm{d}_{\mathbb{D}}(0, r) \right].$$

We note here that, analogously to the upper half-plane $\mathbb{H}$, the *boundary at infinity* of the Poincaré disc $\mathbb{D}$ is the unit circle $\mathbb{S}^1$ in $\mathbb{C}$, which is the circle in $\overline{\mathbb{C}}$ determining $\mathbb{D}$. As with the boundary at infinity $\overline{\mathbb{R}}$ of $\mathbb{H}$, the hyperbolic distance between any point of $\mathbb{S}^1$ and any point of $\mathbb{D}$ is infinite.

One difficulty with the upper half-plane model $\mathbb{H}$ of the hyperbolic plane is that no easily expressed relationship connects the Euclidean distance $|z - w|$ to the hyperbolic distance $\mathrm{d}_{\mathbb{H}}(z, w)$ between a given pair of points. One useful feature of the Poincaré disc model $\mathbb{D}$ is that there does exist such an easily expressed relationship between the Euclidean and hyperbolic distances between a pair of points of $\mathbb{D}$.

We find this relationship by considering functions on $\mathbb{D}$ that are invariant under Möb($\mathbb{D}$), which is reminiscent of the discussion in Section 2.3. Say that a function $g : \mathbb{D} \times \mathbb{D} \to \mathbb{R}$ is *invariant under the action of* Möb($\mathbb{D}$) if for each point (x, y) of $\mathbb{D} \times \mathbb{D}$ and for each element p of Möb($\mathbb{D}$), we have that $g(x, y) = g(p(x), p(y))$.

We already know one such function, namely, the hyperbolic distance $\mathrm{d}_{\mathbb{D}}$. Consequently, for any function $h : [0, \infty) \to \mathbb{R}$, the composition $\varphi = h \circ \mathrm{d}_{\mathbb{D}}$ is invariant under Möb($\mathbb{D}$). Let us try and construct an explicit example.

To begin with, the invariance of hyperbolic lengths of paths in $\mathbb{D}$ under the action of $\text{Möb}^+(\mathbb{D})$ gives that

$$\begin{aligned}
\int_f \frac{2}{1-|z|^2}|\mathrm{d}z| &= \int_a^b \frac{2}{1-|f(t)|^2}|f'(t)|\mathrm{d}t \\
&= \int_a^b \frac{2}{1-|(p\circ f)(t)|^2}|(p\circ f)'(t)|\mathrm{d}t \\
&= \int_a^b \frac{2}{1-|p(f(t))|^2}\,|p'(f(t))|\,|f'(t)|\mathrm{d}t \\
&= \int_f \frac{2|p'(z)|}{1-|p(z)|^2}|\mathrm{d}z| = \int_{p\circ f} \frac{2}{1-|z|^2}|\mathrm{d}z|
\end{aligned}$$

for every piecewise C^1 path $f : [a,b] \to \mathbb{D}$ and every element p of $\text{Möb}^+(\mathbb{D})$. (We have restricted our consideration to an element $p(z)$ of $\text{Möb}^+(\mathbb{D})$ in this argument, as we need to calculate its derivative $p'(z)$.)

As this holds for every piecewise C^1 path $f : [a,b] \to \mathbb{D}$, we may use Lemma 3.10 to conclude that

$$\frac{2}{1-|z|^2} = \frac{2|p'(z)|}{1-|p(z)|^2}$$

for every element p of $\text{Möb}^+(\mathbb{D})$.

We now pause to calculate that

$$(p(x)-p(y))^2 = p'(x)p'(y)(x-y)^2$$

for every element p of $\text{Möb}^+(\mathbb{D})$ and every pair x and y of points of $\mathbb{D}$: Write

$$p(z) = \frac{\alpha z+\beta}{\overline{\beta}z+\overline{\alpha}},$$

where α, $\beta \in \mathbb{C}$ and $|\alpha|^2 - |\beta|^2 = 1$. Then

$$p(z) - p(w) = \frac{z-w}{(\overline{\beta}z+\overline{\alpha})(\overline{\beta}w+\overline{\alpha})}$$

and

$$p'(z) = \frac{1}{(\overline{\beta}z+\overline{\alpha})^2}.$$

Combining these calculations, we can see that

$$\begin{aligned}
\frac{|x-y|^2}{(1-|x|^2)(1-|y|^2)} &= |x-y|^2\left(\frac{|p'(x)|}{1-|p(x)|^2}\right)\left(\frac{|p'(y)|}{1-|p(y)|^2}\right) \\
&= \frac{|p(x)-p(y)|^2}{(1-|p(x)|^2)(1-|p(y)|^2)}.
\end{aligned}$$

Consequently, the function $\varphi : \mathbb{D} \times \mathbb{D} \to \mathbb{R}$ defined by

$$\varphi(x, y) = \frac{|x-y|^2}{(1-|x|^2)(1-|y|^2)}$$

is invariant under the action of $\text{Möb}^+(\mathbb{D})$.

Note that φ is also invariant under the action of complex conjugation, as

$$\varphi(\overline{x}, \overline{y}) = \frac{|\overline{x}-\overline{y}|^2}{(1-|\overline{x}|^2)(1-|\overline{y}|^2)} = \frac{|x-y|^2}{(1-|x|^2)(1-|y|^2)} = \varphi(x, y).$$

Hence, we can conclude that the function $\varphi : \mathbb{D} \times \mathbb{D} \to \mathbb{R}$ defined by

$$\varphi(x, y) = \frac{|x-y|^2}{(1-|x|^2)(1-|y|^2)}$$

is invariant under the action of all of $\text{Möb}(\mathbb{D})$.

The main application of the invariance of φ under the action of $\text{Möb}(\mathbb{D})$ is to provide a link between the Euclidean and hyperbolic distances between a pair of points of $\mathbb{D}$.

Proposition 4.3

For each pair x and y of points of $\mathbb{D}$, we have that

$$\varphi(x, y) = \sinh^2\left(\frac{1}{2}\mathrm{d}_{\mathbb{D}}(x, y)\right) = \frac{1}{2}\left(\cosh(\mathrm{d}_{\mathbb{D}}(x, y)) - 1\right).$$

Proof

The proof of Proposition 4.3 is by direct calculation. Let x and y be a pair of points in $\mathbb{D}$. Choose an element $p(z) = \frac{\alpha z+\beta}{\overline{\beta}z+\overline{\alpha}}$ of $\text{Möb}^+(\mathbb{D})$ (so that α, $\beta \in \mathbb{C}$ and $|\alpha|^2 - |\beta|^2 = 1$) for which $p(x) = 0$ and $p(y)$ is real and positive.

One way to do this is to set $\beta = -\alpha x$, so that

$$p(z) = \frac{\alpha(z-x)}{\overline{\alpha}(-\overline{x}z+1)},$$

where $|\alpha|^2(1-|x|^2) = 1$. Now choose the argument of α so that $p(y) = r$ is real and positive. Then,

$$\begin{aligned} \frac{|x-y|^2}{(1-|x|^2)(1-|y|^2)} &= \varphi(x, y) \\ &= \varphi(p(x), p(y)) = \varphi(0, r) = \frac{r^2}{1-r^2}. \end{aligned}$$

By Exercise 4.2, we have that $r = \tanh(\frac{1}{2}\mathrm{d}_{\mathbb{D}}(0, r))$, and so

$$\varphi(x, y) = \frac{r^2}{1 - r^2} = \sinh^2\left(\frac{1}{2}\mathrm{d}_{\mathbb{D}}(x, y)\right) = \frac{1}{2}\left(\cosh(\mathrm{d}_{\mathbb{D}}(x, y)) - 1\right),$$

as desired. This completes the proof of Proposition 4.3. **QED**

Exercise 4.3

Let ℓ_1 and ℓ_2 be two intersecting hyperbolic lines in $\mathbb{D}$, where the endpoints at infinity of ℓ_1 are z_1 and z_2, and the endpoints at infinity of ℓ_2 are w_1 and w_2, labeled so that the order of the points counterclockwise around $\mathbb{S}^1$ is z_1, w_1, z_2, w_2. Prove that the angle θ between ℓ_1 and ℓ_2 satisfies

$$[z_1, w_1; z_2, w_2]\tan^2\left(\frac{\theta}{2}\right) = -1.$$

We close this section with a discussion of hyperbolic circles.

Definition 4.4

A *hyperbolic circle* in $\mathbb{D}$ is a set in $\mathbb{D}$ of the form

$$C = \{y \in \mathbb{D} \mid \mathrm{d}_{\mathbb{D}}(x, y) = s\},$$

where $x \in \mathbb{D}$ and $s > 0$ are fixed. We refer to x as the *hyperbolic centre* of C and s as the *hyperbolic radius* of C.

We can completely characterize hyperbolic circles in $\mathbb{D}$.

Proposition 4.5

A hyperbolic circle in $\mathbb{D}$ is a Euclidean circle in $\mathbb{D}$ and vice versa, although the hyperbolic and Euclidean centres, and the hyperbolic and Euclidean radii, will in general be different.

Proof

We begin with a specific set of hyperbolic circles in $\mathbb{D}$, namely, those centred at 0. Given $s > 0$, set $r = \tanh(\frac{1}{2}s)$, so that $\mathrm{d}_{\mathbb{D}}(0, r) = s$. As $\mathrm{M\ddot{o}b}(\mathbb{D})$ contains $p(z) = e^{i\theta}z$, we see that p is an isometry of $(\mathbb{D}, \mathrm{d}_{\mathbb{D}})$, and so every point $re^{i\theta}$ in $\mathbb{D}$ satisfies $\mathrm{d}_{\mathbb{D}}(0, re^{i\theta}) = s$ as well. Hence, the Euclidean circle with Euclidean centre 0 and Euclidean radius r and the hyperbolic circle with hyperbolic centre 0 and hyperbolic radius s are the same, where s and r are related by $r = \tanh(\frac{1}{2}s)$.

Now, let C be the hyperbolic circle in $\mathbb{D}$ with hyperbolic centre c and hyperbolic radius s. Let m be an element of $\mathrm{M\ddot{o}b}(\mathbb{D})$ taking c to 0. Then, $m(C)$ is the hyperbolic circle in $\mathbb{D}$ with hyperbolic centre 0 and hyperbolic radius s. In particular, $m(C)$ is also a Euclidean circle. Then, as the elements of $\mathrm{M\ddot{o}b}(\mathbb{D})$ take circles in $\overline{\mathbb{C}}$ to circles in $\overline{\mathbb{C}}$ and as no element of $\mathrm{M\ddot{o}b}(\mathbb{D})$ maps a point of $\mathbb{D}$ to ∞, we see that $C = m^{-1} \circ m(C)$ is also a Euclidean circle in $\mathbb{D}$

Conversely, let C be a Euclidean circle in $\mathbb{D}$. We can assume that the Euclidean centre of C is not 0, as otherwise we would have that C is a hyperbolic circle by the argument given in the first paragraph. Let L be the Euclidean line through 0 and the Euclidean centre of C, and note that C and L meet perpendicularly. Note also that L is also a hyperbolic line in $\mathbb{D}$. Let c be the hyperbolic midpoint on L of the two points of $C \cap L$. Choose an element m of $\mathrm{M\ddot{o}b}(\mathbb{D})$ taking c to 0 and taking L to $\mathbb{R}$; as the two points of $C \cap L$ are equidistant from c and lie on L, m takes them to the two points s, $-s$ for some $s \in \mathbb{R}$. Then, $m(C)$ is a Euclidean circle in $\mathbb{D}$ perpendicular to $\mathbb{R}$, passing through s and $-s$. In particular, $m(C)$ is a Euclidean circle in $\mathbb{D}$ centred at 0, and so it is a hyperbolic circle in $\mathbb{D}$ centred at 0 as well. Hence, C is a hyperbolic circle in $\mathbb{D}$.

This completes the proof of Proposition 4.5. **QED**

Exercise 4.4

Given $s > 0$, let S_s be the hyperbolic circle in $\mathbb{D}$ with hyperbolic centre 0 and hyperbolic radius s. Show that the hyperbolic length of S_s is

$$\mathrm{length}_{\mathbb{D}}(S_s) = 2\pi \sinh(s).$$

In the Euclidean geometry of $\mathbb{C}$, we can draw a circle of any Euclidean centre and any Euclidean radius, and any three noncolinear points determine a circle.

In hyperbolic geometry, Exercise 4.4 and the proof of Proposition 4.5 show that we can find a circle in $\mathbb{D}$ of any hyperbolic centre and any hyperbolic radius.

However, it is not the case in hyperbolic geometry that any three noncolinear points determine a circle. As an explicit example of this, consider the three points $z_1 = \frac{1}{2}$, $z_2 = 0$, and $z_3 = -\frac{1}{2} + i\varepsilon$ in $\mathbb{D}$, for sufficiently small values of ε. We know that the hyperbolic circle through z_1, z_2, z_3, if it exists, is also a Euclidean circle. However, for small values of ε, the Euclidean circle through z_1, z_2, z_3 has a large Euclidean radius, and so it passes outside the Poincaré disc $\mathbb{D}$.

We transported hyperbolic geometry from the upper half-plane model $\mathbb{H}$ to the Poincaré disc $\mathbb{D}$ using any element p of Möb taking $\mathbb{D}$ to $\mathbb{H}$. We have seen that hyperbolic circles and Euclidean circles in the Poincaré disc $\mathbb{D}$ are the same, and so using p, we see that hyperbolic circles and Euclidean circles in the upper half-plane $\mathbb{H}$ are the same as well.

It is not difficult to calculate the hyperbolic centres and hyperbolic radii of circles in either the upper half-plane or Poincaré disc models of the hyperbolic plane.

Exercise 4.5

Let A be the Euclidean circle in the Poincaré disc $\mathbb{D}$ with Euclidean centre $\frac{1}{5} - \frac{1}{4}i$ and Euclidean radius $\frac{1}{10}$. Determine the hyperbolic centre and hyperbolic radius of A.

Exercise 4.6

Let A be the Euclidean circle in the upper half-plane $\mathbb{H}$ with Euclidean centre $1 + 3i$ and Euclidean radius 1. Determine the hyperbolic centre and hyperbolic radius of A.

Exercise 4.7

Let A be a circle in the upper half-plane $\mathbb{H}$. Suppose the Euclidean centre of A is $a + ib$ and the Euclidean radius of A is r. Show that the hyperbolic centre is $a + i\sqrt{b^2 - r^2}$ and the hyperbolic radius R satisfies $r = b\tanh(R)$.

4.2 A General Construction

The construction from Section 4.1, of transferring hyperbolic geometry from the upper half-plane $\mathbb{H}$ to the Poincaré disc $\mathbb{D}$, is actually just a single instance of a more general method of constructing planar models of the hyperbolic plane from the upper half-plane model $\mathbb{H}$ using tools from complex analysis. The purpose of this section is to explore this relationship between complex analysis and planar models of the hyperbolic plane in more detail.

We work in a somewhat restricted setting. Let X be a subset of $\mathbb{C}$ that is *holomorphically equivalent* to $\mathbb{H}$, which means that X is a subset of $\mathbb{C}$ for which there exists a homeomorphism $\xi : X \to \mathbb{H}$ for which ξ and its inverse ξ^{-1} are both holomorphic, as described in Note 3.9. We refer to such a function ξ as a *holomorphic homeomorphism* between $\mathbb{H}$ and X. (We postpone the discussion of which open subsets of $\mathbb{C}$ are holomorphically equivalent to $\mathbb{H}$ until later in this section.)

One example of this sort of function that we have already seen, in Section 4.1, is to take $X = \mathbb{D}$ and to consider

$$\xi(z) = \frac{\frac{i}{\sqrt{2}}z + \frac{1}{\sqrt{2}}}{-\frac{1}{\sqrt{2}}z - \frac{i}{\sqrt{2}}} = \frac{iz+1}{-z-i}.$$

In a crude fashion, we may use ξ to transfer the hyperbolic geometry from $\mathbb{H}$ to X and so to get a model of the hyperbolic plane whose underlying space is X. Specifically, define a *hyperbolic line in* X to be the image in X of a hyperbolic line in $\mathbb{H}$ under ξ^{-1}. So, a hyperbolic line in X has either the form $\{z \in X \mid \mathrm{Re}(\xi(z)) = c\}$ for $c \in \mathbb{R}$ or the form $\{z \in X \mid |\xi(z) - c|^2 = r^2\}$, where $c \in \mathbb{R}$ and $r > 0$.

As a specific example, let X be the quarter-plane

$$X = \{z \in \mathbb{C} \mid \mathrm{Re}(z) > 0 \text{ and } \mathrm{Im}(z) > 0\},$$

and consider the holomorphic homeomorphism $\xi : X \to \mathbb{H}$ given by $\xi(z) = z^2$.

We can explicitly describe the hyperbolic lines in this model X. If we let $w = u + iv$ be the coordinate on X and if we let z be the coordinate on $\mathbb{H}$, then $\xi(w) = z = u^2 - v^2 + 2iuv$. The hyperbolic lines in $\mathbb{H}$ are of two types, those contained in the Euclidean line $L_c = \{z \in \mathbb{H} \mid \mathrm{Re}(z) = c\}$ and those contained in the Euclidean circle $A_{c,r} = \{z \in \mathbb{H} \mid (\mathrm{Re}(z) - c)^2 + (\mathrm{Im}(z))^2 = r^2\}$.

The image of L_c under ξ^{-1} is the curve $\{w \in X \mid u^2 - v^2 = c\}$ in X. For $c = 0$, this curve is the Euclidean ray K from 0 making angle $\frac{\pi}{4}$ with the positive real

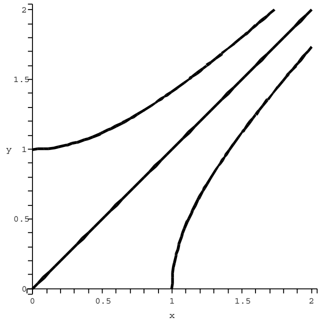

Figure 4.2: Some hyperbolic lines of the form L_c in the quarter-plane X

axis, whereas for $c \neq 0$, this curve is a hyperboloid asymptotic to K. See Figure 4.2.

The image of $A_{c,r}$ under ξ^{-1} is a curve known as an *oval of Cassini*, given by the equation

$$(u^2 + v^2)^2 - 2c(u^2 - v^2) + c^2 = r^2.$$

See Figure 4.3 for some ovals of Cassini for various values of $c \in \mathbb{R}$ and $r > 0$. An oval of Cassini is a variant on an ellipse. Let w_0 and w_1 be two fixed points in $\mathbb{C}$. Whereas an ellipse is the set of points in $\mathbb{C}$ for which the sum of the (Euclidean) distances $|w - w_0| + |w - w_1|$ is constant, an oval of Cassini is the set of points w in $\mathbb{C}$ for which the product of the (Euclidean) distances $|(w-w_0)(w-w_1)|$ is constant. Unlike ellipses, all of which have the same shape, the shapes of ovals of Cassini change as the value of this constant changes.

Although having a description of the hyperbolic lines in such a set X is nice, it is not in general easy to work with. It would often be more computationally useful to transfer the hyperbolic element of arc-length $\frac{1}{\text{Im}(z)}|\text{d}z|$ on $\mathbb{H}$ to X using ξ, to obtain a hyperbolic element of arc-length on X, so that we may actually calculate in this new model of the hyperbolic plane with underlying space X.

We accomplish this transfer of the hyperbolic element of arc-length from $\mathbb{H}$ to X exactly as we accomplished the transfer of the hyperbolic element of arc-length from $\mathbb{H}$ to $\mathbb{D}$, but now using the holomorphic homeomorphism $\xi : X \to \mathbb{H}$. Define the hyperbolic element of arc-length $\text{d}s_X$ on X by declaring that

$$\text{length}_X(f) = \int_f \text{d}s_X = \int_{\xi\circ f} \frac{1}{\text{Im}(z)}\,|\text{d}z| = \text{length}_{\mathbb{H}}(\xi \circ f)$$

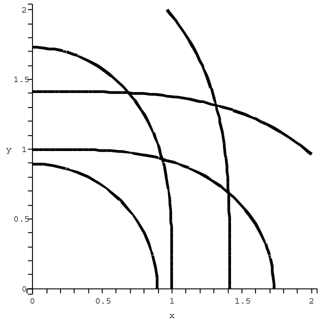

Figure 4.3: Some hyperbolic lines of the form $A_{c,r}$ in the quarter-plane X

for every piecewise C^1 path $f : [a,b] \to X$.

This construction of the element of arc-length $\mathrm{d}s_X$ on X using the holomorphic homeomorphism $\xi : X \to \mathbb{H}$ is often referred to as defining $\mathrm{d}s_X$ to be the *pullback* of the element of arc-length on $\mathbb{H}$ by ξ.

Theorem 4.6

Suppose that X is an open subset of the complex plane $\mathbb{C}$ and that $\xi : X \to \mathbb{H}$ is a holomorphic homeomorphism. The pullback $\mathrm{d}s_X$ in X of the hyperbolic element of arc-length $\frac{1}{\mathrm{Im}(z)}\,|\mathrm{d}z|$ on $\mathbb{H}$ is

$$\mathrm{d}s_X = \frac{1}{\mathrm{Im}(\xi(z))}|\xi'(z)||\mathrm{d}z|.$$

Proof

The proof of Theorem 4.6 is a direct calculation. Proceeding as above, let $f : [a,b] \to X$ be a piecewise C^1 path. The hyperbolic length of f is given by

$$\begin{aligned}\mathrm{length}_X(f) = \int_f \mathrm{d}s_X &= \int_{\xi\circ f} \frac{1}{\mathrm{Im}(z)}|\mathrm{d}z| \\ &= \int_a^b \frac{1}{\mathrm{Im}(\xi(f(t)))}|\xi'(f(t))||f'(t)|\mathrm{d}t \\ &= \int_f \frac{1}{\mathrm{Im}(\xi(z))}|\xi'(z)||\mathrm{d}z|.\end{aligned}$$

Applying Lemma 3.10 completes the proof of Theorem 4.6. **QED**

In exactly the same way that we defined the hyperbolic metric on $\mathbb{D}$ and determined its group of isometries, this construction allows us to define the hyperbolic metric on X and determine its group of isometries.

Specifically, let X be an open subset of $\mathbb{C}$ for which there exists a holomorphic homeomorphism $\xi : X \to \mathbb{H}$. Let $\mathrm{d}s_X$ be the pullback of the hyperbolic element of arc-length $\frac{1}{\mathrm{Im}(z)}|\mathrm{d}z|$ on $\mathbb{H}$ by ξ. Then, we can use $\mathrm{d}s_X$ to define a hyperbolic metric d_X on X by taking the infimum of hyperbolic lengths of piecewise C^1 paths in X.

Using the same proofs as we used in Section 4.1 for the Poincaré disc $\mathbb{D}$, we see that (X, d_X) is a path metric space in which the distance-realizing paths are precisely the almost simple parametrizations of the hyperbolic line segments in X. Also, the group of isometries of (X, d_X) is

$$\mathrm{Isom}(X, \mathrm{d}_X) = \{\xi^{-1} \cdot m \cdot \xi \mid m \in \mathrm{M\ddot{o}b}(\mathbb{H})\}.$$

Define a *hyperbolic structure* on X to be all hyperbolic data coming from this construction, which includes the hyperbolic lines in X, the hyperbolic element $\mathrm{d}s_X$ of arc-length on X, and the resulting hyperbolic metric on X, with its group of isometries.

We continue with the example above, where

$$X = \{z \in \mathbb{C} \mid \mathrm{Re}(z) > 0 \text{ and } \mathrm{Im}(z) > 0\},$$

and with the holomorphic homeomorphism $\xi : X \to \mathbb{H}$ given by $\xi(z) = z^2$.

As

$$\mathrm{Im}(\xi(z)) = \mathrm{Im}(z^2) = 2\,\mathrm{Re}(z)\,\mathrm{Im}(z)$$

and

$$|\xi'(z)| = |2z| = 2|z|,$$

we see that the pullback of $\frac{1}{\mathrm{Im}(z)}|\mathrm{d}z|$ by ξ is

$$\mathrm{d}s_X = \frac{1}{\mathrm{Im}(\xi(z))}|\xi'(z)||\mathrm{d}z| = \frac{|z|}{\mathrm{Re}(z)\,\mathrm{Im}(z)}|\mathrm{d}z|.$$

Note that if we were to choose a different holomorphic homeomorphism $\mu : X \to \mathbb{H}$, we would still get the same hyperbolic element of arc-length on X via the pullback construction. To see this, consider $\mu(z) = \frac{3z^2+5}{z^2+2}$. To see that μ is indeed a holomorphic homeomorphism, note that $\mu(z)$ is the composition $\mu(z) = m \circ \xi(z)$, where $m(z) = \frac{3z+5}{z+2}$ is an element of $\mathrm{M\ddot{o}b}^+(\mathbb{H})$ and hence is a holomorphic homeomorphism of $\mathbb{H}$. Calculating the pullback of $\frac{1}{\mathrm{Im}(z)}|\mathrm{d}z|$ by μ, we see that

$$\mathrm{Im}(\mu(z)) = \mathrm{Im}\left(\frac{3z^2+5}{z^2+2}\right) = \frac{2\mathrm{Re}(z)\mathrm{Im}(z)}{|z^2+2|^2}$$

and (using the Chain Rule)

$$|\mu'(z)| = |m'(\xi(z))|\,|\xi'(z)| = \frac{2|z|}{|z^2+2|^2},$$

and so the pullback on X of the hyperbolic element of arc-length $\frac{1}{\mathrm{Im}(z)}|\mathrm{d}z|$ by μ is

$$\frac{1}{\mathrm{Im}(\mu(z))}|\mu'(z)||\mathrm{d}z| = \frac{|z^2+2|^2}{2\mathrm{Re}(z)\mathrm{Im}(z)}\frac{2|z|}{|z^2+2|^2}|\mathrm{d}z| = \frac{|z|}{\mathrm{Re}(z)\,\mathrm{Im}(z)}|\mathrm{d}z|,$$

as desired.

Exercise 4.8

Let $Y = \{z \in \mathbb{C} \mid \mathrm{Re}(z) > 0\}$, and consider the holomorphic homeomorphism $\xi : Y \to \mathbb{H}$ given by $\xi(z) = iz$. Determine the pullback of $\frac{1}{\mathrm{Im}(z)}|\mathrm{d}z|$ by ξ.

Exercise 4.9

Let $X = \{z \in \mathbb{C} \mid 0 < \mathrm{Im}(z) < \pi\}$, and consider the holomorphic homeomorphism $\xi : X \to \mathbb{H}$ given by $\xi(z) = e^z$. Determine the pullback of $\frac{1}{\mathrm{Im}(z)}|\mathrm{d}z|$ by ξ.

However, even for simple regions, it is not always possible to give an explicit description of the hyperbolic structure. As a specific example, consider the rectangle $R(\ell) = (-\ell, \ell) \times \left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$ in $\mathbb{C}$. For every $\ell > 0$, this rectangle $R(\ell)$ is holomorphically equivalent to $\mathbb{H}$. To date, there is no closed form expression for the hyperbolic element of arc-length on $R(\ell)$. We refer the interested reader to two recent papers of Beardon [8], [9], in which properties of the hyperbolic element of arc-length and the resulting hyperbolic metric on the rectangle $R(\ell)$ are explored in some detail.

This construction is general, and it can be applied to any open subset of $\mathbb{C}$ that is holomorphically equivalent to $\mathbb{H}$. We can characterize such sets. We also wish to show that this construction is independent of the holomorphic homeomorphism between X and $\mathbb{H}$. In doing these things, we will need to use results that are beyond the scope of this book; when we do so, we will provide references for the interested reader. So, let X be any *connected* and *simply connected* open subset of $\mathbb{C}$. Roughly, a connected set is a set that has only one piece.

Definition 4.7

An open set X in $\mathbb{C}$ is *connected* if, given any two points x and y in X, there exists a piecewise C^1 path $f : [a, b] \to X$ with $f(a) = x$ and $f(b) = y$.

Roughly, a simply connected set does not have any holes. Before defining simply connected, we need to know what a *Jordan curve* is.

Definition 4.8

A set C is a *Jordan curve* if there exists a continuous function $f : [0, 2\pi] \to \mathbb{C}$ so that $f(0) = f(2\pi)$, f is injective on $[0, 2\pi)$, and $f([0, 2\pi]) = C$.

The unit circle $\mathbb{S}^1$ in $\mathbb{C}$ is an example of a Jordan curve.

The Jordan curve theorem states that complement in $\mathbb{C}$ of a Jordan curve C has exactly two components, one bounded and the other unbounded. We refer to the bounded component of $\mathbb{C} - C$ as the *disc bounded by* C. We note that some, but not all, Jordan curves can be parametrized by piecewise C^1 paths. A good discussion of the Jordan curve theorem can be found in Guillemin and Pollack [18].

Definition 4.9

A connected open set X in $\mathbb{C}$ is *simply connected* if for every Jordan curve C in X, the disc bounded by C is also contained in X.

Both the upper half-plane $\mathbb{H}$ and the unit disc $\mathbb{D}$ are connected and simply connected. The punctured plane $\mathbb{C} - \{0\}$ is connected but is not simply connected, because the unit circle $\mathbb{S}^1$ is contained in $\mathbb{C} - \{0\}$ but the disc bounded by $\mathbb{S}^1$, namely, the unit disc $\mathbb{D}$, is not contained in $\mathbb{C} - \{0\}$.

Definition 4.10

A *holomorphic disc* is an open subset X of $\mathbb{C}$ that is connected and simply connected, and that is not all of $\mathbb{C}$.

The classical uniformization theorem yields that there are only two possibilities for a connected and simply connected open (nonempty) subset X of $\mathbb{C}$. One possibility is that $X = \mathbb{C}$. The other possibility is that X is a holomorphic disc and that there exists a holomorphic homeomorphism $\xi : X \to \mathbb{H}$. Hence, in the case in which X is a holomorphic disc, it is possible to put a hyperbolic structure on X, and hence to do hyperbolic geometry on X. As we will see a bit later, it is not possible to put a hyperbolic element of arc-length on all of $\mathbb{C}$.

An exact statement and proof of the classical uniformization theorem is beyond the scope of this book. A good exposition can be found in the article of Abikoff [2] and the sources contained in its bibliography.

We would like to argue that the hyperbolic structure on a holomorphic disc X is independent of the choice of the holomorphic homeomorphism $\xi : X \to \mathbb{H}$. To make this argument, we need to use another result that is beyond the scope of this book. This is a slight variation of the classical Schwarz lemma. A proof can be found in most texts on complex analysis, such as Ahlfors [3] or Hille [21]. The interested reader may also wish to consult Ahlfors [4].

We begin with the statement of the Schwarz lemma, which is usually formulated in terms of holomorphic functions from $\mathbb{D}$ into $\mathbb{D}$.

Theorem 4.11

Let $f : \mathbb{D} \to \mathbb{D}$ be holomorphic and satisfy $f(0) = 0$. Then, either $|f(z)| < |z|$ for all $z \in \mathbb{D}$, $z \neq 0$, or $f(z) = e^{i\alpha} z$ for some $\alpha \in \mathbb{R}$.

As a consequence of Theorem 4.11, we can show that any holomorphic homeomorphism of $\mathbb{D}$ is an element of $\text{Möb}^+(\mathbb{D})$.

Corollary 4.12

Let $f : \mathbb{D} \to \mathbb{D}$ be a holomorphic homeomorphism. Then, $f \in \text{Möb}^+(\mathbb{D})$.

Proof

First, choose an element $m \in \text{Möb}^+(\mathbb{D})$ for which $m \circ f(0) = 0$, and let $F = m \circ f$. Then, F is a holomorphic homeomorphism of $\mathbb{D}$ that satisfies the hypothesis of Theorem 4.11. As F is a holomorphic homeomorphism of $\mathbb{D}$, its inverse $G = F^{-1}$ is also a holomorphic homeomorphism of $\mathbb{D}$ and $G(0) = 0$.

If the latter conclusion of Theorem 4.11 holds for either F or G, then it holds for both and implies directly that F and G are elements of $\text{Möb}^+(\mathbb{D})$. Therefore, it suffices to assume that the former conclusion of Theorem 4.11 holds for both F and G, so that $|F(z)| < |z|$ and $|G(z)| < |z|$ for all $z \in \mathbb{D}$, $z \neq 0$. Hence, applying Theorem 4.11 twice, we see that

$$|z| = |G \circ F(z)| < |F(z)| < |z|$$

for all $z \in \mathbb{D}$, $z \neq 0$, which cannot occur. Hence, it must be that $F(z) = e^{i\alpha}z$ for some $\alpha \in \mathbb{R}$, and so $F \in \text{Möb}^+(\mathbb{D})$. Hence, $f = m^{-1} \circ F \in \text{Möb}^+(\mathbb{D})$, as desired. **QED**

As we have chosen to define hyperbolic structures on holomorphic discs using holomorphic homeomorphisms with $\mathbb{H}$ rather than with $\mathbb{D}$, we need the following immediate corollary of Theorem 4.11.

Corollary 4.13

Let $\xi : \mathbb{H} \to \mathbb{H}$ be a holomorphic homeomorphism. Then, $\xi \in \text{Möb}^+(\mathbb{H})$.

We are now ready to show that the hyperbolic structure on a holomorphic disc does not depend on the choice of holomorphic homeomorphism $\xi : X \to \mathbb{H}$.

Theorem 4.14

Let X be a holomorphic disc. Then, the hyperbolic lines in X and the hyperbolic element of arc-length $\mathrm{d}s_X$ on X are independent of the choice of the holomorphic homeomorphism $\xi : X \to \mathbb{H}$.

Proof

Let $\xi : X \to \mathbb{H}$ and $\nu : X \to \mathbb{H}$ be two holomorphic homeomorphisms. Then, $\xi \circ \nu^{-1} : \mathbb{H} \to \mathbb{H}$ is a holomorphic homeomorphism. By Corollary 4.13, we have that $\xi \circ \nu^{-1} = p \in \text{Möb}^+(\mathbb{H})$. Write $\xi = p \circ \nu$.

In particular, let ℓ be a hyperbolic line in $\mathbb{H}$. Then, because p^{-1} takes hyperbolic lines in $\mathbb{H}$ to hyperbolic lines in $\mathbb{H}$, we have that $\xi^{-1}(\ell) = \nu^{-1}(p^{-1}(\ell))$ is a hyperbolic line in X if and only if $\nu^{-1}(\ell)$ is a hyperbolic line in X.

Furthermore, because

$$\frac{1}{\mathrm{Im}(p(z))}|p'(z)| = \frac{1}{\mathrm{Im}(z)}$$

for all $z \in \mathbb{H}$, we see that

$$\frac{1}{\mathrm{Im}(\xi(w))}|\xi'(w)| = \frac{1}{\mathrm{Im}(p(\nu(w)))}|p'(\nu(w))|\,|\nu'(w)| = \frac{1}{\mathrm{Im}(\nu(w))}|\nu'(w)|$$

for all $w \in X$, and so the hyperbolic element of arc-length $\mathrm{d}s_X$ is independent of the choice of the holomorphic homeomorphism between X and $\mathbb{H}$. **QED**

In exactly the same way as was done at the end of Section 3.7, we can define the *boundary at infinity* of any holomorphic disc X, by considering the collection of all hyperbolic rays in X and imposing the same equivalence relation as in Section 3.7. For the upper half-plane $\mathbb{H}$ (respectively, the Poincaré disc $\mathbb{D}$), the boundary at infinity constructed using equivalence classes of hyperbolic rays is the same as the circle in $\overline{\mathbb{C}}$ bounding the disc $\mathbb{H}$ (respectively, $\mathbb{D}$) in $\overline{\mathbb{C}}$. The main difficulty is that, unlike in the case of $\mathbb{H}$ or $\mathbb{D}$, for a general holomorphic disc X, this intrinsic boundary at infinity constructed using hyperbolic rays is not necessarily the same as the topological boundary ∂X of X in $\overline{\mathbb{C}}$. In fact, there need not even be a continuous function from the boundary at infinity to ∂X. The topological characterization of those holomorphic discs X for which the topological boundary ∂X in $\mathbb{C}$ and the boundary at infinity of X are related is subtle, and it is beyond what we can discuss here.

Up to this point, we have been content to describe hyperbolic geometry on a single holomorphic disc in the complex plane $\mathbb{C}$. There are several directions in which we can expand our analysis. One direction is to consider open connected subsets of $\mathbb{C}$ that are not simply connected. Whereas this is an extremely interesting and fruitful direction to pursue, we do not have the means to do so here.

Instead, we must be content to consider the relationship of the hyperbolic structures of two holomorphic discs. That is, we handle several holomorphic discs simultaneously, rather than expanding to consider sets beyond holomorphic discs. We begin with the case of nested holomorphic discs. For a proof of the following theorem, the interested reader is directed to Ahlfors [4].

Theorem 4.15

Let X_1 and X_2 be two holomorphic discs in $\mathbb{C}$ satisfying $X_1 \subset X_2$. For $k = 1, 2$, express the hyperbolic element of arc-length on X_k as $\lambda_k(z)|\mathrm{d}z|$. Then, $\lambda_1(z) \geq \lambda_2(z)$ for every $z \in X_1$.

We can check that this relationship holds in a particular case in which it is easy to calculate the respective hyperbolic elements of arc-length.

Exercise 4.10

Let $D_{a,r}$ denote the open Euclidean disc with Euclidean centre $a \in \mathbb{D}$ and Euclidean radius $r > 0$, where $|a| + r < 1$ (so that $D_{a,r} \subset \mathbb{D}$). Express the hyperbolic element of arc-length on $D_{a,r}$ as $\lambda_{a,r}(z)|\mathrm{d}z|$, and show that $\lambda_{a,r}(z) \geq \frac{2}{1-|z|^2}$ for all $z \in \mathbb{D}$.

For nested pairs of general holomorphic discs, the difficulty in performing this check is that we often cannot explicitly calculate the respective hyperbolic elements of arc-length. However, we can use Theorem 4.15 to estimate the hyperbolic element of arc-length on any holomorphic disc.

For the most basic estimate, let X be a holomorphic disc in $\mathbb{C}$, and express the hyperbolic element of arc-length on X as $\lambda(z)|\mathrm{d}z|$. Consider the function $\delta : X \to (0, \infty)$ defined by setting

$$\delta(z) = \inf\{|z - x| \mid x \in \partial X\},$$

which measures the shortest (Euclidean) distance from z to a point in the topological boundary ∂X of X. Let $w \in X$ be any point. Then, the (Euclidean) disc $D = D_{w,\delta(w)}$ (with (Euclidean) centre w and (Euclidean) radius $\delta(w)$) is contained in X and is the largest (Euclidean) disc with (Euclidean) centre w contained in X. By Exercise 4.10, the hyperbolic element of arc-length on D has the form $\lambda_w(z)|\mathrm{d}z|$, where

$$\lambda_w(z) = \frac{2\delta(w)}{\delta^2(w) - |z - w|^2}.$$

By Theorem 4.15, we then see that

$$\lambda(w) \leq \lambda_w(w) = \frac{2}{\delta(w)}.$$

We note that in this case, there is also an upper bound on $\lambda(w)$, namely, that

$$\frac{1}{2\delta(w)} \leq \lambda(w),$$

but its proof is beyond the scope of the present discussion. Again, we refer the interested reader to Ahlfors [4] for a discussion of this latter inequality.

In particular, these estimates show that on a holomorphic disc X, the conformal distortion $\lambda(z)$ of the hyperbolic element of arc-length $\lambda(z)|\mathrm{d}z|$ from the Euclidean element of arc-length at $z \in X$ is roughly inversely proportional to the Euclidean distance from z to ∂X. For the Poincaré disc $\mathbb{D}$, we can see this relationship explicitly. The conformal distortion of the hyperbolic element of arc-length on $\mathbb{D}$ is

$$\lambda(z) = \frac{2}{1-|z|^2} = \frac{2}{(1-|z|)(1+|z|)}.$$

The Euclidean distance $\delta(z)$ from $z \in \mathbb{D}$ to $\mathbb{S}^1 = \partial\mathbb{D}$ is $1-|z|$, and for all $z \in \mathbb{D}$, we have the bounds

$$1 < \frac{2}{1+|z|} \leq 2.$$

Therefore, on $\mathbb{D}$, we have the estimate

$$\frac{1}{\delta(z)} < \lambda(z) \leq \frac{2}{\delta(z)},$$

which is slightly better than the general estimate described above.

We can also use these estimates on conformal distortion to convince ourselves that there does not exist a hyperbolic element of arc-length on all of $\mathbb{C}$. Suppose to the contrary that there were a hyperbolic element of arc-length $\rho(z)|\mathrm{d}z|$ on $\mathbb{C}$, where ρ is defined on all of $\mathbb{C}$ and $\rho(z) \geq 0$ for all $z \in \mathbb{C}$. For each $n \in \mathbb{N}$, the (Euclidean) disc $D_{0,n} = \{z \in \mathbb{C} \mid |z| < n\}$ is contained in $\mathbb{C}$. Express the hyperbolic element of arc-length on $D_{0,n}$ as $\lambda_n(z)|\mathrm{d}z|$. Then, by Exercise 4.10 and Theorem 4.15, we have that $\rho(z) \leq \lambda_n(z)$ for all $n \in \mathbb{N}$ and all $z \in D_{0,n}$.

Fix $z_0 \in \mathbb{C}$. There exists $n_0 > 0$ so that $z_0 \in D_{0,n}$ for all $n > n_0$. Hence,

$$0 \leq \rho(z_0) \leq \lambda_n(z_0)$$

for all $n > n_0$. Using the formula for $\lambda_n(z_0)$, this equation becomes

$$0 \leq \rho(z_0) \leq \frac{2n}{n^2 - |z_0|^2}.$$

As $n \to \infty$, the right-hand side satisfies $\lim_{n\to\infty} \frac{2n}{n^2-|z_0|^2} = 0$, and so $\rho(z_0) = 0$. As z_0 is arbitrary, this implies that $\rho(z) = 0$ for all $z \in \mathbb{C}$, which contradicts the assumption that $\rho(z)|\mathrm{d}z|$ is a hyperbolic element of arc-length on $\mathbb{C}$.

We now consider the case of two holomorphic discs X_1 and X_2 and a holomorphic function $f : X_1 \to X_2$. Let $\xi_k : X_k \to \mathbb{H}$ be any holomorphic homeomorphism. Recall that the hyperbolic structure on X_k is constructed as the pullback of the hyperbolic structure on $\mathbb{H}$ by ξ_k, and that a different choice of holomorphic homeomorphism from X_k to $\mathbb{H}$ differs from ξ_k by composition with an element of $\text{Möb}^+(\mathbb{H})$, by the proof of Theorem 4.14.

If f is a homeomorphism as well, then f is an isometry from the hyperbolic structure on X_1 to the hyperbolic structure on X_2. To see this, note that because $\xi_2 \circ f$ and ξ_1 are both holomorphic homeomorphisms from X_1 to $\mathbb{H}$, there exists an element m of $\text{Möb}^+(\mathbb{H})$ so that $\xi_2 \circ f = m \circ \xi_1$. In particular, this construction yields that $f = \xi_2^{-1} \circ m \circ \xi_1$. As each of ξ_2^{-1}, m, and ξ_1 is an isometry, we see that f is an isometry.

More interesting is the case in which f is not a homeomorphism. In this case, we begin with a equivariant formulation of Corollary 4.13 due to Pick.

Theorem 4.16

Let $f : \mathbb{H} \to \mathbb{H}$ be holomorphic. If f is a homeomorphism, then $f \in \text{Möb}^+(\mathbb{H})$. If f is not a homeomorphism, then $\text{d}_{\mathbb{H}}(f(z_1), f(z_2)) < \text{d}_{\mathbb{H}}(z_1, z_2)$ for all distinct z_1, $z_2 \in \mathbb{H}$.

To prove Theorem 4.16, we first consider $f : \mathbb{D} \to \mathbb{D}$ holomorphic with $f(0) = 0$. Assuming that f is not a homeomorphism, we have that $|f(z)| < |z|$ for all $z \in \mathbb{D}$, $z \neq 0$, by Theorem 4.11. As the function $h : [0, 1) \to [0, \infty)$ given by $h(t) = \ln\left(\frac{1+t}{1-t}\right)$ is increasing (as can be seen by verifying that $h'(t) > 0$ for all $0 < t < 1$), we see that

$$\text{d}_{\mathbb{D}}(0, f(z)) = \ln\left(\frac{1 + |f(z)|}{1 - |f(z)|}\right) < \text{d}_{\mathbb{D}}(0, z) = \ln\left(\frac{1 + |z|}{1 - |z|}\right)$$

for all $z \in \mathbb{D}$, $z \neq 0$. The proof for a general holomorphic function $f : \mathbb{D} \to \mathbb{D}$ follows by precomposing and postcomposing by appropriate elements of $\text{Möb}^+(\mathbb{D})$.

The proof follows for $\mathbb{H}$ and for a general holomorphic disc X using the same style of argument we have used several times already, when transferring the hyperbolic structure on $\mathbb{H}$ to a hyperbolic structure on a holomorphic disc. Let $f : X_1 \to X_2$ be a function that is holomorphic but is not a homeomorphism, and for $k = 1$, 2, let $\xi_k : X_k \to \mathbb{H}$ be a holomorphic homeomorphism. Then, we have that $\xi_2 \circ f \circ \xi_1^{-1} : \mathbb{H} \to \mathbb{H}$ is holomorphic but is not a homeomorphism. We apply Theorem 4.16 to $\xi_2 \circ f \circ \xi_1^{-1}$ to see that $\xi_2 \circ f \circ \xi_1^{-1}$ decreases hyperbolic distance. As both ξ_1^{-1} and ξ_2 are isometries, this implies that f must decrease hyperbolic distance.

The fact that holomorphic functions are nonincreasing on hyperbolic distance can also be formulated in terms of the respective hyperbolic elements of arclength.

Proposition 4.17

Let $f : \mathbb{H} \to \mathbb{H}$ be holomorphic. Then,

$$\frac{|f'(z)|}{\mathrm{Im}(f(z))} \leq \frac{1}{\mathrm{Im}(z)}$$

for all $z \in \mathbb{H}$.

Proof

If f is constant, then $f'(z) = 0$ for all $z \in \mathbb{H}$, and the inequality follows. So, we can assume that f is nonconstant. We now need to make use of a fact about nonconstant holomorphic functions: If f is a nonconstant holomorphic function on $\mathbb{H}$, then $f(\mathbb{H})$ is open. So, if z and w are sufficiently close in $\mathbb{H}$, the hyperbolic line segment ℓ joining $f(z)$ and $f(w)$ lies in $f(\mathbb{H})$. Choose a piecewise C^1 path $c : [0, L] \to \mathbb{H}$ so that $f \circ c$ is an almost simple parametrization of this hyperbolic line segment ℓ.

Calculating, we see that

$$\mathrm{d}_{\mathbb{H}}(f(z), f(w)) = \int_{f\circ c} \frac{1}{\mathrm{Im}(z)} |\mathrm{d}z| = \int_c \frac{|f'(z)|}{\mathrm{Im}(f(z))} |\mathrm{d}z|$$

and that

$$\mathrm{d}_{\mathbb{H}}(z, w) \leq \int_c \frac{1}{\mathrm{Im}(z)} |\mathrm{d}z|.$$

As $\mathrm{d}_{\mathbb{H}}(f(z), f(w)) < \mathrm{d}_{\mathbb{H}}(z, w)$ by Theorem 4.16, we have that

$$\int_c \frac{|f'(z)|}{\mathrm{Im}(f(z))} |\mathrm{d}z| < \int_c \frac{1}{\mathrm{Im}(z)} |\mathrm{d}z|.$$

As z and w are arbitrary (as long as they are sufficiently close together), this is enough to guarantee that

$$\frac{|f'(z)|}{\mathrm{Im}(f(z))} \leq \frac{1}{\mathrm{Im}(z)},$$

as desired. **QED**

Exercise 4.11

Consider the function $f : \mathbb{D} \to \mathbb{D}$ given by $f(z) = z^2$. Show that f is nonincreasing in terms of hyperbolic distance by showing that, if we express the pullback of the hyperbolic element of arc-length on $\mathbb{D}$ by f as $\lambda(z)|\mathrm{d}z|$, then $\lambda(z) \leq \frac{2}{1-|z|^2}$.

We close this section by introducing the notion of curvature. Let X be an open subset of $\mathbb{C}$, and let $\mathrm{d}s_X = \alpha(z)|\mathrm{d}z|$ be an element of arc-length on X, where α is a positive real-valued differentiable function on X. By the same constructions we have seen several times, this element of arc-length induces a metric on X, where the distance between two points is the infimum of the lengths of all piecewise C^1 paths joining the two points, where the length of a piecewise C^1 path is calculated with respect to this element of arc-length.

There is a numerical quantity associated to the metric that arises from this element of arc-length, called the *curvature* of the metric. The study of metrics and their properties, such as curvature, is properly the subject of differential geometry, but we will say a few words here.

The curvature of the metric induced by $\mathrm{d}s_X = \alpha(z)|\mathrm{d}z|$ is a function curv : $X \to \mathbb{R}$, which is given explicitly by the formula

$$\mathrm{curv}(z) = -\left[\frac{2}{\alpha(z)}\right]^2 \partial\overline{\partial}\log(\alpha(z)).$$

Here, if we write $z = x + iy$ and $\beta(z) = f(x, y) + ig(x, y)$, we set

$$\partial\beta = \frac{1}{2}\left[\frac{\partial}{\partial x}\beta - i\frac{\partial}{\partial y}\beta\right] = \frac{1}{2}\left[\frac{\partial f}{\partial x} + \frac{\partial g}{\partial y} + i\left(\frac{\partial g}{\partial x} - \frac{\partial f}{\partial y}\right)\right]$$

and

$$\overline{\partial}\beta = \frac{1}{2}\left[\frac{\partial}{\partial x}\beta + i\frac{\partial}{\partial y}\beta\right] = \frac{1}{2}\left[\frac{\partial f}{\partial x} - \frac{\partial g}{\partial y} + i\left(\frac{\partial g}{\partial x} + \frac{\partial f}{\partial y}\right)\right].$$

Exercise 4.12

Check that

$$\partial\overline{\partial}\beta = \frac{1}{4}\left[\frac{\partial^2\beta}{\partial x^2} + \frac{\partial^2\beta}{\partial y^2}\right].$$

In particular, note that for the standard Euclidean metric on $\mathbb{C}$, we have $\alpha \equiv 1$, and so the curvature of the Euclidean metric on $\mathbb{C}$ is identically zero.

For the upper half-plane model $\mathbb{H}$, the hyperbolic element of arc-length is $\frac{1}{\mathrm{Im}(z)}|\mathrm{d}z|$, and so the curvature function is

$$\mathrm{curv}(z) = -4\,(\mathrm{Im}(z))^2\,\partial\overline{\partial}\log\left(\frac{1}{\mathrm{Im}(z)}\right) = -1.$$

Note that if we consider the slightly more general hyperbolic element of arc-length $\frac{c}{\mathrm{Im}(z)}|\mathrm{d}z|$ on $\mathbb{H}$, we have that the curvature is $-\frac{1}{c^2}$.

Exercise 4.13

Calculate the curvature of the hyperbolic metric on $\mathbb{D}$ coming from the hyperbolic element of arc-length $\frac{2}{1-|z|^2}|\mathrm{d}z|$.

Exercise 4.14

Calculate the curvature of the metric on $\mathbb{C}$ coming from the element of arc-length $\frac{1}{1+|z|^2}|\mathrm{d}z|$.

5
Convexity, Area, and Trigonometry

In this chapter, we explore some finer points of hyperbolic geometry. We first describe the notion of *convexity* and explore *convex sets*, including the class of *hyperbolic polygons*. Restricting our attention to hyperbolic polygons, we go on to discuss the measurement of *hyperbolic area*, including the *Gauss–Bonnet formula*, which gives a formula for the hyperbolic area of a hyperbolic polygon in terms of its angles. We go on to use the Gauss–Bonnet formula to show that nontrivial *dilations* of the hyperbolic plane do not exist. We close the chapter with a discussion of the *laws of trigonometry* in the hyperbolic plane.

5.1 Convexity

We now have a good working knowledge of the geometry of the hyperbolic plane. We have several different models to work in, and we have fairly explicit descriptions of hyperbolic length and hyperbolic distance in these models. We now begin to explore some of the finer points of hyperbolic geometry.

In this section, we consider the notion of *convexity*. Recall that we know what it means for a set Z in the complex plane $\mathbb{C}$ to be convex, namely, that for each pair z_0 and z_1 of distinct points in Z, the Euclidean line segment joining z_0 and z_1 also lies in Z. In $\mathbb{C}$, this can be expressed formulaically by saying that Z is convex if for each pair z_0 and z_1 of distinct points of Z, the points $z_t = (1-t)z_0 + tz_1$ for $0 \leq t \leq 1$ also lie in Z.

We can consider this definition in the hyperbolic plane.

Definition 5.1

A subset X of the hyperbolic plane is *convex* if for each pair of distinct points x and y in X, the closed hyperbolic line segment ℓ_{xy} joining x to y is contained in X.

Unlike in the Euclidean geometry on the complex plane $\mathbb{C}$, there is not in general, in the models of the hyperbolic plane we have encountered so far, a nice parametrization of the hyperbolic line segment joining two arbitrary points.

Note that because convexity is defined in terms of hyperbolic line segments, it is an immediate consequence of the definition that convexity is preserved by hyperbolic isometries. That is, if X is a convex set in the hyperbolic plane and if γ is an isometry of the hyperbolic plane, then $\gamma(X)$ is also convex.

Proposition 5.2

Hyperbolic lines, hyperbolic rays, and hyperbolic line segments are convex.

Proof

Let ℓ be a hyperbolic line, and let x and y be two points of ℓ. By Proposition 1.2, x and y determine a unique hyperbolic line, namely, ℓ, and so the closed hyperbolic line segment ℓ_{xy} joining x to y is necessarily contained in ℓ. Hence, ℓ is convex.

This same argument also shows that hyperbolic rays and hyperbolic line segments are convex. This completes the proof of Proposition 5.2. **QED**

Convexity behaves well under intersections.

Exercise 5.1

Suppose that $\{X_\alpha\}_{\alpha \in A}$ is a collection of convex subsets of the hyperbolic plane. Prove that the intersection $X = \cap_{\alpha \in A} X_\alpha$ is convex.

Another example, and in a sense the most basic example, of a convex set in the hyperbolic plane is a *half-plane*, as discussed in Section 2.9. To recall the definition, given a hyperbolic line ℓ in the hyperbolic plane, the complement of ℓ in the hyperbolic plane has two components, which are the two *open half-planes determined by* ℓ.

A *closed half-plane determined by* ℓ is the union of ℓ with one of the two open half-planes determined by ℓ. We often refer to ℓ as the *bounding line* for the half-planes it determines. We describe why half-planes can be thought of as the most basic convex sets at the end of this section.

We now show that half-planes are convex.

Proposition 5.3

Open half-planes and closed half-planes in the hyperbolic plane are convex.

Proof

We work in the upper half-plane model $\mathbb{H}$, and we begin with a specific half-plane. Let I be the positive imaginary axis in $\mathbb{H}$, and consider the open right half-plane

$$U = \{z \in \mathbb{H} \mid \mathrm{Re}(z) > 0\}$$

in $\mathbb{H}$ determined by I.

Let x and y be two points of U. If $\mathrm{Re}(x) = \mathrm{Re}(y)$, then the hyperbolic line segment ℓ_{xy} joining x to y is contained in the Euclidean line $L = \{z \in \mathbb{H} \mid \mathrm{Re}(z) = \mathrm{Re}(x)\}$. As $\mathrm{Re}(x) > 0$, we see that L is contained in U, and so ℓ_{xy} is contained in U.

If $\mathrm{Re}(x) \neq \mathrm{Re}(y)$, then the hyperbolic line segment ℓ_{xy} joining x to y lies in the Euclidean circle C with centre on the real axis $\mathbb{R}$. As the intersection of C and I contains at most one point, and as both x and y are contained in U, we see that ℓ_{xy} is contained in U.

So, U is convex. Combining this argument with the fact that $\mathrm{Möb}(\mathbb{H})$ acts transitively on the set of open half-planes of $\mathbb{H}$, by Exercise 2.43, and the fact that $\mathrm{Möb}(\mathbb{H})$ preserves convexity, we have that every open half-plane in $\mathbb{H}$ is convex.

We may repeat this argument without change with a closed half-plane, and obtain that closed half-planes are convex as well. This completes the proof of Proposition 5.3. **QED**

Exercise 5.2

Prove that the open hyperbolic disc D_s in the Poincaré disc $\mathbb{D}$ with hyperbolic centre 0 and hyperbolic radius $s > 0$ is convex. Conclude that all hyperbolic discs are convex.

On the other hand, convexity does not behave well under unions. To take one example, let ℓ_1 and ℓ_2 be two distinct, although not necessarily disjoint, hyperbolic lines.

Take points z_1 on ℓ_1 and z_2 on ℓ_2, chosen only so that neither z_1 nor z_2 is the point of intersection $\ell_1 \cap \ell_2$ of ℓ_1 and ℓ_2. Then, the hyperbolic line segment ℓ_{12} joining z_1 to z_2 does not lie in $\ell_1 \cup \ell_2$, and so $\ell_1 \cup \ell_2$ is not convex.

For an illustration of this phenomenon for the two hyperbolic lines in $\mathbb{H}$ contained in the positive imaginary axis and the unit circle, see Figure 5.1.

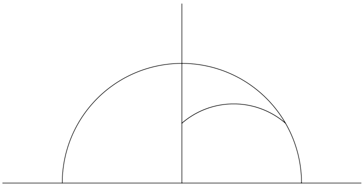

Figure 5.1: The nonconvex union of two hyperbolic lines

Adding the hypothesis of convexity allows us to refine some results from Section 3.7 about the properties of the hyperbolic metric and the hyperbolic distance between sets. For example, consider the following generalization of Exercise 3.19.

Proposition 5.4

Let X be a closed, convex subset of $\mathbb{H}$, and let z be a point of $\mathbb{H}$ not in X. Then, there exists a unique point $x \in X$ with $\mathrm{d}_{\mathbb{H}}(z, x) = \mathrm{d}_{\mathbb{H}}(z, X)$.

Proof

We first show that there exists some point x of X with $\mathrm{d}_{\mathbb{H}}(z, x) = \mathrm{d}_{\mathbb{H}}(z, X)$. As $\mathrm{d}_{\mathbb{H}}(z, X) = \inf\{\mathrm{d}_{\mathbb{H}}(z, x) \mid x \in X\}$, there exists a sequence $\{x_n\}$ of points of

X so that

$$\lim_{n\to\infty} \mathrm{d}_{\mathbb{H}}(z, x_n) = \mathrm{d}_{\mathbb{H}}(z, X).$$

In particular, by the definition of convergence, there is some $N > 0$ so that $\mathrm{d}_{\mathbb{H}}(z, x_n) \leq \mathrm{d}_{\mathbb{H}}(z, X) + 1$ for $n \geq N$. Set $C = \mathrm{d}_{\mathbb{H}}(z, X) + 1$, and let

$$V_C(z) = \{w \in \mathbb{H} \mid \mathrm{d}_{\mathbb{H}}(z, w) \leq C\}$$

denote the closed hyperbolic disc with hyperbolic centre z and hyperbolic radius C.

The subset $X \cap V_C(z)$ of $\mathbb{H}$ is closed and bounded, and hence it is compact. As $\{x_n \mid n \geq N\}$ is a sequence contained in the compact subset $X \cap V_C(z)$ of $\mathbb{H}$, there exists a subsequence $\{x_{n_k}\}$ of $\{x_n \mid n \geq N\}$ that converges to some point x of $\mathbb{H}$. As each x_{n_k} is contained in X and as X is closed, we have that $x \in X$.

As

$$\mathrm{d}_{\mathbb{H}}(z, x) = \lim_{k\to\infty} \mathrm{d}_{\mathbb{H}}(z, x_{n_k}) = \lim_{n\to\infty} \mathrm{d}_{\mathbb{H}}(z, x_n),$$

we have that $\mathrm{d}_{\mathbb{H}}(z, x) = \mathrm{d}_{\mathbb{H}}(z, X)$.

We need now to show that this point x is unique. So, suppose there are two points x_1 and x_2 of X so that

$$\mathrm{d}_{\mathbb{H}}(z, X) = \mathrm{d}_{\mathbb{H}}(z, x_1) = \mathrm{d}_{\mathbb{H}}(z, x_2).$$

Let ℓ_{12} be the hyperbolic line segment joining x_1 to x_2, and let ℓ be the hyperbolic line containing ℓ_{12}.

By Exercise 3.19, there exists a unique point x_0 of ℓ so that $\mathrm{d}_{\mathbb{H}}(z, \ell) = \mathrm{d}_{\mathbb{H}}(z, x_0)$. Moreover, looking at the solution to Exercise 3.19, the hyperbolic distance from z to a point y of ℓ increases monotonically as a function of the hyperbolic distance $\mathrm{d}_{\mathbb{H}}(x_0, y)$ between x_0 and y.

In particular, because $\mathrm{d}_{\mathbb{H}}(z, x_1) = \mathrm{d}_{\mathbb{H}}(z, x_2)$, the point x_0 of ℓ realizing the hyperbolic distance $\mathrm{d}_{\mathbb{H}}(z, \ell)$ must lie between x_1 and x_2. That is, x_0 is contained in ℓ_{12}.

As x_1 and x_2 are both points of X, the convexity of X gives that ℓ_{12} is contained in X, and hence that x_0 is a point of X. However, if $x_1 \neq x_2$, then

$$\mathrm{d}_{\mathbb{H}}(z, x_0) < \mathrm{d}_{\mathbb{H}}(z, x_1) = \mathrm{d}_{\mathbb{H}}(z, X),$$

which is a contradiction. This completes the proof of Proposition 5.4. **QED**

Note that open convex sets do not have the property shown to hold for closed convex sets in Proposition 5.4. For example, let U be an open half-plane determined by a hyperbolic line ℓ. Then, for each point $z \in \ell$, we have that $\mathrm{d}_{\mathbb{H}}(z, U) = 0$, but there does not exist a point $x \in U$ with $\mathrm{d}_{\mathbb{H}}(z, x) = 0$.

The examples of convex sets given to this point have been either open or closed subsets of the hyperbolic plane. There are also convex sets in the hyperbolic plane that are neither open nor closed.

To take a specific example, let U be an open half-plane determined by a hyperbolic line ℓ, let x and y be two points on ℓ, and let ℓ_{xy} be the closed hyperbolic line segment joining x and y. Then, the union $U \cup \ell_{xy}$ is convex, but it is neither open nor closed.

In fact, let U be an open half-plane determined by the hyperbolic line ℓ, and let X be any subset of ℓ. Then, the union $U \cup X$ is convex if and only if X is a convex subset of ℓ. We leave this for the interested reader to verify.

A common way exists of generating convex sets in the hyperbolic plane, namely, by taking convex hulls. Given a subset Y of the hyperbolic plane, the *convex hull* $\mathrm{conv}(Y)$ of Y is the intersection of all convex sets in the hyperbolic plane containing Y.

For example, for the set $Y = \{x, y\}$ containing two distinct points, the convex hull $\mathrm{conv}(Y)$ of Y is the closed hyperbolic line segment ℓ_{xy} joining x and y. By Proposition 5.2, we know that ℓ_{xy} is convex. It remains only to show that there does not exist a convex set containing x and y that is properly contained in ℓ_{xy}. But from the definition of convexity, we see that any convex set containing x and y must contain the closed hyperbolic line segment ℓ_{xy} joining them, and so $\mathrm{conv}(Y) = \ell_{xy}$.

Naively, we should expect the convex hull of a convex set to be the convex set, and this is indeed the case.

Exercise 5.3

Let X be a convex set in the hyperbolic plane. Prove that $\mathrm{conv}(X) = X$.

We saw above that the convex hull of a pair of distinct points in the hyperbolic plane is the closed hyperbolic line segment joining the two points, and so the convex hull of a nonconvex set can be considerably larger than the set.

Exercise 5.4

Let ℓ_1 and ℓ_2 be two distinct hyperbolic lines. Determine the convex hull $\text{conv}(\ell_1 \cup \ell_2)$ of their union.

The reason we said earlier in this section that half-planes are the most basic convex sets in the hyperbolic plane is that convex sets in the hyperbolic plane are generally the sets that can be expressed as the intersection of a collection of half-planes.

To get a feel for the question being considered, we start with a particular example, namely, the positive imaginary axis I in $\mathbb{H}$. As I is a hyperbolic line in $\mathbb{H}$, we know from Proposition 5.2 that I is convex.

To express I as the intersection of a collection of closed half-planes, consider the two closed half-planes A and B determined by I, namely,

$$A = \{z \in \mathbb{H} \mid \text{Re}(z) \geq 0\} \text{ and } B = \{z \in \mathbb{H} \mid \text{Re}(z) \leq 0\}$$

Then, $A \cap B = I = \{z \in \mathbb{H} \mid \text{Re}(z) = 0\}$.

We may also express I as the intersection of a collection of open half-planes, by expressing each closed half-plane above as the intersection of a collection of open half-planes. Specifically, for each $\varepsilon > 0$, let

$$A_\varepsilon = \{z \in \mathbb{H} \mid \text{Re}(z) > -\varepsilon\} \text{ and } B_\varepsilon = \{z \in \mathbb{H} \mid \text{Re}(z) < \varepsilon\}.$$

Then, we can express A as the intersection $A = \cap_{\varepsilon>0} A_\varepsilon$ and B as the intersection $B = \cap_{\varepsilon>0} B_\varepsilon$. Hence, we can express I as

$$I = \cap_{\varepsilon>0}(A_\varepsilon \cap B_\varepsilon).$$

As $\text{Möb}(\mathbb{H})$ acts transitively on the set $\mathcal{L}$ of hyperbolic lines in $\mathbb{H}$, this argument shows that every hyperbolic line can be expressed both as the intersection of a collection of closed half-planes and as the intersection of a collection of open half-planes.

Exercise 5.5

Express a closed hyperbolic ray and a closed hyperbolic line segment both as the intersection of a collection of open half-planes and as the intersection of a collection of closed half-planes.

Theorem 5.5

A closed subset X of the hyperbolic plane is convex if and only if X can be expressed as the intersection of a collection of half-planes.

Proof

We have already proven one direction of Theorem 5.5: We know from Proposition 5.3 that half-planes are convex, and we know from Exercise 5.1 that the intersection of a collection of convex sets is convex. Hence, the intersection of a collection of half-planes is convex.

Suppose now that X is a closed convex set in the hyperbolic plane. It remains only to show that X can be expressed as the intersection of a collection of half-planes. We work in the upper half-plane model $\mathbb{H}$ of the hyperbolic plane.

Let z be a point of $\mathbb{H}$ that is not contained in X. By Proposition 5.4, there exists a unique point $x_z \in X$ with $\mathrm{d}_{\mathbb{H}}(z, x_z) = \mathrm{d}_{\mathbb{H}}(z, X)$.

Let M_z be the hyperbolic line segment joining x_z and z, and let L_z be the hyperbolic line perpendicular to M_z and passing through x_z. The hyperbolic line L_z is the bounding line for two half-planes, the open half-plane A_z containing z and the closed half-plane B_z not containing z.

We show that X is contained in B_z. Suppose not. As A_z and B_z are disjoint half-planes whose union is $\mathbb{H}$, there must then exist a point p_z of $X \cap A_z$. Let ℓ_z be the hyperbolic line segment joining x_z to p_z, and let ℓ be the hyperbolic line containing ℓ_z.

As M_z is perpendicular to L_z, and as M_z and L_z intersect at x_z, we have by Proposition 5.4 and the solution to Exercise 3.19 that

$$\mathrm{d}_{\mathbb{H}}(z, y) \geq \mathrm{d}_{\mathbb{H}}(z, x_z)$$

for every point y in L_z, with equality if and only if $y = x_z$.

Also, for any point y of B_z that is not contained in L_z, the hyperbolic line segment joining y to z intersects L_z, and so we have that

$$\mathrm{d}_{\mathbb{H}}(z, y) > \mathrm{d}_{\mathbb{H}}(L_z, z) = \mathrm{d}_{\mathbb{H}}(x_z, z)$$

as well. That is, we have that

$$\mathrm{d}_{\mathbb{H}}(y, z) \geq \mathrm{d}_{\mathbb{H}}(x_z, z)$$

for every point y of the closed half-plane B_z, with equality if and only if $y = x_z$.

Now apply Proposition 5.4 to the point z and the hyperbolic line ℓ. The only hyperbolic line through x_z that intersects M_z perpendicularly is L_z. As p_z is contained in A_z and as A_z and L_z are disjoint, we have that $\ell \neq L_z$, and so ℓ and M_z cannot intersect perpendicularly.

As ℓ and M_z do not intersect perpendicularly, the solution to Exercise 3.19 implies that there exists a point a of ℓ so that

$$\mathrm{d}_{\mathbb{H}}(a, z) < \mathrm{d}_{\mathbb{H}}(x_z, z).$$

By the argument just given, this point a cannot lie in B_z, and so there exists a point a of $\ell \cap A_z$ so that

$$\mathrm{d}_{\mathbb{H}}(a, z) < \mathrm{d}_{\mathbb{H}}(x_z, z).$$

Let a_ℓ be the point of ℓ given by the solution to Exercise 3.19 that satisfies

$$\mathrm{d}_{\mathbb{H}}(z, a_\ell) \leq \mathrm{d}_{\mathbb{H}}(z, a)$$

for every point a of ℓ. As $\mathrm{d}_{\mathbb{H}}(a, z)$ is monotone increasing as a function of $\mathrm{d}_{\mathbb{H}}(a, a_\ell)$, we see that a_ℓ is contained in A_z as well.

In particular, regardless of whether a_ℓ is or is not contained in ℓ_z, there exists a point b_z of ℓ_z that satisfies

$$\mathrm{d}_{\mathbb{H}}(z, b_z) < \mathrm{d}_{\mathbb{H}}(z, x_z).$$

However, as X is convex and as both endpoints p_z and x_z of ℓ_z are points of X, we have that ℓ_z is contained in X. Here is where we make use of the convexity of X.

So, we have constructed a point b_z of X for which

$$\mathrm{d}_{\mathbb{H}}(z, b_z) < \mathrm{d}_{\mathbb{H}}(z, x_z).$$

This construction contradicts the choice of x_z. This contradiction completes the proof of the claim that X is contained in the closed half-plane B_z.

Finally, note that we can express X as the intersection of closed half-planes

$$X = \cap\{B_z \mid z \in \mathbb{H} \text{ and } z \notin X\}.$$

This completes the proof of Theorem 5.5. **QED**

One consequence of Theorem 5.5 is that we can generalize the scope of our definition. Namely, we can define what it means for a subset of the union of the hyperbolic plane and its boundary at infinity to be convex, without actually altering the definition in any essential way. To make this explicit, we work in the upper half-plane model $\mathbb{H}$, whose boundary at infinity is $\overline{\mathbb{R}}$.

Each hyperbolic line ℓ in $\mathbb{H}$ determines a pair of points in $\overline{\mathbb{R}}$, namely, its endpoints at infinity, and each half-plane determined by ℓ is naturally associated with one of the two arcs in $\overline{\mathbb{R}}$ determined by this pair of points. So, with only a slight abuse of language, we can speak of a half-plane in $\mathbb{H}$ containing a point in $\overline{\mathbb{R}}$.

In particular, for a subset X of $\mathbb{H} \cup \overline{\mathbb{R}}$, define the *convex hull* $\text{conv}(X)$ of X in $\mathbb{H}$ to be the intersection of all half-planes in $\mathbb{H}$ containing X.

For example, if x and y are two distinct points in $\overline{\mathbb{R}}$, the convex hull $\text{conv}(Y)$ of the set $Y = \{x, y\}$ is the hyperbolic line determined by x and y. Similarly, if z is a point of $\mathbb{H}$ and if x is a point of $\overline{\mathbb{R}}$, the convex hull $\text{conv}(Z)$ of the set $Z = \{z, x\}$ is the closed hyperbolic ray determined by z and x.

5.2 Hyperbolic Polygons

As in Eulidean geometry, the *polygon* is one of the basic objects in hyperbolic geometry. In the Euclidean plane, a polygon is a closed convex set that is bounded by Euclidean line segments. We would like to mimic this definition as much as possible in the hyperbolic plane.

Starting from the definition of convexity and its characterization in Section 5.1, namely, that a convex set is the intersection of a collection of half-planes, we need to impose a condition on this collection. The condition we impose is *local finiteness.*

Definition 5.6

Let $\mathcal{H} = \{H_\alpha\}_{\alpha \in A}$ be a collection of half-planes in the hyperbolic plane, and for each $\alpha \in A$, let ℓ_α be the bounding line for H_α. The collection $\mathcal{H}$ is *locally finite* if for each point z in the hyperbolic plane, there exists some $\varepsilon > 0$ so that only finitely many bounding lines ℓ_α of the half-planes in $\mathcal{H}$ intersect the open hyperbolic disc $U_\varepsilon(z)$ of hyperbolic radius ε and hyperbolic centre z.

In other words, even though the collection $\{H_\alpha\}$ may be infinite, near each point it looks as though it is a finite collection when viewed in the hyperbolic disc of radius ε. Note that the value of ε needed will in general depend on the point z.

It is easy to see that every finite collection $\mathcal{H} = \{H_k\}_{1 \leq k \leq n}$ of half-planes is locally finite, because every open disc $U_\varepsilon(z)$ in the hyperbolic plane can intersect at most n bounding lines, because there are only n half-planes in the collection.

Less easy to see is that there cannot exist an uncountable collection of half-planes that is locally finite.

Exercise 5.6

Prove that an uncountable collection of distinct half-planes in the hyperbolic plane cannot be locally finite.

One example of an infinite collection of half-planes that is locally finite is the collection $\{H_n\}_{n\in\mathbb{Z}}$ in $\mathbb{H}$, where the bounding line ℓ_n of H_n lies in the Euclidean circle with Euclidean centre n and Euclidean radius 1, and where H_n is the closed half-plane determined by ℓ_n that contains the point $2i$. Part of this collection of bounding lines is shown in Figure 5.2.

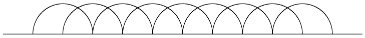

Figure 5.2: Some bounding lines

To see that $\{H_n\}$ is locally finite, take some point $x \in \mathbb{H}$. For each $\varepsilon > 0$, consider the open hyperbolic disc $U_\varepsilon(x)$ with hyperbolic centre x and hyperbolic radius ε. The hyperbolic distance between x and the hyperbolic line ℓ_μ contained in the Euclidean line $\{z \in \mathbb{H} \mid \mathrm{Re}(z) = \mathrm{Re}(x) + \mu\}$ satisfies

$$\mathrm{d}_{\mathbb{H}}(x, \ell_\mu) < \frac{\mu}{\mathrm{Im}(x)},$$

because the right-hand side is the hyperbolic length of the Euclidean line segment joining x to $x + \mu$.

In particular, the hyperbolic disc $U_\varepsilon(x)$ is contained in the strip

$$\{z \in \mathbb{H} \mid \mathrm{Re}(x) - \varepsilon\,\mathrm{Im}(x) < \mathrm{Re}(z) < \mathrm{Re}(x) + \varepsilon\,\mathrm{Im}(x)\}.$$

As for each $\varepsilon > 0$ this strip intersects only finitely many ℓ_n, we see that the collection $\{H_n\}$ is locally finite.

However, just because a collection of half-planes is countable does not imply that it is locally finite. For example, consider the collection $\mathcal{H} = \{H_n\}_{n\in\mathbb{N}}$ of closed half-planes in $\mathbb{H}$, where the bounding line ℓ_n of H_n is the hyperbolic line in $\mathbb{H}$ contained in the Euclidean circle of Euclidean radius 1 and Euclidean centre $\frac{1}{n}$, and where H_n is the closed half-plane determined by ℓ_n that contains $2i$.

To see that $\mathcal{H}$ is not locally finite, we observe that for each $\varepsilon > 0$ the open hyperbolic disc $U_\varepsilon(i)$ intersects infinitely many ℓ_n, including those for which the hyperbolic distance $\mathrm{d}_{\mathbb{H}}(i, \frac{1}{n} + i)$ satisfies $\mathrm{d}_{\mathbb{H}}(i, \frac{1}{n} + i) < \varepsilon$.

Definition 5.7

A *hyperbolic polygon* is a closed convex set in the hyperbolic plane that can be expressed as the intersection of a locally finite collection of closed half-planes.

One thing to note about this definition is that for a given hyperbolic polygon P, there will always be many different locally finite collections of closed half-planes whose intersection is P. Also, we use closed half-planes in the definition, because a closed subset of $\mathbb{H}$ cannot be expressed as the intersection of a locally finite collection of open half-planes.

We have already seen one example of a hyperbolic polygon in $\mathbb{H}$, namely, $\cap_{n\in\mathbb{Z}} H_n$, where H_n is the closed half-plane determined by the hyperbolic line ℓ_n contained in the Euclidean circle with Euclidean centre $n \in \mathbb{Z}$ and Euclidean radius 1.

Another example of a hyperbolic polygon in $\mathbb{H}$ is shown in Figure 5.3. It is the intersection of the five closed half-planes $H_1 = \{z \in \mathbb{H} \mid \mathrm{Re}(z) \leq 1\}$, $H_2 = \{z \in \mathbb{H} \mid \mathrm{Re}(z) \geq -1\}$, $H_3 = \{z \in \mathbb{H} \mid |z| \geq 1\}$, $H_4 = \{z \in \mathbb{H} \mid |z-1| \geq 1\}$, and $H_5 = \{z \in \mathbb{H} \mid |z+1| \geq 1\}$.

In addition to individual hyperbolic polygons, we can also consider families of hyperbolic polygons. For this example, we work in the Poincaré disc $\mathbb{D}$. For $r > 1$, consider the hyperbolic polygon P_r that is the intersection of the four closed half-planes

$$H_k = \{z \in \mathbb{D} \mid |z - ri^k| \geq \sqrt{r^2 - 1}\}$$

for $k = 0, 1, 2, 3$. For an illustration of such a P_r with $r = 1.5$, see Figure 5.4.

Up to this point, none of our definitions have made use of any intrinsic property of any specific model of the hyperbolic plane. In fact, everything we have said

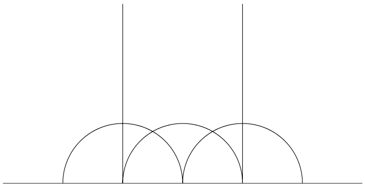

Figure 5.3: A hyperbolic polygon in $\mathbb{H}$

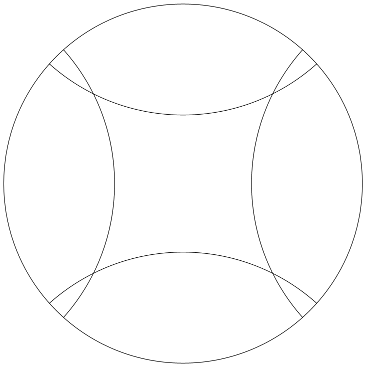

Figure 5.4: A hyperbolic polygon in $\mathbb{D}$

makes sense in every model, and so we are free to apply these definitions in whichever model is most convenient or most comfortable.

Note that by the definition of hyperbolic polygon we have chosen, there are some subsets of the hyperbolic plane that satisfy the definition of a hyperbolic polygon, but we do not want to consider them to be hyperbolic polygons.

For example, a hyperbolic line ℓ is a hyperbolic polygon, because it is a closed convex set in the hyperbolic plane that can be expressed as the intersection $A_\ell \cap B_\ell$, where A_ℓ and B_ℓ are the two closed half-planes determined by ℓ.

It is a bit uncomfortable having a hyperbolic line as a hyperbolic polygon. One way to get around this possibility is to impose another condition. Recall that the *interior* of a set X in the hyperbolic plane is the largest open set contained in X. The interior of a hyperbolic line is empty, because a hyperbolic line does not contain an open subset of the hyperbolic plane.

Definition 5.8

A hyperbolic polygon is *nondegenerate* if it has nonempty interior. A hyperbolic polygon is *degenerate* if it has empty interior.

Unless explicitly stated otherwise, we assume that *all hyperbolic polygons are nondegenerate.* For instance, the examples of hyperbolic polygons shown in the figures in this section are all nondegenerate. And as it turns out, the degenerate hyperbolic polygons are easy to understand.

Exercise 5.7

> Prove that a degenerate hyperbolic polygon is either a hyperbolic line, a closed hyperbolic ray, a closed hyperbolic line segment, or a point.

Let P be a hyperbolic polygon in the hyperbolic plane. The boundary ∂P of P has a nice decomposition. To see this decomposition, let ℓ be a hyperbolic line that intersects P. It may be that ℓ intersects the interior of P. In this case, the intersection $P \cap \ell$ is a closed convex subset of ℓ that is not a point, and so it is either a closed hyperbolic line segment in ℓ, a closed hyperbolic ray in ℓ, or all of ℓ.

On the other hand, it may be that ℓ does not pass through the interior of P. In this case, P is contained in a closed half-plane determined by ℓ. The proof of this result is similar to the analysis carried out in detail in Section 5.1. The intersection $P \cap \ell$ is again a closed convex subset of ℓ, and so it is either a point in ℓ, a closed hyperbolic line segment in ℓ, a closed hyperbolic ray in ℓ, or all of ℓ. All four possibilities can occur, as is shown in Figure 5.5.

Here, the hyperbolic polygon P is the intersection of four closed half-planes, namely, $H_1 = \{z \in \mathbb{H} \mid \mathrm{Re}(z) \leq 1\}$, $H_2 = \{z \in \mathbb{H} \mid \mathrm{Re}(z) \geq -1\}$, $H_3 = \{z \in \mathbb{H} \mid |z| \geq 1\}$, and $H_4 = \{z \in \mathbb{H} \mid |z+1| \geq 1\}$. The bounding lines of P intersect P in turn in a hyperbolic line, a closed hyperbolic ray, a closed hyperbolic ray, and a closed hyperbolic line segment. The hyperbolic line ℓ whose endpoints at infinity are -3 and $-\frac{1}{2}$ intersects P in a single point.

In general, let P be a hyperbolic polygon and let ℓ be a hyperbolic line so that P intersects ℓ and so that P is contained in a closed half-plane determined by ℓ. If the intersection $P \cap \ell$ is a point, we say that this point is a *vertex* of P. In the other cases, namely, that the intersection $P \cap \ell$ is either a closed hyperbolic

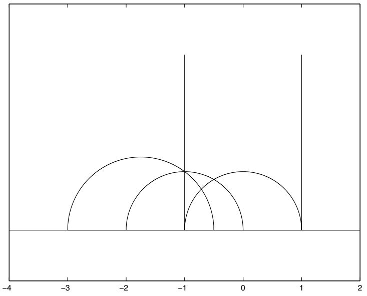

Figure 5.5: Intersections of hyperbolic lines with a hyperbolic polygon

line segment, a closed hyperbolic ray, or all of ℓ, we say that this intersection is a *side* of P. The sides and vertices of a hyperbolic polygon are closely related.

Lemma 5.9

Let P be a hyperbolic polygon. Each vertex of P is an endpoint of a side of P.

Proof

Lemma 5.9 is a fairly direct consequence of our definition of a hyperbolic polygon as the intersection of a locally finite collection of closed half-planes of the hyperbolic plane.

To start the proof of Lemma 5.9, express P as the intersection of a locally finite collection $\mathcal{H}$ of distinct closed half-planes. Write $\mathcal{H} = \{H_n\}_{n\in A}$, where A is a (necessarily) countable set, and let ℓ_n be the bounding line of H_n.

Let p be a point of ∂P. The local finiteness of $\mathcal{H}$ implies that there exists some $\varepsilon_0 > 0$ so that only finitely many ℓ_n intersect the open hyperbolic disc $U_{\varepsilon_0}(p)$.

For $\delta < \varepsilon_0$, the number of bounding lines that intersect $U_\delta(p)$ is bounded above by the number of bounding lines that intersect $U_{\varepsilon_0}(p)$. In particular, as $\delta \to 0$, the number of bounding lines intersecting $U_\delta(p)$ either stays constant or decreases.

As there are only finitely many bounding lines that intersect $U_{\varepsilon_0}(p)$, there exists some $\varepsilon < \varepsilon_0$ so that all bounding lines that intersect $U_\varepsilon(p)$ actually pass through p. This is the crucial point at which we make use of the local finiteness of the collection $\mathcal{H}$.

Let $H_1, \ldots, H_n$ be the closed half-planes in $\mathcal{H}$ whose bounding lines contain p, and consider their intersection. As P is nondegenerate, P is not contained in a hyperbolic line, and so no two of these closed half-planes can have the same bounding line.

The n bounding lines break the hyperbolic disc $U_\varepsilon(p)$ into $2n$ wedge-shaped regions. The intersection $\cap_{k=1}^n H_k$ is one of these wedge-shaped regions. An illustration of this phenomenon in the Poincaré disc $\mathbb{D}$ with the vertex $p = 0$ is given in Figure 5.6.

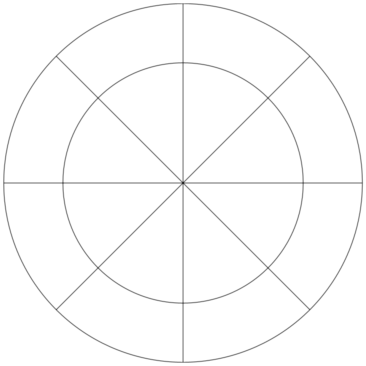

Figure 5.6: Wedges of a hyperbolic disc

Note that there are necessarily two half-planes H_j and H_m in the collection $\mathcal{H}$ so that $\cap_{k=1}^n H_k = H_j \cap H_m$.

In particular, the vertex p is the point of intersection of the two bounding lines ℓ_j and ℓ_m, and the two sides of P that contain p are the sides of P contained in ℓ_j and ℓ_m. This completes the proof of Lemma 5.9. **QED**

The proof of Lemma 5.9 shows that there exists a very good local picture of the structure of the boundary of a hyperbolic polygon P. In fact, given a hyperbolic polygon, we can make use of this proof to construct a canonical locally finite collection of closed half-planes $\mathcal{H}$ whose intersection is P.

Namely, let P be a hyperbolic polygon in the hyperbolic plane. Construct a collection $\mathcal{H}$ of closed half-planes as follows: Enumerate the sides of P as

$s_1, \ldots, s_k, \ldots$. For each s_k, let ℓ_k be the hyperbolic line that contains s_k, and let H_k be the closed half-plane determined by ℓ_k that contains P. Then $\mathcal{H} = \{H_k\}$ is a locally finite collection of closed half-planes, and

$$P = \cap_{H \in \mathcal{H}} H.$$

One consequence of this analysis is that each vertex v of a hyperbolic polygon P is the intersection of two adjacent sides of P. In particular, we can measure the *interior angle* inside P at v.

Definition 5.10

Let P be a hyperbolic polygon, and let v be a vertex of P that is the intersection of two sides s_1 and s_2 of P. Let ℓ_k be the hyperbolic line containing s_k. The union $\ell_1 \cup \ell_2$ divides the hyperbolic plane into four components, one of which contains P. The *interior angle* of P at v is the angle between ℓ_1 and ℓ_2, measured in the component of the complement of $\ell_1 \cup \ell_2$ containing P.

Let P be a hyperbolic polygon. With the definition of vertex we have given, a vertex of P cannot lie inside a side of P, and so any interior angle of P lies in the range $(0, \pi)$.

We can relax this definition of vertex a bit.

Definition 5.11

A hyperbolic polygon P in the hyperbolic plane has an *ideal vertex* at v if there are two adjacent sides of P that are either closed hyperbolic rays or hyperbolic lines and that share v as an endpoint at infinity.

Let P be a hyperbolic polygon with an ideal vertex. Then the interior angle of P at the ideal vertex is 0, which is easy to see if we look in the upper half-plane $\mathbb{H}$. If the hyperbolic polygon P in $\mathbb{H}$ has an ideal vertex at v, and if s_1 and s_2 are the sides of P sharing the ideal vertex at v, then the circles in $\overline{\mathbb{C}}$ containing s_1 and s_2 are tangent at v, and hence the angle between them is 0. See Figure 5.7 for a hyperbolic polygon in $\mathbb{H}$ with an ideal vertex at ∞.

We close this section by discussing some basic types of hyperbolic polygons. We begin with the following definition, which we use to restrict the class of hyperbolic polygons we will work with.

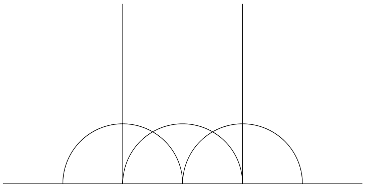

Figure 5.7: A hyperbolic polygon with an ideal vertex at ∞

Definition 5.12

A finite-sided polygon P in the hyperbolic plane is *reasonable* if P does not contain an open half-plane.

Note that if P is a reasonable hyperbolic polygon with sides $s_1, \ldots, s_m$ (labelled counterclockwise), then each adjacent pair of sides s_j and s_{j+1} (where $s_{m+1} = s_1$) shares either a vertex or an ideal vertex. In particular, the sum of the number of vertices of P and the number of ideal vertices of P is equal to the number of sides of P.

For a general finite-sided hyperbolic polygon Q, the best that can be said is that sum of the number of vertices of Q and the number of ideal vertices Q is bounded above by the number of sides of Q, as the endpoint at infinity of a side of a polygon need not be an ideal vertex. The example to keep in mind is a closed hyperbolic half-plane, which has one side and no vertices or ideal vertices.

Now, consider compact hyperbolic polygons. As a compact hyperbolic polygon P is necessarily bounded (by the definition of compactness), as P necessarily has only many finite sides (by the local finiteness of the collection of half-planes whose intersection is P), and as hyperbolic rays and hyperbolic lines are not bounded, all sides of P are closed hyperbolic line segments. In particular, a compact hyperbolic polygon has no ideal vertices, and the number of vertices equals the number of sides. Hence, a compact hyperbolic polygon is necessarily reasonable. In fact, a bit more is true.

Exercise 5.8

Let P be a compact hyperbolic polygon. Prove that P is the convex hull of its vertices.

Although we focus the bulk of our attention on compact hyperbolic polygons, noncompact reasonable hyperbolic polygons will play an important role in the later sections of this chapter.

A *hyperbolic n-gon* is a reasonable hyperbolic polygon with n sides. A compact hyperbolic n-gon is *regular* if its sides have equal length and if its interior angles are equal. For each integer $n \geq 3$, an *ideal n-gon* is a reasonable hyperbolic polygon P that has n sides and n ideal vertices.

As in the Euclidean plane, several hyperbolic polygons have particular names. A *hyperbolic triangle* is a reasonable hyperbolic polygon with three sides. See Figure 5.8 for an ideal hyperbolic triangle and a three-sided hyperbolic polygon that is not a hyperbolic triangle. A *hyperbolic quadrilateral* is a reasonable hyperbolic polygon with four sides, and a *hyperbolic rhombus* is a hyperbolic quadrilateral whose sides have equal length.

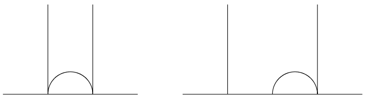

Figure 5.8: An ideal hyperbolic triangle and a three-sided hyperbolic polygon

A *hyperbolic parallelogram* is a hyperbolic quadrilateral whose opposite sides are contained in parallel or ultraparallel hyperbolic lines. Note that, as parallelism is a much different condition in the hyperbolic plane than it is in the Euclidean plane, there is a much greater variety of possible hyperbolic parallelograms than there are Euclidean parallelograms.

Exercise 5.9

For $s > 2$, let Q_s be the hyperbolic quadrilateral in $\mathbb{H}$ with vertices $x_1 = i - 1$, $x_2 = 2i - 1$, $x_3 = i + 1$, and $x_4 = si + 1$. Determine the values of s for which Q_s is a hyperbolic parallelogram.

Exercise 5.10

Let P be an ideal polygon, and let $\{p_1, \ldots, p_k\}$ be its ideal vertices. Prove that $P = \text{conv}(\{p_1, \ldots, p_k\})$.

Exercise 5.11

Let T be a hyperbolic triangle in $\mathbb{H}$ with sides A, B, and C. For any point $x \in A$, prove that

$$\mathrm{d}_{\mathbb{H}}(x, B \cup C) \le \ln(1 + \sqrt{2}).$$

Exercise 5.12

Let A be a hyperbolic circle in the Poincaré disc $\mathbb{D}$ with the property that there exists a hyperbolic ideal triangle T circumscribing A. Show that, for every point $z \in \mathbb{S}^1$, there exists a hyperbolic ideal triangle T_z with one ideal vertex at z circumscribing A.

5.3 The Definition of Hyperbolic Area

In addition to those we have already mentioned, one of the nice properties of hyperbolic convex sets in general, and hyperbolic polygons in particular, is that it is easy to calculate their hyperbolic area. But first, we need to define hyperbolic area. For now, we work in the upper half-plane model $\mathbb{H}$.

Recall that in $\mathbb{H}$, the hyperbolic length of a piecewise C^1 path, and from this the hyperbolic distance between a pair of points, is calculated by integrating the hyperbolic element of arc-length $\frac{1}{\mathrm{Im}(z)}|\mathrm{d}z|$ along the path. The hyperbolic area of a set X in $\mathbb{H}$ is given by integrating the square of the hyperbolic element of arc-length over the set.

Definition 5.13

The *hyperbolic area* $\mathrm{area}_{\mathbb{H}}(X)$ of a set X in $\mathbb{H}$ is given by the integral

$$\mathrm{area}_{\mathbb{H}}(X) = \int_X \frac{1}{\mathrm{Im}(z)^2}\,\mathrm{d}x\,\mathrm{d}y = \int_X \frac{1}{y^2}\,\mathrm{d}x\,\mathrm{d}y,$$

where $z = x + iy$.

For example, consider the region X in $\mathbb{H}$ that is bounded by the three Euclidean lines $\{z \in \mathbb{H} \,|\, \mathrm{Re}(z) = -1\}$, $\{z \in \mathbb{H} \,|\, \mathrm{Re}(z) = 1\}$, and $\{z \in \mathbb{H} \,|\, \mathrm{Im}(z) = 1\}$. Note that as $\{z \in \mathbb{H} \,|\, \mathrm{Im}(z) = 1\}$ is not contained in a hyperbolic line, the region X is not a hyperbolic polygon, although it is convex.

The hyperbolic area of X is then

$$\mathrm{area}_{\mathbb{H}}(X) = \int_X \frac{1}{y^2}\, \mathrm{d}x\, \mathrm{d}y = \int_{-1}^{1} \int_{1}^{\infty} \frac{1}{y^2}\, \mathrm{d}y\, \mathrm{d}x = \int_{-1}^{1} \mathrm{d}x = 2.$$

Exercise 5.13

For $s > 0$, let X_s be the region in $\mathbb{H}$ bounded by the three Euclidean lines $\{z \in \mathbb{H} \,|\, \mathrm{Re}(z) = -1\}$, $\{z \in \mathbb{H} \,|\, \mathrm{Re}(z) = 1\}$, and $\{z \in \mathbb{H} \,|\, \mathrm{Im}(z) = s\}$. Calculate the hyperbolic area $\mathrm{area}_{\mathbb{H}}(X)$ of X_s.

In our discussion of hyperbolic lengths of piecewise C^1 paths, we actually derived the hyperbolic element of arc-length under the assumption that it was invariant under $\mathrm{M\ddot{o}b}(\mathbb{H})$. It then followed immediately that hyperbolic length was naturally invariant under $\mathrm{M\ddot{o}b}(\mathbb{H})$.

However, as we will see in Exercise 5.15, we cannot derive the formula for hyperbolic area by assuming invariance under the action of $\mathrm{M\ddot{o}b}(\mathbb{H})$, as the group of transformations of $\mathbb{H}$ preserving hyperbolic area is much larger than $\mathrm{M\ddot{o}b}(\mathbb{H})$. So, we spend the remainder of this section giving a direct proof that hyperbolic area is invariant under the action of $\mathrm{M\ddot{o}b}(\mathbb{H}) = \mathrm{Isom}(\mathbb{H}, \mathrm{d}_{\mathbb{H}})$.

Theorem 5.14

Hyperbolic area in $\mathbb{H}$ is invariant under the action of $\mathrm{M\ddot{o}b}(\mathbb{H})$. That is, if X be a set in $\mathbb{H}$ whose hyperbolic area $\mathrm{area}_{\mathbb{H}}(X)$ is defined and if A is an element of $\mathrm{M\ddot{o}b}(\mathbb{H})$, then

$$\mathrm{area}_{\mathbb{H}}(X) = \mathrm{area}_{\mathbb{H}}(A(X)).$$

Proof

The proof of Theorem 5.14 is an application of the change of variables theorem from multivariable calculus, which we recall here. Let $F : \mathbb{R}^2 \to \mathbb{R}^2$ be a differentiable function, which we write as

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

and consider its derivative DF, written in matrix form as

$$DF(x,y) = \left(\begin{array}{cc} \frac{\partial f}{\partial x}(x,y) & \frac{\partial g}{\partial x}(x,y) \\ \frac{\partial f}{\partial y}(x,y) & \frac{\partial g}{\partial y}(x,y) \end{array} \right).$$

The change of variables theorem states that, under fairly mild conditions on a set X in $\mathbb{R}^2$ and a function h on X, we have

$$\int_{F(X)} h(x,y)\, \mathrm{d}x\, \mathrm{d}y = \int_X h \circ F(x,y)\, |\det(DF)|\, \mathrm{d}x\, \mathrm{d}y.$$

We do not give the most general statement of the conditions for the change of variables theorem. For our purposes, it suffices to note that the change of variables theorem applies to convex subsets X of $\mathbb{H}$ and to continuous functions h.

We begin by applying the change of variables theorem to an element A of $\text{Möb}^+(\mathbb{H})$. We first rewrite A in terms of x and y as

$$\begin{aligned} A(z) = \frac{az+b}{cz+d} &= \frac{(az+b)(c\overline{z}+d)}{(cz+d)(c\overline{z}+d)} \\ &= \frac{acx^2+acy^2+bd+bcx+adx}{(cx+d)^2+c^2y^2} + i\frac{y}{(cx+d)^2+c^2y^2}, \end{aligned}$$

where a, b, c, $d \in \mathbb{R}$ and $ad - bc = 1$.

So, consider the function $A : \mathbb{H} \to \mathbb{H}$ given by

$$A(x,y) = \left(\frac{acx^2+acy^2+bd+bcx+adx}{(cx+d)^2+c^2y^2}, \frac{y}{(cx+d)^2+c^2y^2} \right).$$

Calculating, we see that

$$DA(x,y) = \left(\begin{array}{cc} \frac{(cx+d)^2-c^2y^2}{((cx+d)^2+c^2y^2)^2} & \frac{2cy(cx+d)}{((cx+d)^2+c^2y^2)^2} \\ \frac{-2cy(cx+d)}{((cx+d)^2+c^2y^2)^2} & \frac{(cx+d)^2-c^2y^2}{((cx+d)^2+c^2y^2)^2} \end{array} \right).$$

In particular, we have that

$$\det(DA(x,y)) = \frac{1}{((cx+d)^2+c^2y^2)^2}.$$

For the calculation of hyperbolic area in $\mathbb{H}$, we integrate the function $h(x,y) = \frac{1}{y^2}$, and so we need to calculate the composition

$$h \circ A(x,y) = \frac{((cx+d)^2+c^2y^2)^2}{y^2}.$$

Hence, the change of variables theorem yields that

$$\begin{aligned}\mathrm{area}_{\mathbb{H}}(A(X)) &= \int_{A(X)} \frac{1}{y^2}\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_X h\circ A(x,y)\,|\det(DA)|\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_X \frac{((cx+d)^2+c^2y^2)^2}{y^2}\frac{1}{((cx+d)^2+c^2y^2)^2}\,\mathrm{d}x\,\mathrm{d}y \\ &= \int_X \frac{1}{y^2}\,\mathrm{d}x\,\mathrm{d}y = \mathrm{area}_{\mathbb{H}}(X),\end{aligned}$$

as desired.

To complete the proof of Theorem 5.14, we need only show that hyperbolic area is invariant under $B(z) = -\overline{z}$, which is the content of the following exercise.

Exercise 5.14

Use the change of variables theorem to prove that hyperbolic area in $\mathbb{H}$ is invariant under $B(z) = -\overline{z}$.

This completes the proof of Theorem 5.14. **QED**

As mentioned earlier in this section, unlike in the case of hyperbolic length, in which Möb($\mathbb{H}$) is exactly the group of transformations of $\mathbb{H}$ preserving hyperbolic length, there are transformations of $\mathbb{H}$ that preserve hyperbolic area but that do not lie in Möb($\mathbb{H}$).

Exercise 5.15

Consider the homeomorphism f of $\mathbb{H}$ given by $f(z) = z + \mathrm{Im}(z)$. Use the change of variables theorem to prove that f preserves hyperbolic area. Show further that f is not an element of Möb($\mathbb{H}$).

Exercise 5.16

Let $F : \mathbb{H} \to \mathbb{H}$ have the form $F(x, y) = (x, g(x, y))$. Determine the conditions on $g(x, y)$ which imply that F preserves hyperbolic area. Give a geometric interpretation of the result of your calculation.

Although we do not prove it, we note that this definition of hyperbolic area makes sense for every convex set in $\mathbb{H}$, and for many nonconvex sets. We do not address the general question of determining the sets in $\mathbb{H}$ for which this definition of hyperbolic area makes sense.

In the same way that we defined the hyperbolic element of area in $\mathbb{H}$, this entire discussion can be carried out in any of the other models of the hyperbolic plane, such as the Poincaré disc model $\mathbb{D}$.

In the Poincaré disc $\mathbb{D}$, there are two natural coordinate systems that come from the fact that $\mathbb{D}$ is a subset of $\mathbb{C}$, namely, the standard cartesian coordinates and polar coordinates. In the cartesian coordinates x and y, the hyperbolic area of a set X in $\mathbb{D}$ is written

$$\text{area}_{\mathbb{D}}(X) = \int_X \frac{4}{(1-|z|^2)^2}\,\mathrm{d}x\,\mathrm{d}y = \int_X \frac{4}{(1-x^2-y^2)^2}\,\mathrm{d}x\,\mathrm{d}y.$$

In polar coordinates, using the standard conversion $x = r\cos(\theta)$ and $y = r\sin(\theta)$ from cartesian to polar coordinates, this integral becomes

$$\text{area}_{\mathbb{D}}(X) = \int_X \frac{4r}{(1-r^2)^2}\,\mathrm{d}r\,\mathrm{d}\theta.$$

Exercise 5.17

Given $s > 0$, let D_s be the open hyperbolic disc in $\mathbb{D}$ with hyperbolic centre 0 and hyperbolic radius s. Show that the hyperbolic area $\text{area}_{\mathbb{D}}(D_s)$ of D_s is

$$\text{area}_{\mathbb{D}}(D_s) = 4\pi\sinh^2\left(\frac{1}{2}s\right).$$

Exercise 5.18

Let D_s be as defined in Exercise 5.17. Describe the behaviour of the quantity

$$q_{\mathbb{D}}(s) = \frac{\text{length}_{\mathbb{D}}(S_s)}{\text{area}_{\mathbb{D}}(D_s)}.$$

Compare the behaviour of $q_{\mathbb{D}}(s)$ with the corresponding quantity $q_{\mathbb{C}}$ calculated using a Euclidean circle and a Euclidean disc in $\mathbb{C}$.

We close this section with the following observation. Although we do not give a proof of it, there is a general formula relating the hyperbolic length $\mathrm{length}_{\mathbb{D}}(C)$ of a simple closed curve C in $\mathbb{D}$ and the hyperbolic area $\mathrm{area}_{\mathbb{D}}(D)$ of the region D in $\mathbb{D}$ bounded by C. Specifically,

$$[\mathrm{length}_{\mathbb{D}}(C)]^2 - 4\pi \,\mathrm{area}_{\mathbb{D}}(D) - [\mathrm{area}_{\mathbb{D}}(D)]^2 \geq 0.$$

Note that in $\mathbb{D}$, it follows from Exercise 4.4 and Exercise 5.17 that the minimum of 0 is achieved when C is a hyperbolic circle and D is the hyperbolic disc bounded by C. However, we cannot conclude solely from this calculation that any region in $\mathbb{D}$ that achieves the minimum of 0 in this inequality is in fact a hyperbolic disc.

Such an inequality is called an *isoperimetric inequality*, as it can be viewed as an equation describing the maximum area of all regions bounded by simple closed curves of a fixed length. This sort of inequality is not specific to the hyperbolic plane; variants hold for a wide range of spaces. For more information about isoperimetric inequalities in general, the interested reader is referred to the encyclopedic work of Burago and Zalgaller [13] and the references contained therein.

5.4 Area and the Gauss–Bonnet Formula

Now that we have shown that hyperbolic area in $\mathbb{H}$ is invariant under the action of $\mathrm{Möb}(\mathbb{H})$, we can more easily calculate the hyperbolic area of relatively simple sets in the hyperbolic plane, such as hyperbolic polygons. We begin by considering hyperbolic triangles.

One approach would be to proceed by direct calculation. That is, for a hyperbolic triangle P, we would write explicit expressions for the Euclidean lines and Euclidean circles containing the sides of P and use these as the limits of integration to calculate the hyperbolic area of P. However, even for a specific hyperbolic triangle, this approach is not effective and is far too unwieldy to use to derive the formula for the hyperbolic area of a general hyperbolic triangle.

Another approach to try is to express our given hyperbolic triangle somehow in terms of hyperbolic triangles whose hyperbolic areas are significantly easier to calculate. We take this approach.

We begin with a simple example. Consider a hyperbolic triangle P with one ideal vertex v_1, and with two other vertices v_2 and v_3, which might or might not be ideal vertices. Let ℓ_{jk} be the hyperbolic line determined by v_j and v_k.

We now make use of the transitivity properties of Möb($\mathbb{H}$) as described in Section 2.9. Namely, let γ be an element of Möb($\mathbb{H}$) that takes v_1 to ∞ and that takes ℓ_{23} to the hyperbolic line contained in the unit circle, so that $v_2 = e^{i\varphi}$ and $v_3 = e^{i\theta}$, where $0 \leq \theta < \varphi \leq \pi$. (We allow $\theta = 0$ and $\varphi = \pi$ to allow for the possibility that one or both of v_2 and v_3 is an ideal vertex.) See Figure 5.9.

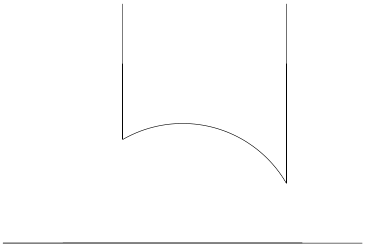

Figure 5.9: The case of one ideal vertex

As hyperbolic area is invariant under the action of Möb($\mathbb{H}$), we may thus assume P to be the hyperbolic triangle with an ideal vertex at ∞, and with two other vertices at $e^{i\theta}$ and $e^{i\varphi}$, where $0 \leq \theta < \varphi \leq \pi$. As P has at least one ideal vertex, it is not compact, but we can still easily calculate its hyperbolic area.

Calculating, we see that

$$\text{area}_{\mathbb{H}}(P) = \int_P \frac{1}{y^2}\,\mathrm{d}x\,\mathrm{d}y = \int_{\cos(\varphi)}^{\cos(\theta)} \int_{\sqrt{1-x^2}}^{\infty} \frac{1}{y^2}\mathrm{d}y\,\mathrm{d}x = \int_{\cos(\varphi)}^{\cos(\theta)} \frac{1}{\sqrt{1-x^2}}\mathrm{d}x.$$

Making the substitution $x = \cos(w)$, so that $\mathrm{d}x = -\sin(w)\,\mathrm{d}w$, this calculation becomes

$$\int_{\cos(\varphi)}^{\cos(\theta)} \frac{1}{\sqrt{1-x^2}}\mathrm{d}x = \int_{\varphi}^{\theta} -\mathrm{d}w = \varphi - \theta.$$

At this point, with a hint of foreshadowing, we observe that the interior angle of P at the ideal vertex $v_1 = \infty$ is $\alpha_1 = 0$, the interior angle at the vertex $v_2 = e^{i\theta}$ is $\alpha_2 = \theta$, and the interior angle at the vertex $v_3 = e^{i\varphi}$ is $\alpha_3 = \pi - \varphi$. Hence, we have proven the following proposition.

Proposition 5.15

Let P be a hyperbolic triangle with one ideal vertex, and let α_2 and α_3 be the interior angles at the other two vertices, which might or might not be ideal vertices. Then,

$$\text{area}_{\mathbb{H}}(P) = \pi - (\alpha_2 + \alpha_3).$$

One consequence of Proposition 5.15 is that the hyperbolic area of an ideal triangle in $\mathbb{H}$ is π, which follows from the observation that the interior angle at each ideal vertex of an ideal triangle is 0.

Suppose now that P is a compact hyperbolic triangle with vertices v_1, v_2, and v_3. Let α_k be the interior angle of P at v_k. Let ℓ be the hyperbolic ray from v_1 passing through v_2, and let x be the endpoint at infinity of ℓ. See Figure 5.10.

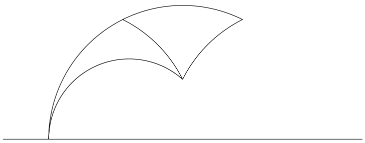

Figure 5.10: The case of no ideal vertices

The hyperbolic triangle T with vertices v_1, v_3, and x has one ideal vertex at x and two nonideal vertices at v_1 and v_3. The interior angle of T at v_1 is α_1 and the interior angle of T at v_3 is $\delta > \alpha_3$. So, by Proposition 5.15, the hyperbolic area of T is

$$\text{area}_{\mathbb{H}}(T) = \pi - (\alpha_1 + \delta).$$

The hyperbolic triangle T' with vertices v_2, v_3, and x has one ideal vertex at x and two nonideal vertices at v_2 and v_3. The interior angle of T' at v_2 is $\pi - \alpha_2$, and the interior angle of T' at v_3 is $\delta - \alpha_3$. So, the hyperbolic area of T' is

$$\text{area}_{\mathbb{H}}(T') = \pi - (\pi - \alpha_2 + \delta - \alpha_3).$$

As T is the union of T' and P, and as T' and P overlap only along a side, we see that

$$\text{area}_{\mathbb{H}}(T) = \text{area}_{\mathbb{H}}(T') + \text{area}_{\mathbb{H}}(P).$$

Substituting in the calculations of the previous two paragraphs, we see that

$$\begin{aligned}
\text{area}_{\mathbb{H}}(P) &= \text{area}_{\mathbb{H}}(T) - \text{area}_{\mathbb{H}}(T') \\
&= \pi - (\alpha_1 + \delta) - (\pi - (\pi - \alpha_2 + \delta - \alpha_3)) \\
&= \pi - (\alpha_1 + \alpha_2 + \alpha_3).
\end{aligned}$$

This completes the proof of the following theorem.

Theorem 5.16

Let P be a hyperbolic triangle with interior angles α, β, and γ. Then,

$$\text{area}_{\mathbb{H}}(P) = \pi - (\alpha + \beta + \gamma).$$

Theorem 5.16 is known as the *Gauss–Bonnet formula*.

Exercise 5.19

Consider the hyperbolic triangle P in $\mathbb{H}$ with vertices i, $4+i$, and $2+2i$. Calculate the hyperbolic area of P by determining the three interior angles of P.

We can generalize Theorem 5.16 to all reasonable hyperbolic polygons.

Theorem 5.17

Let P be a reasonable hyperbolic polygon with vertices and ideal vertices $v_1, \ldots, v_n$. Let α_k be the interior angle at v_k. Then,

$$\text{area}_{\mathbb{H}}(P) = (n-2)\pi - \sum_{k=1}^{n} \alpha_k.$$

Proof

We prove Theorem 5.17 by decomposing P into hyperbolic triangles, using Theorem 5.16 to calculate the hyperbolic area of each hyperbolic triangle in this decomposition, and then summing to get the hyperbolic area of P.

Choose a point x in the interior of P. As P is convex, the hyperbolic line segment (or hyperbolic ray, in the case in which v_k is an ideal vertex) ℓ_k joining x to v_k is contained in P. The hyperbolic line segments $\ell_1, \ldots, \ell_n$ break P into n triangles $T_1, \ldots, T_n$. See Figure 5.11.

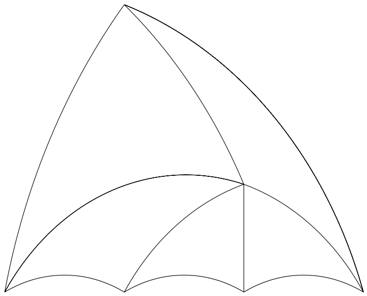

Figure 5.11: Decomposing a hyperbolic pentagon into hyperbolic triangles

Label these hyperbolic triangles at $T_1, \ldots, T_n$, so that T_k has vertices x, v_k, and v_{k+1} for $1 \leq k \leq n$, where in a slight abuse of notation, we set $v_{n+1} = v_1$ and $T_{n+1} = T_1$.

Let μ_k be the interior angle of T_k at x, and note that

$$\sum_{k=1}^{n} \mu_k = 2\pi.$$

Let β_k be the interior angle of T_k at v_k, and let δ_k be the interior angle of T_k at v_{k+1}. As both T_k and T_{k+1} have a vertex at v_{k+1}, we see that

$$\alpha_{k+1} = \delta_k + \beta_{k+1}.$$

Applying Theorem 5.16 to T_k yields that

$$\mathrm{area}_{\mathbb{H}}(T_k) = \pi - (\mu_k + \beta_k + \delta_k).$$

As the union $T_1 \cup \cdots \cup T_n$ is equal to P and as the hyperbolic triangles $T_1, \ldots, T_n$ overlap only along on their sides, we have that

$$\begin{aligned} \mathrm{area}_{\mathbb{H}}(P) &= \sum_{k=1}^{n} \mathrm{area}_{\mathbb{H}}(T_k) = \sum_{k=1}^{n} [\pi - (\mu_k + \beta_k + \delta_k)] \\ &= n\pi - \left[\sum_{k=1}^{n} \mu_k + \sum_{k=1}^{n} \beta_k + \sum_{k=1}^{n} \delta_k \right]. \end{aligned}$$

As $\alpha_{k+1} = \delta_k + \beta_{k+1}$ for each k, we have that

$$\sum_{k=1}^{n} \beta_k + \sum_{k=1}^{n} \delta_k = \sum_{k=1}^{n} \alpha_k.$$

Hence,

$$\text{area}_{\mathbb{H}}(P) = \sum_{k=1}^{n} \text{area}_{\mathbb{H}}(T_k) = (n-2)\pi - \sum_{k=1}^{n} \alpha_k.$$

This completes the proof of Theorem 5.17. **QED**

Note that we could have taken the point x around which we decomposed P to be a point on a side of P, or to be a vertex of P. In either case, the particulars of the calculation would be slightly different, but we would still obtain Theorem 5.17 in the end. In the former case, we would decompose P into $n-1$ hyperbolic triangles, and the sum of the interior angles of the hyperbolic triangles at x would equal π. In the latter case, we would decompose P into $n-2$ hyperbolic triangles, and the sum of the interior angles of the hyperbolic triangles at x would equal the interior angle of P at x.

We note here that although the specifics of the calculations are different, Theorem 5.17 holds in any model of the hyperbolic plane.

5.5 Applications of the Gauss–Bonnet Formula

In this section, we describe two applications of Theorem 5.17 in the hyperbolic plane. One application is positive, in that it asserts the existence of a large number and variety of different regular compact hyperbolic polygons. The other application is negative, in that it asserts the nonexistence of a certain type of transformation of the hyperbolic plane.

We begin with a fact about the Euclidean plane $\mathbb{C}$, namely, that for each integer $n \geq 3$, there exists only one regular Euclidean n-gon, up to scaling, rotation, and translation.

Here is one construction of such a regular Euclidean n-gon P_n in $\mathbb{C}$. Start by choosing a basepoint x in $\mathbb{C}$, and let $\ell_1, \ldots, \ell_n$ be n Euclidean rays from x, where the angle between consecutive rays is $\frac{2\pi}{n}$. Choose some $r > 0$, and for each $1 \leq k \leq n$, consider the point y_k on ℓ_k that is Euclidean distance r from x. These points $y_1, \ldots, y_n$ are the vertices of a regular Eulidean n-gon P_n.

To see that P_n is unique up to scaling, rotation, and translation, we repeat the construction. That is, choose a different basepoint x' in $\mathbb{C}$. Let $\ell'_1, \ldots, \ell'_n$ be n Euclidean rays from x', where the angle between consecutive rays is $\frac{2\pi}{n}$.

Choose some $r' > 0$, and let y'_k be the point on ℓ'_k that is Euclidean distance r' from x'. Then, the points $y'_1, \ldots, y'_n$ are the vertices of a regular Euclidean n-gon P'_n.

We now construct a transformation of $\mathbb{C}$ that takes P_n to P'_n. Let θ be the angle between ℓ_1 and the positive real axis, and let θ' be the angle between ℓ'_1 and the positive real axis. Then the homeomorphism B of $\mathbb{C}$ given by

$$B(z) = e^{i(\theta' - \theta)}\, \frac{r'}{r}\, (z - x + x')$$

is the composition of a rotation, a dilation, and a translation of $\mathbb{C}$ that satisfies $B(P_n) = P'_n$.

In particular, the interior angles of P_n at its vertices depend only on the number of sides n, and not on the choice of the basepoint x or the Euclidean rays ℓ_k or the Euclidean distance r of the vertices of P_n from x. In fact, the interior angle at a vertex of P_n is $\frac{n-2}{n}\pi$.

In the hyperbolic plane, the situation is considerably different.

Proposition 5.18

For each $n \geq 3$ and for each α in the interval $(0, \frac{n-2}{n}\pi)$, there is a compact regular hyperbolic n-gon whose interior angle is α.

Proof

We work in the Poincaré disc $\mathbb{D}$ and start with the same construction just given for regular Euclidean n-gons in $\mathbb{C}$. Given $n \geq 3$, consider the n hyperbolic rays $\ell_0, \ldots, \ell_{n-1}$ from 0, where ℓ_k is the hyperbolic ray determined by 0 and $p_k = \exp\left(\frac{2\pi i}{n} k\right)$.

For each $0 < r < 1$, the n points $rp_0 = r, \ldots, rp_{n-1} = r \exp\left(\frac{2\pi i}{n}(n-1)\right)$ in $\mathbb{D}$ are the vertices of a regular hyperbolic n-gon $P_n(r)$. Specifically, construct $P_n(r)$ as the intersection of a locally finite collection of closed half-planes as follows: For $0 \leq k \leq n-1$, let ℓ_k be the hyperbolic line passing through p_k and p_{k+1}, where $p_n = p_0$. Let H_k be the closed half-plane determined by ℓ_k that contains 0, and note that

$$P_n(r) = \cap_{k=0}^{n-1} H_k.$$

Let s_k be the side of $P_n(r)$ contained in ℓ_k.

To see that $P_n(r)$ is regular, we use the elliptic Möbius transformation

$$m(z) = \exp\left(\frac{2\pi i}{n}\right) z,$$

which is contained in $\text{Möb}(\mathbb{D})$. For each $0 \leq k \leq n-1$, we have that $m^k(rp_0) = rp_k$ and that $m^k(\ell_0) = \ell_k$. Hence, $m^k(P_n(r)) = P_n(r)$. Moreover, as $m^k(rp_1) = rp_{k+1}$ as well, we see that $m^k(s_0) = s_k$. As each m^k takes the sides of $P_n(r)$ to $P_n(r)$, the hyperbolic lengths of the sides of $P_n(r)$ are equal.

As $m^k(s_0) = s_k$ and $m^k(s_{n-1}) = s_{k-1}$, we also have that m^k takes the two sides s_{n-1} and s_0 of $P_n(r)$ that intersect at rp_0 to the two sides s_{k-1} and s_k of $P_n(r)$ that intersect at rp_k. In particular, the interior angles of $P_n(r)$ at any two vertices are equal.

For $0 < r < 1$, let $\alpha(r)$ denote the interior angle of $P_n(r)$ at $r = rp_0$. We now analyze the behaviour of $\alpha(r)$ as r varies. We note that $\alpha(r)$ is a continuous function of r, by the calculation in Exercise 5.20.

Exercise 5.20

Express the interior angle of $P_n(r)$ at $r = rp_0$ in terms of n and r. Fix n. Conclude that $\alpha(r)$ is a continuous function of r.

Theorem 5.17 yields that the hyperbolic area of $P_n(r)$ is

$$\text{area}_{\mathbb{D}}(P_n(r)) = (n-2)\pi - \sum_{k=0}^{n-1} \alpha(r) = (n-2)\pi - n\alpha(r).$$

For each value of r, $0 < r < 1$, the hyperbolic polygon $P_n(r)$ is contained in the hyperbolic disc D_r in $\mathbb{D}$ with hyperbolic centre 0 and Euclidean radius r. (Note that this implies the boundedness and, hence, the compactness, of $P_n(r)$.) The hyperbolic area of D_r is

$$\text{area}_{\mathbb{D}}(D_r) = \frac{2\pi r}{1-r^2}.$$

As $P_n(r)$ is contained in D_r, we have that $\text{area}_{\mathbb{D}}(P_n(r)) \leq \text{area}_{\mathbb{D}}(D_r)$, and so

$$\lim_{r\to 0^+} \text{area}_{\mathbb{D}}(P_n(r)) \leq \lim_{r\to 0^+} \frac{2\pi r}{1-r^2} = 0.$$

Substituting in the expression for $\text{area}_{\mathbb{D}}(P_n(r))$, we see that

$$\lim_{r\to 0^+} [(n-2)\pi - n\alpha(r)] = 0,$$

and so

$$\lim_{r \to 0^+} \alpha(r) = \frac{n-2}{n}\pi.$$

Hence, as $r \to 0^+$, the interior angle of $P_n(r)$ is converging to the interior angle of a regular Euclidean n-gon.

As r increases, we can make two observations, both of which we can get either from Exercise 5.20 or from direct observation.

First, for $0 < s < r < 1$, the vertices of $P_n(s)$ lie in the interior of $P_n(r)$. The convexity of $P_n(r)$ then forces $P_n(s)$ to be contained in $P_n(r)$, and so

$$\text{area}_{\mathbb{D}}(P_n(s)) < \text{area}_{\mathbb{D}}(P_n(r))$$

for $0 < s < r < 1$. In other words, the hyperbolic area of $P_n(r)$ is monotonically increasing in r. As

$$\text{area}_{\mathbb{D}}(P_n(r)) = (n-2)\pi - n\alpha(r),$$

we have that the interior angle $\alpha(r)$ is monotonically decreasing in r.

Second, as $r \to 1^-$, the compact hyperbolic polygon $P_n(r)$ is becoming more and more like the ideal hyperbolic n-gon P_n^∞ with ideal vertices at $p_0 = 1$, $p_1 = \exp(\frac{2\pi i}{n}), \ldots, \ldots, p_{n-1} = \exp(\frac{2\pi i}{n}(n-1))$. In particular, we have that

$$\lim_{r \to 1^-} \text{area}_{\mathbb{D}}(P_n(r)) = \text{area}_{\mathbb{D}}(P_n^\infty).$$

Expressing $\text{area}_{\mathbb{D}}(P_n(r))$ and $\text{area}_{\mathbb{D}}(P_n^\infty)$ in terms of the interior angles of $P_n(r)$ and P_n^∞, respectively, we have that

$$\lim_{r \to 1^-} [(n-2)\pi - n\alpha(r)] = (n-2)\pi,$$

and so

$$\lim_{r \to 1^-} \alpha(r) = 0.$$

Combining these observations, we see that for $n \geq 3$, the interior angle $\alpha(r)$ of the compact regular hyperbolic n-gon $P_n(r)$ lies in the interval $(0, \frac{n-2}{n}\pi)$. Moreover, the monotonicity and continuity of α imply that every number in this interval is the interior angle of one and only one hyperbolic polygon $P_n(r)$. This completes the proof of Proposition 5.18. **QED**

One specific way in which the behaviour of hyperbolic polygons is much different from the behaviour of Euclidean polygons is that, in the Euclidean plane, there is one and only one regular n-gon with all right angles, namely, the square.

However, in the hyperbolic plane, not only do hyperbolic squares not exist, but for each $n \geq 5$, there exists a compact regular hyperbolic n-gon with all right angles. To see that hyperbolic squares do not exist, we use Proposition 5.18 in the case $n = 4$. The interval of possible interior angles of a compact regular hyperbolic 4-gon is $(0, \frac{1}{2}\pi)$. In particular, there is no hyperbolic 4-gon with all right angles.

The proof that there exist compact regular hyperbolic n-gons with all right angles for $n \geq 5$ is left as an exercise.

Exercise 5.21

Prove that for $n \geq 5$, there exists a compact regular hyperbolic n-gon all of whose interior angles are right angles.

In addition to the fact that the interior angle $\alpha(r)$ of the compact regular hyperbolic n-gon $P_n(r)$ is a continuous function of r, we also have that the hyperbolic length of a side of $P_n(r)$ is continuous as a function of r.

Exercise 5.22

Given $0 < r < 1$, explicitly calculate the hyperbolic length of a side of $P_n(r)$ in terms of n and r.

For each $n \geq 5$, Exercise 5.21 gives one compact hyperbolic n-gon with all right angles, namely, the compact regular hyperbolic n-gon with all right angles. In fact, for each $n \geq 5$, there are many nonregular compact hyperbolic n-gons with all right angles, although we do not prove this fact here.

Also, it is possible to construct reasonable hyperbolic polygons with prescribed interior angles that are not necessarily right angles. In fact, the only restriction on the possible internal angles is that the hyperbolic area, as given by the Gauss–Bonnet formula, should be positive. Again, we do not prove this fact here. The interested reader is referred to Beardon [7] for the proof of Theorem 5.19.

Theorem 5.19

Let $\alpha_1, \ldots, \alpha_n$ be a collection of n real numbers in the interval $[0, \pi)$. Then, there exists a hyperbolic n-gon in the hyperbolic plane with interior angles

$\alpha_1, \ldots, \alpha_n$ if and only if

$$\alpha_1 + \cdots + \alpha_n < (n-2)\pi.$$

There is a second application of Theorem 5.17 we consider here. Recall that in the construction of regular Euclidean n-gons given at the beginning of this section, we remarked that even though the Euclidean n-gons constructed were of different area, any two regular Euclidean n-gons are related by a homeomorphism of $\mathbb{C}$ that is the composition of an isometry of $\mathbb{C}$ and a dilation of $\mathbb{C}$.

Here, a *dilation* of $\mathbb{C}$ is a homeomorphism of $\mathbb{C}$ that takes Euclidean lines to Euclidean lines and is conformal, in that it preserves the angles between pairs of lines. Dilations are not isometries, as they do not preserve Euclidean length or area. In fact, every dilation of $\mathbb{C}$ is of the form $f(z) = az + b$ for some $a \in \mathbb{C} - \{0\}$ and $b \in \mathbb{C}$. The function f is a Euclidean isometry if and only if $|a| = 1$, a fact that is essentially contained in the solution to Exercise 3.15.

Definition 5.20

A *hyperbolic dilation* is a homeomorphism of the hyperbolic plane that takes hyperbolic lines to hyperbolic lines and is conformal, in that it preserves the angles between pairs of hyperbolic lines.

As in the case of $\mathbb{C}$, every isometry of the hyperbolic plane is a hyperbolic dilation. However, unlike in the case of $\mathbb{C}$, there are no hyperbolic dilations other than hyperbolic isometries.

The key fact in the proof of this fact, stated as Proposition 5.21, is that because a hyperbolic dilation is conformal and so preserves angles, Theorem 5.17 immediately implies that a hyperbolic dilation preserves the hyperbolic area of a hyperbolic polygon. In particular, if g is a dilation of the hyperbolic plane and if P is a hyperbolic polygon, then the hyperbolic areas of P and of $g(P)$ are equal.

Proposition 5.21

Let f be a hyperbolic dilation of the hyperbolic plane. Then, f is a hyperbolic isometry.

Proof

The proof of Proposition 5.21 is similar in spirit to the proof of Theorem 3.19. The main technical tool used in the proof of Proposition 5.21 is Theorem 5.17.

We work in the Poincaré disc model $\mathbb{D}$ of the hyperbolic plane. Let f be a hyperbolic dilation of $\mathbb{D}$, so that by definition f is a homeomorphism of $\mathbb{D}$ that takes hyperbolic lines to hyperbolic lines and that preserves angles.

We begin by using the transitivity properties of Möb($\mathbb{D}$) to normalize f. First compose f with an element m of Möb($\mathbb{D}$) that takes $f(0)$ to 0. By definition, every element of Möb($\mathbb{D}$) is a hyperbolic isometry, and hence, it is a hyperbolic dilation. Therefore, the composition $m \circ f$ is a hyperbolic dilation of $\mathbb{D}$ that fixes 0.

The hyperbolic dilation $m \circ f$ of $\mathbb{D}$ takes hyperbolic rays from 0 to hyperbolic rays from 0 and preserves angles between hyperbolic rays. So, there exists an element n of Möb($\mathbb{D}$) fixing 0 so that the composition $n \circ m \circ f$ is a hyperbolic dilation of $\mathbb{D}$ that fixes 0 and that takes every hyperbolic ray from 0 to itself. (This element n will be either an elliptic Möbius transformation fixing 0 or the composition of an elliptic Möbius transformation fixing 0 and $C(z) = \overline{z}$.)

Set $g = n \circ m \circ f$. To complete the proof that f is an element of Möb($\mathbb{D}$), we show that g is the identity.

Let z_0 be a point of $\mathbb{D} - \{0\}$. Let ℓ_0 be the hyperbolic ray from 0 passing through z_0. Let ℓ_1 be the hyperbolic ray from 0 making angle $\frac{2\pi}{3}$ with ℓ_0, and let ℓ_2 be the hyperbolic ray from 0 making angle $\frac{4\pi}{3}$ with ℓ_0. Let T be the hyperbolic triangle with vertices $v_0 = z_0$, $v_1 = \exp\left(\frac{2\pi i}{3}\right) z_0$, and $v_2 = \exp\left(\frac{4\pi i}{3}\right) z_0$. Let s_{jk} be the side of T joining v_j to v_k. We consider the image $g(T)$ of T under g.

Set $r = |z_0|$ and $s = |g(z_0)|$, so that $g(z_0) = \frac{s}{r} z_0$. We first show that $g(v_1) = \frac{s}{r} \exp\left(\frac{2\pi i}{3}\right) z_0$ and that $g(v_2) = \frac{s}{r} \exp\left(\frac{4\pi i}{3}\right) z_0$. By our assumptions on g, we have that v_k, and hence $g(v_k)$, lies on the hyperbolic ray ℓ_k, as g takes each hyperbolic ray from 0 to itself. Hence, $g(v_k)$ is a positive real multiple of v_k.

As the angle of intersection of s_{0k} with ℓ_0 is equal to the angle of intersection of s_{0k} with ℓ_k, we have that the angle of intersection of $g(s_{0k})$ with $g(\ell_0) = \ell_0$ is equal to the angle of intersection of $g(s_{0k})$ with $g(\ell_k) = \ell_k$. In particular, the point of intersection of $g(s_{0k})$ with ℓ_0 and the point of intersection of $g(s_{0k})$ with ℓ_k are the same Euclidean distance from the origin. As the point of intersection of $g(s_{0k})$ with ℓ_0 is $g(v_0) = \frac{s}{r} z_0$, we have that $g(v_k) = \frac{s}{r} \exp\left(\frac{2k\pi i}{3}\right) z_0$ for $k = 1$ and 2, as desired.

So, the image $g(T)$ of T under g is the hyperbolic triangle with vertices $g(v_0) = \frac{s}{r} z_0$, $g(v_1) = \frac{s}{r} \exp\left(\frac{2\pi i}{3}\right) z_0$, and $g(v_2) = \frac{s}{r} \exp\left(\frac{4\pi i}{3}\right) z_0$. As g is a hyperbolic dilation, angles between hyperbolic lines are preserved by g, and so the interior angles of T and of $g(T)$ are equal. By Theorem 5.17, we then have that

$$\text{area}_{\mathbb{D}}(T) = \text{area}_{\mathbb{D}}(g(T)).$$

However, if $s = |g(z_0)| > r = |z_0|$, then T is properly contained in $g(T)$ and so $\text{area}_{\mathbb{D}}(T) < \text{area}_{\mathbb{D}}(g(T))$, a contradiction. If $s = |g(z_0)| < r < |z_0|$, then $g(T)$ is properly contained in T, and so $\text{area}_{\mathbb{D}}(T) > \text{area}_{\mathbb{D}}(g(T))$, which is again a contradiction.

Hence, we have that $g(z) = z$ for every point z of $\mathbb{D}$, and so g is the identity. This completes the proof of Proposition 5.21. **QED**

5.6 Trigonometry in the Hyperbolic Plane

Let T be a compact hyperbolic triangle in the hyperbolic plane. As in the case for a Euclidean triangle, there are trigonometric laws in the hyperbolic plane relating the interior angles of T and the hyperbolic lengths of the sides of T.

The way we derive the trigonometric laws in the hyperbolic plane is to link the Euclidean and hyperbolic distances between a pair of points. As the hyperbolic and Euclidean measurement of the angles of T are the same, we may then make use of the Euclidean trigonometric laws. We note here that there are intrinsic ways of deriving the hyperbolic trigonometric laws that do not start from the Euclidean trigonometric laws.

As we saw in Exercise 4.2, the relationship between Euclidean and hyperbolic lengths involves the use of the hyperbolic trigonometric functions. Before going any further, we state some identities involving the hyperbolic trigonometric functions that arise over the course of the section, leaving their verification as an exercise.

Exercise 5.23

Verify each of the following identities.

1. $\cosh^2(x) - \sinh^2(x) = 1;$

2. $2\cosh(x)\sinh(x) = \sinh(2x)$;

3. $\sinh^2(x) = \frac{1}{2}\cosh(2x) - \frac{1}{2}$;

4. $\cosh^2(x) = \frac{1}{2}\cosh(2x) + \frac{1}{2}$;

5. $\sinh^2(x)\cosh^2(y) + \cosh^2(x)\sinh^2(y) = \frac{1}{2}(\cosh(2x)\cosh(2y) - 1)$.

We work in the Poincaré disc model $\mathbb{D}$. Let T be a compact hyperbolic triangle in $\mathbb{D}$ with vertices v_1, v_2, and v_3. Let a, b, and c be the hyperbolic lengths of its sides, and let α, β, and γ be its interior angles, where α is the interior angle at the vertex v_1 opposite the side of hyperbolic length a, β is the interior angle at the vertex v_2 opposite the side of hyperbolic length b, and γ is the interior angle at the vertex v_3 opposite the side of hyperbolic length c.

As the interior angles at the vertices of T and the hyperbolic lengths of the sides of T are invariant under the action of $\text{Möb}(\mathbb{D})$, we may use the transitivity properties of $\text{Möb}(\mathbb{D})$ to assume that $v_1 = 0$, that $v_2 = r > 0$ lies on the positive real axis, and that $v_3 = se^{i\alpha}$, where $0 < \alpha < \pi$. By Exercise 4.2, we have that

$$r = \tanh\left(\frac{1}{2}c\right) \text{ and } s = \tanh\left(\frac{1}{2}b\right).$$

On the one hand, we may apply the Euclidean law of cosines to the Euclidean triangle with vertices v_1, v_2, and v_3 to see that

$$\begin{aligned} &|v_2 - v_3|^2 \\ &\quad = r^2 + s^2 - 2rs\cos(\alpha) \\ &\quad = \tanh^2\left(\frac{1}{2}c\right) + \tanh^2\left(\frac{1}{2}b\right) - 2\tanh\left(\frac{1}{2}c\right)\tanh\left(\frac{1}{2}b\right)\cos(\alpha). \end{aligned}$$

On the other hand, by Proposition 4.3, we have that

$$\begin{aligned} \frac{|v_2 - v_3|^2}{(1 - |v_2|^2)(1 - |v_3|^2)} &= \frac{|v_2 - v_3|^2}{(1 - r^2)(1 - s^2)} \\ &= \sinh^2\left(\frac{1}{2}\mathrm{d}_{\mathbb{D}}(v_2, v_3)\right) = \sinh^2\left(\frac{1}{2}a\right), \end{aligned}$$

and so

$$\begin{aligned} |v_2 - v_3|^2 &= (1 - r^2)(1 - s^2)\sinh^2\left(\frac{1}{2}a\right) \\ &= \operatorname{sech}^2\left(\frac{1}{2}c\right)\operatorname{sech}^2\left(\frac{1}{2}c\right)\sinh^2\left(\frac{1}{2}a\right). \end{aligned}$$

Equating the two expressions for $|v_2 - v_3|^2$, we obtain

$$\begin{aligned}&\operatorname{sech}^2\left(\frac{1}{2}c\right)\operatorname{sech}^2\left(\frac{1}{2}b\right)\sinh^2\left(\frac{1}{2}a\right) = \\ &\qquad \tanh^2\left(\frac{1}{2}c\right) + \tanh^2\left(\frac{1}{2}b\right) - 2\tanh\left(\frac{1}{2}c\right)\tanh\left(\frac{1}{2}b\right)\cos(\alpha).\end{aligned}$$

And now we simplify. Multiplying through by $\cosh^2\left(\frac{1}{2}c\right)\cosh^2\left(\frac{1}{2}b\right)$, we obtain

$$\begin{aligned}&\sinh^2\left(\frac{1}{2}a\right) = \\ &\qquad \sinh^2\left(\frac{1}{2}c\right)\cosh^2\left(\frac{1}{2}b\right) + \sinh^2\left(\frac{1}{2}b\right)\cosh^2\left(\frac{1}{2}c\right) \\ &\qquad -2\sinh\left(\frac{1}{2}c\right)\sinh\left(\frac{1}{2}b\right)\cosh\left(\frac{1}{2}c\right)\cosh\left(\frac{1}{2}b\right)\cos(\alpha).\end{aligned}$$

Using the identities given in Exercise 5.23, this calculation becomes

$$\frac{1}{2}\cosh(a) - \frac{1}{2} = \frac{1}{2}\cosh(b)\cosh(c) - \frac{1}{2} - \frac{1}{2}\sinh(c)\sinh(b)\cos(\alpha),$$

and so we obtain the hyperbolic **law of cosines I:**

$$\cosh(a) = \cosh(b)\cosh(c) - \sinh(c)\sinh(b)\cos(\alpha).$$

Unlike in the Euclidean plane, there are three basic trigonometric laws in the hyperbolic plane. One is the law of cosines I, which we have just derived. The other two, the hyperbolic law of sines and the hyperbolic law of cosines II, are stated below.

law of sines:

$$\frac{\sinh(a)}{\sin(\alpha)} = \frac{\sinh(b)}{\sin(\beta)} = \frac{\sinh(c)}{\sin(\gamma)}.$$

law of cosines II:

$$\cos(\gamma) = -\cos(\alpha)\cos(\beta) + \sin(\alpha)\sin(\beta)\cosh(c).$$

The hyperbolic law of cosines I and the hyperbolic law of sines are the direct analogues of the Euclidean law of cosines and the Euclidean law of sines. In fact, as we have just seen, the proof of the law of cosines I follows fairly quickly from the Euclidean law of cosines and some algebraic manipulation.

In much the same way that the Euclidean law of sines can be derived from the Euclidean law of cosines by algebraic manipulation, the hyperbolic law of sines and the hyperbolic law of cosines II can be derived from the hyperbolic law of cosines I.

Exercise 5.24

Derive the hyperbolic law of cosines II and the hyperbolic law of sines from the hyperbolic law of cosines I.

Exercise 5.25

State and prove the hyperbolic Pythagorean theorem, relating the hyperbolic lengths of the sides of a hyperbolic right triangle.

Exercise 5.26

For $\lambda > 1$, consider the loxodromic transformation $m(z) = \lambda z$. Let A be the Eulidean ray in $\mathbb{H}$ from 0 making angle θ with the positive real axis. Calculate the translation distance of m along A as a function of λ and θ.

Exercise 5.27

Fix $0 < r < 1$. For each $s > 0$, consider the set

$$C_r(s) = \{z \in \mathbb{D} \mid \cosh(\mathrm{d}_{\mathbb{D}}(z, r)) + \cosh(\mathrm{d}_{\mathbb{D}}(z, -r)) = s\}.$$

Describe $C_r(s)$.

The most surprising of the hyperbolic trigonometric laws is the law of cosines II, which states that the hyperbolic length of a side of a hyperbolic triangle is determined by the interior angles of the triangle. In particular, this trigonometric law implies that there is a canonical unit of hyperbolic length, which is unlike length in the Euclidean plane.

For example, consider the compact hyperbolic triangle T with interior angles $\alpha = \frac{1}{2}\pi$, $\beta = \frac{1}{3}\pi$, and $\gamma = \frac{1}{7}\pi$ at its vertices. Let a be the hyperbolic length of the side of T opposite the vertex with angle α, let b be the hyperbolic length of the side of T opposite the vertex with angle β, and let c be the hyperbolic length of the side of T opposite the vertex with angle γ.

By the law of cosines II, the hyperbolic lengths of the three sides of T satisfy

$$\cosh(a) = \frac{\cos(\alpha)+\cos(\beta)\cos(\gamma)}{\sin(\beta)\sin(\gamma)} = \cot\left(\frac{\pi}{3}\right)\cot\left(\frac{\pi}{7}\right) \sim 1.1989;$$

$$\cosh(b) = \frac{\cos(\beta)+\cos(\alpha)\cos(\gamma)}{\sin(\alpha)\sin(\gamma)} = \cos\left(\frac{\pi}{3}\right)\csc\left(\frac{\pi}{7}\right) \sim 1.1524;$$

$$\cosh(c) = \frac{\cos(\gamma)+\cos(\alpha)\cos(\beta)}{\sin(\alpha)\sin(\beta)} = \cos\left(\frac{\pi}{7}\right)\csc\left(\frac{\pi}{3}\right) \sim 1.0403.$$

We pause to insert a note about actually solving for hyperbolic lengths. To solve

$$\cosh(a) = \frac{1}{2}\left(e^a + e^{-a}\right) = x,$$

we see by the quadratic formula (after multiplying through by e^a) that e^a satisfies

$$e^a = x \pm \sqrt{x^2-1},$$

and so either

$$a = \log(x+\sqrt{x^2-1}) \text{ or } a = \log(x-\sqrt{x^2-1}).$$

However, because

$$(x+\sqrt{x^2-1})(x-\sqrt{x^2-1}) = 1,$$

we have that

$$\log(x-\sqrt{x^2-1}) = -\log(x+\sqrt{x^2-1}).$$

As hyperbolic length is positive, we have that

$$a = \log(x+\sqrt{x^2-1}).$$

The hyperbolic law of cosines II has no Euclidean analogue and in fact is false in Euclidean geometry. Indeed, one reason that the interior angles of a Euclidean triangle cannot determine the side lengths is that Euclidean geometry admits dilations. As hyperbolic geometry does not admit dilations, as we have seen in Section 5.5, it is not unreasonable to have expected a result like the law of cosines II to hold in the hyperbolic plane.

Although we will not take this approach, we mention here that there is a unified proof of the three hyperbolic trigonometric laws, as might be suggested by the similarity of the forms of the hyperbolic laws of cosines I and II. We refer the interested reader to Thurston [35] for this approach, and for much more.

As we have seen on several occasions, including in the derivation of the hyperbolic law of cosines I, the calculation of the hyperbolic distance between points in $\mathbb{D}$ is fairly easy. However, as we have also seen on several occasions, such as in Exercise 5.20, calculations of angles in $\mathbb{D}$ can in general be tedious.

One application of the hyperbolic trigonometric laws is to make these calculations of angle much more tractible. For instance, we may rework Exercises 5.20 and 5.22 using the two hyperbolic laws of cosines.

For $n \geq 3$ and $0 < r < 1$, we consider the compact regular hyperbolic n-gon $P_n(r)$ in the Poincaré disc $\mathbb{D}$ with vertices at $p_k = r \exp\left(\frac{2\pi i}{n} k\right)$ for $0 \leq k \leq n-1$, as constructed in Section 5.5.

Let T be the hyperbolic triangle with vertices at 0, $p_0 = r$, and $p_1 = r \exp\left(\frac{2\pi i}{n}\right)$. The interior angle of T at 0 is $\frac{2\pi}{n}$. Also, the hyperbolic lengths of the two sides of T adjacent to 0 are equal to the hyperbolic distance from 0 to $p_0 = r$, which is

$$b = \mathrm{d}_{\mathbb{D}}(0, p_0) = \ln\left[\frac{1+r}{1-r}\right].$$

In particular,

$$\cosh(b) = \frac{1+r^2}{1-r^2} \text{ and } \sinh(b) = \frac{2r}{1-r^2}.$$

By the hyperbolic law of cosines I, the hyperbolic length a of the side of T opposite 0 satisfies

$$\cosh(a) = \cosh^2(b) - \sinh^2(b)\cos\left(\frac{2\pi}{n}\right) = \frac{(1+r^2)^2 - 4r^2\cos\left(\frac{2\pi}{n}\right)}{(1-r^2)^2}.$$

Now that we have an explicit formula for the hyperbolic length a of the side of T opposite 0, we can use the hyperbolic law of sines to determine the interior angle β of T at p_0, namely,

$$\sin(\beta) = \frac{\sinh(b)\sin\left(\frac{2\pi}{n}\right)}{\sinh(a)}.$$

The interior angle of $P_n(r)$ at p_0 is then 2β.

Exercise 5.28

Let T be a compact hyperbolic triangle whose sides have hyperbolic length a. Prove that the three interior angles of T are equal. Furthermore, if we let α be the interior angle of T at a vertex, prove that

$$2\cosh\left(\frac{1}{2}a\right)\sin\left(\frac{1}{2}\alpha\right) = 1.$$

Exercise 5.29

Let T be a compact hyperbolic triangle. Show that the three angle bisectors of T intersect in a single point. (Here, an *angle bisector* is a hyperbolic ray into the triangle from a vertex that bisects the angle at that vertex.)

Exercise 5.30

Let R be a compact hyperbolic quadrilateral with angles $\frac{1}{2}\pi$, $\frac{1}{2}\pi$, $\frac{1}{2}\pi$, and φ. Starting at the vertex with angle φ and moving counterclockwise around R, label the sides of R as A, B, C, D. Show that

$$\sinh(C)\sinh(B) = \cos(\varphi)$$

and that

$$\cosh(C) = \cosh(A)\sin(\varphi).$$

6
Nonplanar models

In this final chapter, we consider two nonplanar models of hyperbolic geometry. The first is the *hyperboloid model* of the hyperbolic plane, which is a model of the hyperbolic plane that sits naturally in $\mathbb{R}^3$, defined in terms of linear algebra, which we present in some detail. The other is a discussion of *generalizations to higher dimensions* of the models of hyperbolic geometry discussed in this book. This last section has a slightly different flavor from the rest of the book. It is intended to be a taster for topics that we do not have the space to cover in detail, and hence, it is largely expository.

6.1 The Hyperboloid Model of the Hyperbolic Plane

Up to this point, we have only considered models of the hyperbolic plane whose base space is a holomorphic disc in the complex plane $\mathbb{C}$ and whose hyperbolic element of arc-length is a conformal distortion $\lambda(z)|\mathrm{d}z|$ of the standard Euclidean metric on $\mathbb{C}$. The purpose of this section is to describe a different model of the hyperbolic plane, the *hyperboloid model*, which sits as a subset of $\mathbb{R}^3$. The planar models of the hyperbolic plane we have considered are closely tied to complex analysis, whereas the hyperboloid model is much more closely tied to linear algebra. For basic facts about linear algebra, we refer the interested reader to any undergraduate linear algebra textbook, such as Anton and Busby [6] or Strang [34].

We develop the basic properties of the hyperboloid model in a slightly different order than that taken earlier in the book. When we developed the upper half-plane model $\mathbb{H}$, we began with the definition of hyperbolic line, then determined a group of homeomorphisms of $\mathbb{H}$ that took hyperbolic lines to hyperbolic lines, and then derived the hyperbolic element of arc-length assuming its invariance under the action of this group. For the hyperboloid model $\mathbb{U}$, we begin by defining the hyperbolic length of a piecewise C^1 path, then develop a natural group of homeomorphisms of $\mathbb{U}$ preserving this hyperbolic length, and only then define hyperbolic lines.

For the remainder of this chapter, we view the elements of $\mathbb{R}^3$ as column vectors, with coordinates $\mathbf{x} = [x_0, x_1, x_2]^{\mathrm{T}}$. (Although this notation is slightly cumbersome, it is preferable to the alternatives when we express matrices and vectors in coordinates.)

So, we need to describe the base space of the hyperboloid model and define the hyperbolic length of a piecewise C^1 path. We begin by describing a (loose) way of measuring the size of a vector in $\mathbb{R}^3$.

Definition 6.1

A *quadratic form* on $\mathbb{R}^3$ is a function $q : \mathbb{R}^3 \to \mathbb{R}$ of the form $q(\mathbf{x}) = \mathbf{x}^{\mathrm{T}} B\mathbf{x}$, where B is a symmetric 3×3 matrix (with real entries).

If we are given the symmetric matrix B, then it is easy to directly calculate the quadratic form $q(\mathbf{x}) = \mathbf{x}^{\mathrm{T}} B\mathbf{x}$. Conversely, if we are given the quadratic form $q(\mathbf{x})$, we can determine the unique symmetric matrix B for which $q(\mathbf{x}) = \mathbf{x}^{\mathrm{T}} B\mathbf{x}$ by noting that

$$\mathbf{x}^{\mathrm{T}} \left(\begin{array}{ccc} \alpha & \beta & \gamma \\ \beta & \delta & \mu \\ \gamma & \mu & \eta \end{array} \right) \mathbf{x} = \alpha x_0^2 + 2\beta x_0 x_1 + \delta x_1^2 + 2\gamma x_0 x_2 + \eta x_2^2 + 2\mu x_1 x_2$$

and equating the coefficients of corresponding terms in $q(\mathbf{x})$ and $\mathbf{x}^{\mathrm{T}} B\mathbf{x}$.

We can give a crude classification of quadratic forms on $\mathbb{R}^3$ in terms of the eigenvalues of their associated symmetric matrices. (In a slight abuse of language, we sometimes refer to the eigenvalues of the symmetric matrix B as the eigenvalues of the associated quadratic form $q(\mathbf{x}) = \mathbf{x}^{\mathrm{T}} B\mathbf{x}$.)

Definition 6.2

Let $q(\mathbf{x}) = \mathbf{x}^{\mathrm{T}} B \mathbf{x}$ be a quadratic form on $\mathbb{R}^3$, where B is a symmetric 3×3 matrix.

1. If the eigenvalues of B are all positive, say that q is *positive definite*. In this case, $q(\mathbf{x}) > 0$ for all $\mathbf{x} \in \mathbb{R}^3$, $\mathbf{x} \neq \mathbf{0}$.

2. If the eigenvalues of B are all negative, say that q is *negative definite*. In this case, $q(\mathbf{x}) < 0$ for all $\mathbf{x} \in \mathbb{R}^3$, $\mathbf{x} \neq \mathbf{0}$.

3. If either B has 0 as an eigenvalue, or if B has both a positive eigenvalue and a negative eigenvalue, say that q is *indefinite*. In this case, there exists some $\mathbf{x} \in \mathbb{R}^3$, $\mathbf{x} \neq \mathbf{0}$, for which $q(\mathbf{x}) = 0$.

Exercise 6.1

Determine the eigenvalues of the quadratic form

$$q(\mathbf{x}) = -4x_0^2 + 14x_0x_1 - 2x_0x_2 + 2x_1^2 - 16x_1x_2 + 10x_2^2.$$

Exercise 6.2

For any 3×3 matrix A, consider the function $f : \mathbb{R}^3 \to \mathbb{R}$ given by $f(\mathbf{x}) = \mathbf{x}^{\mathrm{T}} A \mathbf{x}$. Write $f(\mathbf{x}) = \mathbf{x}^{\mathrm{T}} B \mathbf{x}$ for a symmetric 3×3 matrix B. Describe the relationship between the matrices A and B.

Multiplication by a 3×3 matrix A yields a linear map $A : \mathbb{R}^3 \to \mathbb{R}^3$. We can compose this linear map with a quadratic form $q : \mathbb{R}^3 \to \mathbb{R}$ to get a new quadratic form $q \circ A : \mathbb{R}^3 \to \mathbb{R}$. A natural question to ask at this point is to determine the matrices A for which q and $q \circ A$ are equal. These matrices will be important in our construction of the hyperboloid model of the hyperbolic plane.

Definition 6.3

For a quadratic form q on $\mathbb{R}^3$, let $\mathcal{O}(q)$ be the collection of all 3×3 matrices keeping q invariant: That is, $A \in \mathcal{O}(q)$ if and only if $q(\mathbf{x}) = q(A\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^3$.

We note a couple of properties of $\mathcal{O}(q)$. First, regardless of the quadratic form q, the 3×3 identity matrix I_3 is an element of $\mathcal{O}(q)$. Also, regardless of the quadratic form q, if A, $B \in \mathcal{O}(q)$, then $AB \in \mathcal{O}(q)$. This follows from the observation that, for any $\mathbf{x} \in \mathbb{R}^3$, we have that

$$q(AB\mathbf{x}) = q(A(B\mathbf{x})) = q(B\mathbf{x}) = q(\mathbf{x}),$$

where the second equality follows from $A \in \mathcal{O}(q)$ and the third equality follows from $B \in \mathcal{O}(q)$. Using a variant of this argument, we can see that, for the invertible elements of $\mathcal{O}(q)$, we have that $\mathcal{O}(q)$ contains their inverses as well.

Exercise 6.3

Let q be a quadratic form on $\mathbb{R}^3$, and let A be an invertible element of $\mathcal{O}(q)$. Show that $A^{-1} \in \mathcal{O}(q)$.

We note here, though, that there exist quadratic forms q for which $\mathcal{O}(q)$ contains a noninvertible element, and hence, it is not a subgroup of the group $\mathrm{GL}_3(\mathbb{R})$ of invertible 3×3 matrices. For example, if we take

$$q(\mathbf{x}) = 2x_0^2 + 3x_1^2 = \mathbf{x}^{\mathrm{T}} \begin{pmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 0 \end{pmatrix} \mathbf{x},$$

then

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \in \mathcal{O}(q).$$

There is a condition we can impose on a quadratic form q on $\mathbb{R}^3$ to ensure that the group $\mathcal{O}(q)$ of matrices preserving it is a subgroup of $\mathrm{GL}_3(\mathbb{R})$.

Definition 6.4

A quadratic form $q(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}B\mathbf{x}$, where B is a symmetric 3×3 matrix, is *nondegenerate* if B is invertible, and it is *degenerate* otherwise.

The quadratic form $q(\mathbf{x}) = 2x_0^2 + 3x_1^2$ is degenerate, because its associated symmetric matrix has 0 as an eigenvalue and hence is not invertible. By definition, positive definite and negative definite quadratic forms are nondegenerate, and so any degenerate quadratic form is indefinite. However, as we will see, not all indefinite quadratic forms are degenerate. One reason for considering nondegenerate quadratic forms is the following proposition.

Proposition 6.5

Let q be a nondegenerate quadratic form on $\mathbb{R}^3$. Then, every element of $\mathcal{O}(q)$ is invertible. Hence, $\mathcal{O}(q)$ is a subgroup of $\mathrm{GL}_3(\mathbb{R})$.

Proof

As we already have that $\mathcal{O}(q)$ contains the 3×3 identity matrix I_3 and is closed under composition, the only thing we need to know to show that $\mathcal{O}(q)$ is a group is that $\mathcal{O}(q)$ is closed under taking inverses.

Let q be a nondegenerate quadratic form on $\mathbb{R}^3$, and write $q(\mathbf{x}) = \mathbf{x}^{\mathrm{T}} B\mathbf{x}$, where B is a symmetric 3×3 matrix. Let A be an element of $\mathcal{O}(q)$. As $q(A\mathbf{x}) = q(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^3$, we have that

$$\mathbf{x}^{\mathrm{T}} B\mathbf{x} = q(\mathbf{x}) = q(A\mathbf{x}) = (A\mathbf{x})^{\mathrm{T}} BA\mathbf{x} = \mathbf{x}^{\mathrm{T}}(A^{\mathrm{T}} BA)\mathbf{x}$$

for all $\mathbf{x} \in \mathbb{R}^3$. As $A^{\mathrm{T}} BA$ is symmetric, the uniqueness of the symmetric matrix determining a quadratic form implies that $B = A^{\mathrm{T}} BA$. In particular, if there exists a nonzero vector $\mathbf{v}$ for which $A\mathbf{v} = \mathbf{0}$, then $A^{\mathrm{T}} BA\mathbf{v} = B\mathbf{v} = \mathbf{0}$ as well, which contradicts the assumption that B is invertible. Hence, $\ker(A) = \{\mathbf{0}\}$ and so A is invertible. By Exercise 6.3, $A^{-1} \in \mathcal{O}(q)$.

As every element of $\mathcal{O}(q)$ is invertible, we have that $\mathcal{O}(q)$ is a subgroup of $\mathrm{GL}_3(\mathbb{R})$. **QED**

Exercise 6.4

Let q be a degenerate quadratic form. Show that $\mathcal{O}(q)$ contains a noninvertible element and, hence, it cannot be a subgroup of $\mathrm{GL}_3(\mathbb{R})$.

We now make use of some multivariable calculus. Like any real-valued function on $\mathbb{R}^3$, a quadratic form q on $\mathbb{R}^3$ naturally determines a collection of subsets of $\mathbb{R}^3$. For $c \in \mathbb{R}$, consider the *level set*

$$S_c = \{\mathbf{x} \in \mathbb{R}^3 \mid q(\mathbf{x}) = c\}$$

of q in $\mathbb{R}^3$. Even though we do not yet know the composition of $\mathcal{O}(q)$, we do have that $A(S_c) \subset S_c$ for every $A \in \mathcal{O}(q)$ and every $c \in \mathbb{R}$. This result follows from the observation that if $\mathbf{x} \in S_c$, then by definition $q(\mathbf{x}) = c$; for an element $A \in \mathcal{O}(q)$, we have that $q(A\mathbf{x}) = q(\mathbf{x}) = c$, and so $A\mathbf{x} \in S_c$.

If q is nondegenerate, then we can say more. Namely, because every element of $\mathcal{O}(q)$ is invertible, we also have that $A^{-1}S_c \subset S_c$ for every $A \in \mathcal{O}(q)$ and every $c \in \mathbb{R}$. Multiplying through by A, we obtain that $S_c \subset AS_c$. Combining this observation with the observation above that $A(S_c) \subset S_c$ for every $A \in \mathcal{O}(q)$ and every $c \in \mathbb{R}$, we see that $AS_c = S_c$ for every $A \in \mathcal{O}(q)$ and every $c \in \mathbb{R}$.

The quadratic form that will be of most interest to us through this chapter is the indefinite, nondegenerate quadratic form

$$Q(\mathbf{x}) = -x_0^2 + x_1^2 + x_2^2 = \mathbf{x}^{\mathrm{T}} \begin{pmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \mathbf{x}.$$

In particular, we have that $\mathcal{O}(Q)$ is a subgroup of the group $\mathrm{GL}_3(\mathbb{R})$ of all invertible 3×3 matrices.

The analysis that follows can be carried out for any quadratic form, although the specifics may differ depending on the particular quadratic form we choose to work with.

There is a particular level set we are interested in here, namely,

$$S_{-1} = \{\mathbf{x} \in \mathbb{R}^3 \mid Q(\mathbf{x}) = -1\}.$$

The level set S_{-1} is disjoint from the x_1x_2-plane in $\mathbb{R}^3$, because $Q([0, x_1, x_2]^{\mathrm{T}}) = -1$ implies that $x_1^2 + x_2^2 = -1$, and this equation has no solutions for real numbers x_1 and x_2. Note that $[1, 0, 0]^{\mathrm{T}} \in S_{-1}$, because $Q([1, 0, 0]^{\mathrm{T}}) = -(1^2) + 0^2 + 0^2 = -1$, and that each level set S_c is invariant under reflection in the x_1x_2-plane, because $Q([x_0, x_1, x_2]^{\mathrm{T}}) = Q([-x_0, x_1, x_2]^{\mathrm{T}}) = -x_0^2 + x_1^2 + x_2^2$.

In particular, the level set S_{-1} has two components: the *upper sheet*

$$\mathbb{U} = \{\mathbf{x} = [x_0, x_1, x_2]^{\mathrm{T}} \in \mathbb{R}^3 \mid Q(\mathbf{x}) = -1 \text{ and } x_0 > 0\}$$

and the *lower sheet*

$$\mathbb{L} = \{\mathbf{x} = [x_0, x_1, x_2]^{\mathrm{T}} \in \mathbb{R}^3 \mid Q(\mathbf{x}) = -1 \text{ and } x_0 < 0\}.$$

The upper sheet $\mathbb{U}$ of S_{-1} is the base space for the hyperboloid model of the hyperbolic plane. For a side view, see Figure 6.1.

In fact, for any $c \neq 0$, the level set S_c is a hyperboloid. If $c < 0$, then this hyperboloid has two sheets, which are interchanged by reflection in the x_1x_2-plane. If $c > 0$, then this hyperboloid has one sheet. The level set S_0 is a cone.

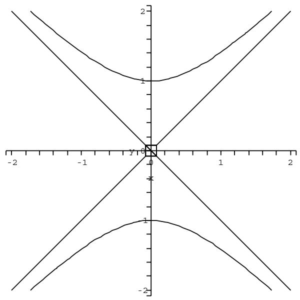

Figure 6.1: Side view of hyperboloid and cone, with x_0-axis vertical

Let $\mathcal{O}^+(Q)$ be the subgroup of $\mathcal{O}(Q)$ preserving the upper sheet $\mathbb{U}$ of S_{-1}. We can describe $\mathcal{O}^+(Q)$ as the group of invertible 3×3 matrices A for which $Q(\mathbf{x}) = Q(A\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^3$ and for which the x_0-coordinate of $A([1,0,0]^{\mathrm{T}})$ is positive. Note that $\mathcal{O}^+(Q)$ is an index 2 subgroup of $\mathcal{O}(Q)$.

One task before us is to determine which matrices lie in $\mathcal{O}^+(Q)$ and to analyze their action on $\mathbb{U}$. However, before doing this task, we show that it is possible to define the hyperbolic length of a piecewise C^1 path in $\mathbb{U}$ and that it is invariant under the action of $\mathcal{O}^+(Q)$, without an explicit description of the elements of $\mathcal{O}^+(Q)$. We begin with an examination of tangent vectors to $\mathbb{U}$.

Proposition 6.6

Let $\mathbf{u}$ be any point of $\mathbb{U}$, and let $\mathbf{t}$ be a nonzero vector tangent to $\mathbb{U}$ at $\mathbf{u}$. Then, $Q(\mathbf{t}) > 0$.

Proof

We begin by noting that the level set $\{\mathbf{x} \in \mathbb{R}^3 \mid Q(\mathbf{x}) = 0\}$ is the right-angled cone

$$C = \{[x_0, x_1, x_2]^{\mathrm{T}} \in \mathbb{R}^3 \mid x_0^2 = x_1^2 + x_2^2\}.$$

The region $\{\mathbf{x} \in \mathbb{R}^3 | Q(\mathbf{x}) < 0\}$ is the disconnected region containing the x_0-axis (except for the origin), whereas the region $\{\mathbf{x} \in \mathbb{R}^3 \,|\, Q(\mathbf{x}) > 0\}$ is the connected region containing the x_1x_2-plane (except for the origin). If we draw $\mathbb{R}^3$ so that the x_0-axis is vertical, then the (disconnected) region $\{\mathbf{x} \in \mathbb{R}^3 \mid Q(\mathbf{x}) < 0\}$ lies above and below the cone C, whereas the (connected) region $\{\mathbf{x} \in \mathbb{R}^3 | Q(\mathbf{x}) > 0\}$ sits between the top and bottom parts of the cone C. Again, see Figure 6.1 for a side view.

As $\mathbb{U}$ is contained in the level set $\{\mathbf{x} \in \mathbb{R}^3 \mid Q(\mathbf{x}) = -1\}$, every normal vector to $\mathbb{U}$ at a point $\mathbf{u} = [u_0, u_1, u_2]^{\mathrm{T}} \in \mathbb{U}$ is a multiple of the gradient

$$\nabla Q(\mathbf{u}) = \begin{bmatrix} -2u_0 \\ 2u_1 \\ 2u_2 \end{bmatrix}$$

evaluated at $\mathbf{u}$. So, a nonzero normal vector $\mathbf{n}$ to $\mathbb{U}$ at a point $\mathbf{u}$ of $\mathbb{U}$ has the form

$$\mathbf{n} = c \begin{bmatrix} -2u_0 \\ 2u_1 \\ 2u_2 \end{bmatrix}$$

for some $c \in \mathbb{R}$, $c \neq 0$ and, hence, it satisfies

$$Q(\mathbf{n}) = Q\left(c \begin{bmatrix} -2u_0 \\ 2u_1 \\ 2u_2 \end{bmatrix} \right) = 4c^2(-u_0^2 + u_1^2 + u_2^2) = -4c^2 < 0.$$

(As $\mathbf{u} \in \mathbb{U}$, we have that $-u_0^2 + u_1^2 + u_2^2 = -1$.) Therefore, because any nonzero tangent vector $\mathbf{t}$ to $\mathbb{U}$ at $\mathbf{u} \in \mathbb{U}$ is perpendicular to the normal vector $\mathbf{n}$ to $\mathbb{U}$ at $\mathbf{u}$, nonzero normal vectors lie in the region $\{\mathbf{x} \in \mathbb{R}^3 \mid Q(\mathbf{x}) < 0\}$, and the cone C is a right-angled cone, we see that nonzero tangent vectors to $\mathbb{U}$ lie in the region $\{\mathbf{x} \in \mathbb{R}^3 \mid Q(\mathbf{x}) > 0\}$, Hence, we have that $Q(\mathbf{t}) > 0$. **QED**

As a consequence, let $f : [a, b] \to \mathbb{U}$ be a piecewise C^1 path. As $Q(f'(t)) > 0$, as $f'(t)$ is tangent to $\mathbb{U}$ at $f(t)$, it is tempting to define the hyperbolic length of f to be

$$\mathrm{length}_{\mathbb{U}}(f) = \int_a^b \sqrt{Q(f'(t))}\mathrm{d}t.$$

For this to be a reasonable definition, we need to show that this putative definition of hyperbolic length is invariant under the action of $\mathcal{O}^+(Q)$.

Proposition 6.7

For any piecewise C^1 path $f : [a, b] \to \mathbb{U}$ and for any $A \in \mathcal{O}^+(Q)$, we have that $\mathrm{length}_{\mathbb{U}}(f) = \mathrm{length}_{\mathbb{U}}(A \circ f)$.

Proof

As

$$\text{length}_{\mathbb{U}}(A \circ f) = \int_a^b \sqrt{Q((A \circ f)'(t))}\mathrm{d}t,$$

we need to consider $Q((A \circ f)'(t))$. Viewing A as a linear map $A : \mathbb{R}^3 \to \mathbb{R}^3$, we observe that $(A \circ f)'(t) = A \circ f'(t)$. This equality can be checked either using the definition of the derivative of a linear map from $\mathbb{R}^3$ to $\mathbb{R}^3$ or writing $f(t)$ and A in terms of their coordinates, calculating $(A \circ f)'(t)$ and $A \circ f'(t)$ directly, and comparing the resulting expressions. We do the latter.

So, write

$$f(t) = \begin{bmatrix} x_0(t) \\ x_1(t) \\ x_2(t) \end{bmatrix}$$

and

$$A = \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & k \end{pmatrix}.$$

Then,

$$\begin{aligned}
(A \circ f)'(t) &= \left(\begin{pmatrix} a & b & c \\ d & e & f \\ g & h & k \end{pmatrix} \begin{bmatrix} x_0(t) \\ x_1(t) \\ x_2(t) \end{bmatrix} \right)' \\
&= \begin{bmatrix} ax_0(t) + bx_1(t) + cx_2(t) \\ dx_0(t) + ex_1(t) + fx_2(t) \\ gx_0(t) + hx_1(t) + kx_2(t) \end{bmatrix}' \\
&= \begin{bmatrix} ax_0'(t) + bx_1'(t) + cx_2'(t) \\ dx_0'(t) + ex_1'(t) + fx_2'(t) \\ gx_0'(t) + hx_1'(t) + kx_2'(t) \end{bmatrix} \\
&= \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & k \end{pmatrix} \begin{bmatrix} x_0'(t) \\ x_1'(t) \\ x_2'(t) \end{bmatrix} = A \circ f'(t),
\end{aligned}$$

as desired.

Hence, using that $A \in \mathcal{O}^+(Q)$, we see that

$$Q((A \circ f)'(t)) = Q(A \circ f'(t)) = Q(f'(t)),$$

and so

$$\text{length}_{\mathbb{U}}(A \circ f) = \int_a^b \sqrt{Q((A \circ f)'(t))}\mathrm{d}t = \int_a^b \sqrt{Q(f'(t))}\mathrm{d}t = \text{length}_{\mathbb{U}}(f),$$

as desired. **QED**

As an example, consider the set T_r, defined to be the intersection of $\mathbb{U}$ with the Euclidean plane $P_r = \{\mathbf{x} \in \mathbb{R}^3 \mid x_0 = r\}$ for $r > 1$. If $[r, s, t]^{\mathrm{T}} \in T_r$, then $s^2 + t^2 = r^2 - 1$, and so we can express s and t as $s = \sqrt{r^2-1}\cos(\alpha)$ and $t = \sqrt{r^2-1}\sin(\alpha)$ for some $\alpha \in \mathbb{R}$. As α varies over $[0, 2\pi]$, we cover all of T_r. In particular, we see that T_r is a Euclidean circle in the plane P_r in $\mathbb{R}^3$. Moreover, the (Euclidean) radius of T_r is $\sqrt{r^2-1}$, and so the (Euclidean) length of T_r is $2\pi\sqrt{r^2-1}$.

To calculate the hyperbolic length of T_r, we begin by parametrizing it by the path $f : [0, 2\pi] \to \mathbb{U}$ given by

$$f(t) = \begin{bmatrix} r \\ \sqrt{r^2-1}\cos(t) \\ \sqrt{r^2-1}\sin(t) \end{bmatrix}.$$

For this choice of f, we have that

$$f'(t) = \begin{bmatrix} 0 \\ -\sqrt{r^2-1}\sin(t) \\ \sqrt{r^2-1}\cos(t) \end{bmatrix},$$

and so $Q(f'(t)) = r^2 - 1$. Hence,

$$\text{length}_{\mathbb{U}}(f) = \int_0^{2\pi} \sqrt{r^2-1}\mathrm{d}t = 2\pi\sqrt{r^2-1},$$

which is equal to the Euclidean length of T_r, when T_r is viewed as a circle in P_r.

As we can calculate the hyperbolic length of a piecewise C^1 path in $\mathbb{U}$, we can define the hyperbolic metric $\mathrm{d}_{\mathbb{U}}$ on $\mathbb{U}$ as we have done earlier with the upper half-plane $\mathbb{H}$, by defining the hyperbolic distance $\mathrm{d}_{\mathbb{U}}(\mathbf{u}, \mathbf{v})$ between points $\mathbf{u}$ and $\mathbf{v}$ of $\mathbb{U}$ to be the infimum of the hyperbolic lengths of all piecewise C^1 paths $f : [a, b] \to \mathbb{U}$ with $f(a) = \mathbf{u}$ and $f(b) = \mathbf{v}$. As before, the invariance of hyperbolic length $\text{length}_{\mathbb{U}}$ under the action of $\mathcal{O}^+(Q)$ immediately implies the invariance of hyperbolic distance $\mathrm{d}_{\mathbb{U}}$ under the action of $\mathcal{O}^+(Q)$, using the same argument as given in the proof of Proposition 3.17.

We are now ready to determine the elements of $\mathcal{O}^+(Q)$.

Proposition 6.8

The group $\mathcal{O}^+(Q)$ is generated by the family of matrices

$$A_\alpha = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos(\alpha) & -\sin(\alpha) \\ 0 & \sin(\alpha) & \cos(\alpha) \end{pmatrix}$$

for $\alpha \in \mathbb{R}$, the family of matrices

$$B_\mu = \begin{pmatrix} \cosh(\mu) & 0 & \sinh(\mu) \\ 0 & 1 & 0 \\ \sinh(\mu) & 0 & \cosh(\mu) \end{pmatrix}$$

for $\mu \in \mathbb{R}$, and the single matrix

$$C_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix}.$$

Proof

We could proceed from the fact, as discussed in the proof of Proposition 6.5, that $\mathcal{O}(Q)$ consists of all invertible 3×3 matrices A for which $B = A^{\mathrm{T}}BA$, but this calculation quickly becomes unwieldy. Instead, we proceed in several steps.

We begin by determining the subgroup of $\mathcal{O}^+(Q)$ fixing the point $[1,0,0]^{\mathrm{T}}$. So, let

$$A = \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & k \end{pmatrix}$$

be any element of $\mathcal{O}^+(Q)$ satisfying $A\,[1,0,0]^{\mathrm{T}} = [1,0,0]^{\mathrm{T}}$. As

$$\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = A \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & k \end{pmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} a \\ d \\ g \end{bmatrix},$$

we have immediately that $a = 1$ and $d = g = 0$.

We now use that $Q(\mathbf{x}) = Q(A\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^3$. Calculating, we see that

$$A\mathbf{x} = \begin{pmatrix} 1 & b & c \\ 0 & e & f \\ 0 & h & k \end{pmatrix} \begin{bmatrix} x_0 \\ x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x_0 + bx_1 + cx_2 \\ ex_1 + fx_2 \\ hx_1 + kx_2 \end{bmatrix},$$

and so

$$\begin{aligned} Q(A\mathbf{x}) \quad = \quad & -x_0^2 + (-b^2 + e^2 + h^2)x_1^2 + (-c^2 + f^2 + k^2)x_2^2 \\ & -2bx_0x_1 - 2cx_0x_2 + 2(-bc + ef + hk)x_1x_2. \end{aligned}$$

Equating the coefficients of corresponding terms of $Q(\mathbf{x}) = -x_0^2 + x_1^2 + x_2^2$ and $Q(A\mathbf{x})$, we see that

$$b = c = 0,$$

that

$$e^2 + h^2 = 1,$$

that

$$f^2 + k^2 = 1,$$

and that

$$ef + hk = 0.$$

In particular, there are real numbers α and β so that

$$e = \cos(\alpha) \text{ and } h = \sin(\alpha)$$

and

$$f = \cos(\beta) \text{ and } k = \sin(\beta).$$

Thus, A has the form

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & e & f \\ 0 & h & k \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos(\alpha) & \cos(\beta) \\ 0 & \sin(\alpha) & \sin(\beta) \end{pmatrix}.$$

The equation $ef + hk = 0$ then becomes

$$\cos(\alpha)\cos(\beta) + \sin(\alpha)\sin(\beta) = 0,$$

and so $\cos(\alpha - \beta) = 0$. Hence, $\alpha - \beta = \frac{\pi}{2} + k\pi$ for $k \in \mathbb{Z}$. There are two cases to consider.

If $\alpha - \beta = \frac{\pi}{2} + 2k\pi$ for $k \in \mathbb{Z}$, then $\beta = \alpha - \frac{\pi}{2} - 2k\pi$. In this case, we see that $\cos(\beta) = \sin(\alpha)$ and $\sin(\beta) = -\cos(\alpha)$, and so

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & e & f \\ 0 & h & k \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos(\alpha) & \sin(\alpha) \\ 0 & \sin(\alpha) & -\cos(\alpha) \end{pmatrix}.$$

If $\alpha - \beta = \frac{3\pi}{2} + 2k\pi$ for $k \in \mathbb{Z}$, then $\beta = \alpha - \frac{3\pi}{2} - 2k\pi$. In this case, we see that $\cos(\beta) = -\sin(\alpha)$ and $\sin(\beta) = \cos(\alpha)$, and so

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & e & f \\ 0 & h & k \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos(\alpha) & -\sin(\alpha) \\ 0 & \sin(\alpha) & \cos(\alpha) \end{pmatrix}.$$

Note that

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos(\alpha) & \sin(\alpha) \\ 0 & \sin(\alpha) & -\cos(\alpha) \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos(\alpha) & -\sin(\alpha) \\ 0 & \sin(\alpha) & \cos(\alpha) \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix},$$

and so the subgroup of $\mathcal{O}^+(Q)$ fixing $[1,0,0]^{\mathrm{T}}$ is generated by the family of matrices

$$A_\alpha = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos(\alpha) & -\sin(\alpha) \\ 0 & \sin(\alpha) & \cos(\alpha) \end{pmatrix}$$

for $\alpha \in \mathbb{R}$, and the single matrix

$$C_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix}.$$

Set

$$\begin{aligned} C_1 = A_{\frac{\pi}{2}} C_2 A_{\frac{\pi}{2}}^{-1} &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \end{aligned}$$

Geometrically, each A_α is a rotation of $\mathbb{U}$ about the x_0-axis. The transformations C_1 and C_2 are reflections, where C_1 is reflection in the plane $\{\mathbf{x} \in \mathbb{R}^3 \mid x_1 = 0\}$, which is the x_0x_2-plane, and C_2 is reflection in the plane $\{\mathbf{x} \in \mathbb{R}^3 \mid x_2 = 0\}$, which is the x_0x_1-plane. In general, if S_α is the Euclidean plane containing the x_0-axis that makes angle α with the x_0x_1-plane, then $A_\alpha C_2 A_\alpha^{-1}$ is reflection in S_α.

For the next step of this argument, let

$$A = \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & k \end{pmatrix}$$

be any element of $\mathcal{O}^+(Q)$ preserving the x_0x_2-plane. Any element of the x_0x_2-plane in $\mathbb{R}^3$ has the form

$$\begin{bmatrix} * \\ 0 \\ * \end{bmatrix},$$

where $*$ represents any real number. As

$$A \begin{bmatrix} * \\ 0 \\ * \end{bmatrix} = \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & k \end{pmatrix} \begin{bmatrix} * \\ 0 \\ * \end{bmatrix} = \begin{bmatrix} * \\ 0 \\ * \end{bmatrix},$$

we see that A has the form

$$A = \begin{pmatrix} a & b & c \\ 0 & e & 0 \\ g & h & k \end{pmatrix}.$$

As before, we now use that $Q(\mathbf{x}) = Q(A\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^3$. Calculating, we see that

$$A\mathbf{x} = \begin{pmatrix} a & b & c \\ 0 & e & 0 \\ g & h & k \end{pmatrix} \begin{bmatrix} x_0 \\ x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} ax_0 + bx_1 + cx_2 \\ ex_1 \\ gx_0 + hx_1 + kx_2 \end{bmatrix},$$

and so

$$\begin{aligned} Q(A\mathbf{x}) &= (-a^2 + g^2)x_0^2 + (-b^2 + e^2 + h^2)x_1^2 + (-c^2 + k^2)x_2^2 \\ & \quad 2(-ab + gh)x_0x_1 + 2(-ac + gk)x_0x_2 + 2(-bc + hk)x_1x_2. \end{aligned}$$

Again equating the coefficients of corresponding terms of $Q(\mathbf{x}) = -x_0^2 + x_1^2 + x_2^2$ and $Q(A\mathbf{x})$, we see that

$$a^2 - g^2 = 1,$$

that

$$k^2 - c^2 = 1,$$

that

$$-b^2 + e^2 + h^2 = 1,$$

and that

$$ab - gh = ac - gk = bc - hk = 0.$$

In particular, there are real numbers μ and η so that

$$a = \pm \cosh(\mu) \text{ and } g = \pm \sinh(\mu)$$

and

$$k = \pm \cosh(\eta) \text{ and } c = \pm \sinh(\eta).$$

Note that

$$A \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{pmatrix} a & b & c \\ 0 & e & 0 \\ g & h & k \end{pmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} a \\ 0 \\ g \end{bmatrix} \in \mathbb{U},$$

and so $a > 0$. Hence, $a = \cosh(\mu)$. Note that $\cosh(\mu) = \cosh(\mu^{-1})$ and $-\sinh(\mu) = \sinh(\mu^{-1})$; these identities can easily be derived by expressing $\cosh(x)$ and $\sinh(x)$ in terms of $\exp(x)$. Hence, by replacing μ by μ^{-1}, if necessary, we can assume that $g = \sinh(\mu)$.

So, we now have that

$$A = \begin{pmatrix} \cosh(\mu) & b & c \\ 0 & e & 0 \\ \sinh(\mu) & h & k \end{pmatrix},$$

and so

$$AC_2 = \begin{pmatrix} \cosh(\mu) & b & c \\ 0 & e & 0 \\ \sinh(\mu) & h & k \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix} = \begin{pmatrix} \cosh(\mu) & b & -c \\ 0 & e & 0 \\ \sinh(\mu) & h & -k \end{pmatrix}.$$

Replace A by AC_2 if $k < 0$ to ensure that $k = \cosh(\eta) > 0$. Then, replace η by η^{-1} if $c < 0$ to ensure that $c = \sinh(\eta)$. So, after possibly replacing μ by μ^{-1}, possibly replacing A by AC_2, and possibly replacing η by η^{-1}, we see that A has the form

$$A = \begin{pmatrix} \cosh(\mu) & b & \sinh(\eta) \\ 0 & e & 0 \\ \sinh(\mu) & h & \cosh(\eta) \end{pmatrix}.$$

The equation $ac - gk = 0$ then becomes the equation

$$\cosh(\mu)\sinh(\eta) - \cosh(\eta)\sinh(\mu) = 0.$$

Rewrite this equation as $\tanh(\mu) = \tanh(\eta)$. As $\tanh(t)$ is an increasing function of t, we see that $\mu = \eta$.

Thus, A has the form

$$A = \begin{pmatrix} a & b & c \\ 0 & e & 0 \\ g & h & k \end{pmatrix} = \begin{pmatrix} \cosh(\mu) & b & \sinh(\mu) \\ 0 & e & 0 \\ \sinh(\mu) & h & \cosh(\mu) \end{pmatrix}.$$

The equation $ab = gh$ becomes the equation $\cosh(\mu)b = \sinh(\mu)h$, and the equation $bc = hk$ becomes the equation $\sinh(\mu)b = \cosh(\mu)h$. Hence,

$$b = \frac{\sinh(\mu)}{\cosh(\mu)}h = \left(\frac{\sinh(\mu)}{\cosh(\mu)}\right)^2 b.$$

There are two possibilities. The first possibility is that $b = 0$, which then forces $h = 0$, so that A has the form

$$A = \begin{pmatrix} \cosh(\mu) & 0 & \sinh(\mu) \\ 0 & e & 0 \\ \sinh(\mu) & 0 & \cosh(\mu) \end{pmatrix},$$

where $e^2 = 1$. Therefore, $e = \pm 1$. If $e = 1$, then $A = B_\mu$. If $e = -1$, then $A = C_1 B_\mu$.

The other possibility is that $\sinh(\mu) = 0$, which then forces $\mu = 0$, so that A has the form

$$A = \begin{pmatrix} 1 & b & 0 \\ 0 & e & 0 \\ 0 & h & 1 \end{pmatrix}.$$

However, in this case, we calculate to see that $A([1, 0, 0]^{\mathrm{T}}) = [1, 0, 0]^{\mathrm{T}}$. As A lies in the subgroup of $\mathcal{O}^+(Q)$ fixing $[1, 0, 0]^{\mathrm{T}}$, we have that $b = 0$, that $h = 0$, and that $e = \pm 1$, using our earlier work. If $e = 1$, then $A = I_3$, the 3×3 identity matrix. If $e = -1$, then $A = C_1$.

To summarize, the subgroup of $\mathcal{O}^+(Q)$ preserving the x_0x_2-plane is generated by the family of matrices

$$B_\mu = \begin{pmatrix} \cosh(\mu) & 0 & \sinh(\mu) \\ 0 & 1 & 0 \\ \sinh(\mu) & 0 & \cosh(\mu) \end{pmatrix},$$

and by the single matrices

$$C_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix}$$

and

$$C_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Note that C_1 is reflection in the x_0x_2-plane and so fixes the x_0x_2-plane pointwise. C_2 is reflection in the x_0x_1-plane, which is perpendicular to the x_0x_2-plane, and so C_2 takes the x_0x_2-plane to itself and acts as reflection across the x_0-axis in the x_0x_2-plane. We postpone the geometric interpretation of B_μ until we have defined the notion of hyperbolic line in $\mathbb{U}$.

To complete the proof, let A be any element of $\mathcal{O}^+(Q)$. We show that A can be expressed as a composition of the A_α for $\alpha \in \mathbb{R}$, the B_μ for $\mu \in \mathbb{R}$, and C_2, by following the image of a single point. So, consider the point $A([1,0,0]^{\mathrm{T}}) = [r,s,t]^{\mathrm{T}}$. We will construct an element of $\mathcal{O}^+(Q)$ taking $[r,s,t]^{\mathrm{T}}$ to $[1,0,0]^{\mathrm{T}}$.

As $[r,s,t]^{\mathrm{T}} \in \mathbb{U}$, we have that $-r^2+s^2+t^2 = -1$ and that $r \geq 1$. If $r = 1$, then $s = t = 0$ and so $[r,s,t]^{\mathrm{T}} = [1,0,0]^{\mathrm{T}}$. In this case, our analysis of the elements of $\mathcal{O}^+(Q)$ fixing $[1,0,0]^{\mathrm{T}}$ yields that $A = A_\alpha C_2^\varepsilon$ for some $\alpha \in \mathbb{R}$ and for $\varepsilon = 0$ or 1.

Hence, we can assume that $r > 1$. We first wish to find α so that $A_\alpha([r,s,t]^{\mathrm{T}})$ lies in the x_0x_2-plane. (Geometrically, it is obvious that we can rotate $\mathbb{U}$ about the x_0-axis so that the image of $[r,s,t]^{\mathrm{T}}$ under this rotation lies in the x_0x_2-plane. However, we take this opportunity to see how to find the specific value of α required to achieve this result.) Calculating, we see that

$$A_\alpha \begin{bmatrix} r \\ s \\ t \end{bmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos(\alpha) & -\sin(\alpha) \\ 0 & \sin(\alpha) & \cos(\alpha) \end{pmatrix} \begin{bmatrix} r \\ s \\ t \end{bmatrix} = \begin{bmatrix} r \\ s\cos(\alpha) - t\sin(\alpha) \\ s\sin(\alpha) + t\cos(\alpha) \end{bmatrix}.$$

As $s^2+t^2 = r^2-1 > 0$, we can write $s = \sqrt{r^2-1}\cos(\beta)$ and $t = \sqrt{r^2-1}\sin(\beta)$ for some $\beta \in \mathbb{R}$, so that

$$\begin{aligned} s\cos(\alpha) - t\sin(\alpha) &= \sqrt{r^2-1}(\cos(\beta)\cos(\alpha) - \sin(\beta)\sin(\alpha)) \\ &= \sqrt{r^2-1}\cos(\beta+\alpha). \end{aligned}$$

If we set $\alpha = \frac{\pi}{2} - \beta$, then $s\cos(\alpha) - t\sin(\alpha) = 0$.

So, we have found a value of α so that

$$A_\alpha A \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = A_\alpha \begin{bmatrix} r \\ s \\ t \end{bmatrix} = \begin{bmatrix} r \\ 0 \\ T \end{bmatrix}$$

for some T. Note that because $r > 1$ by assumption, we have that $T \neq 0$. To complete the argument, we find a value of μ so that

$$B_\mu A_\alpha A \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = B_\mu \begin{bmatrix} r \\ 0 \\ T \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}.$$

Calculating, we see that

$$B_\mu \begin{bmatrix} r \\ 0 \\ T \end{bmatrix} = \begin{pmatrix} \cosh(\mu) & 0 & \sinh(\mu) \\ 0 & 1 & 0 \\ \sinh(\mu) & 0 & \cosh(\mu) \end{pmatrix} \begin{bmatrix} r \\ 0 \\ T \end{bmatrix} = \begin{bmatrix} \cosh(\mu)r + \sinh(\mu)T \\ 0 \\ \sinh(\mu)r + \cosh(\mu)T \end{bmatrix}.$$

Hence, we need to find μ so that

$$\cosh(\mu)r + \sinh(\mu)T = 1$$

and

$$\sinh(\mu)r + \cosh(\mu)T = 0.$$

The second of this pair of equations yields that $T = -\tanh(\mu)r$, and so

$$\cosh(\mu)r + \sinh(\mu)T = (\cosh(\mu) - \sinh(\mu)\tanh(\mu))r = \frac{r}{\cosh(\mu)}.$$

Hence, we see that $\cosh(\mu)r + \sinh(\mu)T = 1$ if and only if $r = \cosh(\mu)$. So, given r, we need only find μ so that $\cosh(\mu) = r$, which is always possible as $r > 1$.

We now complete the argument by using the first part of this proof. As $B_\mu A_\alpha A$ fixes $[1, 0, 0]^{\mathrm{T}}$ for the values of α and μ we have just found, there exists $\theta \in \mathbb{R}$ and $\varepsilon = 0$ or 1 so that $B_\mu A_\alpha A = A_\theta C_2^\varepsilon$. Hence, we have that $A = A_\alpha^{-1} B_\mu^{-1} A_\theta C_2^\varepsilon$. **QED**

One consequence of the proof of Proposition 6.8 is that if $A \in \mathcal{O}^+(Q)$, then $\det(A) = \pm 1$, as each of the generating matrices A_α, B_μ, and C_2 for $\mathcal{O}^+(Q)$ has determinant ± 1. The notation we have chosen here is slightly nonstandard. A common notation for the group of matrices preserving the quadratic form $Q(\mathbf{x}) = -x_0^2 + x_1^2 + x_2^2$ is $\mathrm{O}(2,1)$. Let $\mathrm{SO}(2,1)$ be the subgroup of $\mathrm{O}(2,1)$ consisting of those elements of $\mathrm{O}(2,1)$ of determinant 1, and note that $\mathrm{SO}(2,1)$ is generated by the A_α for $\alpha \in \mathbb{R}$, and the B_μ for $\mu \in \mathbb{R}$.

Define a *hyperbolic line* in $\mathbb{U}$ to be the intersection of $\mathbb{U}$ with a Euclidean plane in $\mathbb{R}^3$ through the origin $\mathbf{0}$. With this definition, it becomes immediately clear that any two distinct points in $\mathbb{U}$ determine a unique hyperbolic line in $\mathbb{U}$, because two distinct points in $\mathbb{U}$ and the origin $\mathbf{0}$ (which is not in $\mathbb{U}$) determine a unique Euclidean plane in $\mathbb{R}^3$. Also, because every element of $\mathcal{O}^+(Q)$, when viewed as a linear map from $\mathbb{R}^3$ to $\mathbb{R}^3$, takes Euclidean planes through $\mathbf{0}$ to Euclidean planes through $\mathbf{0}$, we see immediately that every element of $\mathcal{O}^+(Q)$ takes hyperbolic lines in $\mathbb{U}$ to hyperbolic lines in $\mathbb{U}$. The proof that distance-realizing paths in $\mathbb{U}$ are the almost simple parametrizations of hyperbolic line segments in $\mathbb{U}$ is similar in spirit to the proof given for the upper half-plane $\mathbb{H}$. We leave the pursuit of the details for the interested reader.

We can now give a geometric interpretation of the element B_μ of $\mathcal{O}^+(Q)$. The intersection of the x_0x_2-plane with $\mathbb{U}$ is a hyperbolic line ℓ, and B_μ acts as translation along ℓ. The reflection C_2 in the x_0x_1-plane takes ℓ to itself, but it conjugates B_μ to B_μ^{-1}, as the reflection C_2 reverses the direction of travel along ℓ. In fact, C_2 fixes $[1,0,0]^{\mathrm{T}}$ and interchanges the two hyperbolic rays in ℓ from $[1,0,0]^{\mathrm{T}}$. The reflection C_1 in the x_0x_2-plane fixes ℓ pointwise and so conjugates B_μ to B_μ.

The proof of Proposition 6.8 gives a means of addressing the transitivity properties of the action of $\mathcal{O}^+(Q)$ on $\mathbb{U}$. All transitivity properties we derived for the upper half-plane model hold true here. We give proofs of two of the most important and leave the remaining properties for the interested reader.

Proposition 6.9

First, $\mathcal{O}^+(Q)$ acts transitively on $\mathbb{U}$. Second, $\mathcal{O}^+(Q)$ acts transitively on pairs $(\mathbf{u}, \ell)$, where $\mathbf{u}$ is a point of $\mathbb{U}$ and ℓ is a hyperbolic line passing through $\mathbf{u}$.

Proof

We first consider the action of $\mathcal{O}^+(Q)$ on $\mathbb{U}$. In the third part of the proof of Proposition 6.8, we saw that if $\mathbf{u}$ is any point of $\mathbb{U}$, then there exists an element E of $\mathcal{O}^+(Q)$ so that $E\mathbf{u} = [1,0,0]^{\mathrm{T}}$. Applying Lemma 2.8 completes the proof that $\mathcal{O}^+(Q)$ acts transitively on the points of $\mathbb{U}$.

For the second claim, let ℓ_1 and ℓ_2 be any two hyperbolic lines and let $\mathbf{u}_k$ be any point of ℓ_k for $k = 1$, 2. Again, using the third part of the proof of Proposition 6.8, there exist elements E_1 and E_2 of $\mathcal{O}^+(Q)$ so that $E_k\mathbf{u}_k = [1,0,0]^{\mathrm{T}}$. The first part of the proof of Proposition 6.8 shows that, because $E_1\ell_1$ and $E_2\ell_2$ are two hyperbolic lines through $[1,0,0]^{\mathrm{T}}$, there exists $\alpha \in \mathbb{R}$ so that the element A_α of $\mathcal{O}^+(Q)$ fixes $[1,0,0]^{\mathrm{T}}$ and satisfies $A_\alpha(E_1\ell_1) = E_2\ell_2$. Hence, we see that $E_2^{-1}A_\alpha E_1\mathbf{u}_1 = \mathbf{u}_2$ and $E_2^{-1}A_\alpha E_1\ell_1 = \ell_2$. **QED**

Formally, we should check that $\mathbb{U}$ with this definition of hyperbolic line satisfies the parallelism condition that we expect the hyperbolic plane to satisfy, as we have not yet shown that $\mathbb{U}$ with this definition of hyperbolic line is a model of the hyperbolic plane. In particular, we have not related it to any other model of the hyperbolic plane that we have discussed in this book. So, let ℓ be any hyperbolic line in $\mathbb{U}$, and let $\mathbf{u}$ be any point of $\mathbb{U}$ that is not contained in ℓ.

Using the transitivity properties of the action of $\mathcal{O}^+(Q)$ acting on $\mathbb{U}$, we can assume that ℓ is the intersection of $\mathbb{U}$ with the x_0x_1-plane P_{01}. Let L be the Euclidean line in $\mathbb{R}^3$ through $\mathbf{0}$ and $\mathbf{u}$. Any Euclidean plane through $\mathbf{0}$ and $\mathbf{u}$ must contain L. One such Euclidean plane is the Euclidean plane P containing both L and the x_1-axis.

The intersection of this Euclidean plane P and the x_0x_1-plane P_{01} is the x_1-axis, which is disjoint from $\mathbb{U}$, and so $P \cap \mathbb{U}$ is a hyperbolic line in $\mathbb{U}$ parallel to $\ell = P_{01} \cap \mathbb{U}$. In fact, if P' is any Euclidean plane in $\mathbb{R}^3$ whose intersection with the x_0x_1-plane P_{01} is disjoint from $\mathbb{U}$, then $P' \cap \mathbb{U}$ is a hyperbolic line in $\mathbb{U}$ parallel to ℓ. (This observation follows from the observation that any point in the intersection $(P' \cap \mathbb{U}) \cap (P_{01} \cap \mathbb{U})$ of these two hyperbolic lines is a point of $P' \cap P_{01}$, which we have assumed to be empty.)

Let P'' be any Euclidean plane containing L whose angle with P is very small. The intersection of P'' with the x_0x_1-plane P_{01} will be very close to the x_1-axis, and so it will also be disjoint from $\mathbb{U}$. Hence, the intersection of such a Euclidean plane P'' with $\mathbb{U}$ gives rise to a hyperbolic line through $\mathbf{u}$ and parallel to ℓ. As there are infinitely many such Euclidean planes, there are infinitely many hyperbolic lines through $\mathbf{u}$ and parallel to ℓ.

Another consequence of the proof of Proposition 6.8 is the observation that the group $\mathrm{Isom}(\mathbb{U}, \mathrm{d}_{\mathbb{U}})$ of all isometries of the metric space $(\mathbb{U}, \mathrm{d}_{\mathbb{U}})$ is

$$\mathrm{Isom}(\mathbb{U}, \mathrm{d}_{\mathbb{U}}) = \mathcal{O}^+(Q).$$

The proof of this fact has the same structure as the proof of Theorem 3.19, which states that $\mathrm{Isom}(\mathbb{H}, \mathrm{d}_{\mathbb{H}}) = \mathrm{M\ddot{o}b}(\mathbb{H})$. Namely, given an isometry f of $(\mathbb{U}, \mathrm{d}_{\mathbb{U}})$, we use the transitivity properties of the action of $\mathcal{O}^+(Q)$ on $\mathbb{U}$ to normalize f by finding an element A of $\mathcal{O}^+(Q)$ so that $A \circ f$ fixes a point of $\mathbb{U}$, so that $A \circ f$ fixes pointwise a hyperbolic line in $\mathbb{U}$ through this point, and so that $A \circ f$ does not interchange the two half-planes in $\mathbb{U}$ determined by this hyperbolic line. We then show that the normalized isometry$A \circ f$ is the identity, and so $f = A^{-1} \in \mathcal{O}^+(Q)$. We have already given this argument in great detail for $(\mathbb{H}, \mathrm{d}_{\mathbb{H}})$, and so we do not give it for $(\mathbb{U}, \mathrm{d}_{\mathbb{U}})$ as the details are essentially the same.

We close this section by returning to the Euclidean circle T_r defined earlier. We start by arguing that T_r is a hyperbolic circle as well. This follows from the fact that T_r is invariant under the action of the subgroup $\langle A_\alpha \mid \alpha \in \mathbb{R} \rangle$ of $\mathcal{O}^+(Q)$ fixing $[1,0,0]^{\mathrm{T}}$, and each A_α is a hyperbolic isometry as well as a Euclidean isometry. So, the hyperbolic centre of T_r is the point $[1,0,0]^{\mathrm{T}}$.

We now calculate the hyperbolic radius of T_r. The hyperbolic distance between $[1,0,0]^{\mathrm{T}}$ and $[r, \sqrt{r^2-1}, 0]^{\mathrm{T}}$ is equal to the length of the hyperbolic line segment joining $[1,0,0]^{\mathrm{T}}$ to $[r, \sqrt{r^2-1}, 0]^{\mathrm{T}}$, which in turn can be parametrized by the simple path $g : [1,r] \to \mathbb{U}$, $g(t) = [t, \sqrt{t^2-1}, 0]^{\mathrm{T}}$. The hyperbolic length of g is then

$$\begin{aligned} \mathrm{length}_{\mathbb{U}}(g) &= \int_1^r \sqrt{Q\left(\left[1, \frac{t}{\sqrt{t^2-1}}, 0\right]^{\mathrm{T}}\right)} \mathrm{d}t \\ &= \int_1^r \frac{1}{\sqrt{t^2-1}} \mathrm{d}t = \ln(r + \sqrt{r^2-1}). \end{aligned}$$

As a check that we have defined hyperbolic length appropriately, note that we have calculated that the hyperbolic length of T_r is $2\pi\sqrt{r^2-1}$ and that the hyperbolic radius of T_r is $\ln(r+\sqrt{r^2-1})$. Checking the relationship expressed in Exercise 4.4, we calculate:

$$\begin{aligned} 2\pi \sinh(\text{hyperbolic radius of } T_r) &= 2\pi \sinh(\ln(r + \sqrt{r^2-1})) \\ &= 2\pi \frac{1}{2}\left(r + \sqrt{r^2-1} - \left(\frac{1}{r+\sqrt{r^2-1}}\right)\right) \\ &= \pi(r + \sqrt{r^2-1} - (r - \sqrt{r^2-1})) \\ &= 2\pi\sqrt{r^2-1} = \mathrm{length}_{\mathbb{U}}(T_r), \end{aligned}$$

as desired.

For a slightly different presentation of the hyperboloid model, including a unified proof of the hyperbolic laws of cosines, we refer the interested reader to the book of Thurston [35].

6.2 Higher Dimensional Hyperbolic Spaces

Up to this point, we have considered models of the hyperbolic plane. The purpose of this last short section is to discuss how to extend each model to a model of higher dimensional hyperbolic spaces. We will not prove all of the specifics in detail, but instead we will attempt to point out the similarities with the construction of the two-dimensional equivalent of each model.

We begin by discussing what we mean by the phrase, *higher dimensional hyperbolic space*. There are several different directions in which we can attempt to generalize the two-dimensional hyperbolic geometry we have been studying in this book. One way is to generalize the notion of *hyperbolic*, whereas another way is to consider the same models as we have been considering, only with a higher dimensional base space.

One possible generalization of the notion of hyperbolic, which was first proposed by Gromov and which has been explored by a great many mathematicians since, begins with the result of Exercise 5.11. This exercise states that hyperbolic triangles in the hyperbolic plane are uniformly *thin*: If T is a hyperbolic triangle in the upper half-plane $\mathbb{H}$, say, with sides A, B, and C, and if $x \in A$ is any point, then

$$\mathrm{d}_{\mathbb{H}}(x, B \cup C) \leq \ln(\sqrt{2} + 1).$$

Let (X, d_X) be a path metric space. (For ease of exposition, we only consider the case in which X is simply connected.) We can define a *triangle* T in X as follows: take three points x_1, x_2, x_3 in X, and for each $1 \leq j \neq k \leq 3$, let s_{jk} be a curve in X that is the image of a distance-realizing path joining x_j and x_k. The s_{jk} are the *sides* of the triangle. For a real number $\delta > 0$, say that T is *δ-thin* if each point on one side of T is distance at most δ from the union of the other two sides. Specifically, we require that

$$\mathrm{d}_X(x, s_{12} \cup s_{13}) \leq \delta \text{ for all } x \in s_{23},$$

$$\mathrm{d}_X(x, s_{12} \cup s_{23}) \leq \delta \text{ for all } x \in s_{13},$$

and

$$\mathrm{d}_X(x, s_{23} \cup s_{13}) \leq \delta \text{ for all } x \in s_{12}.$$

We then define a (simply connected) path metric space (X, d_X) to be *δ-hyperbolic* if there exists a $\delta > 0$ so that all triangles in X are δ-thin. Exercise 5.11 shows that the hyperbolic plane is $\ln(\sqrt{2}+1)$-thin, as all triangles in the hyperbolic plane are δ-thin with $\delta = \ln(\sqrt{2}+1)$.

The interested reader is invited to check that the complex plane $\mathbb{C}$ with the usual Euclidean metric $\mathrm{n}(z, w) = |z - w|$ is not δ-hyperbolic for any $\delta > 0$, and that a tree is 0-hyperbolic. We do not have the space here to discuss this generalization in any more detail, but we refer the (brave and) interested reader to the original article of Gromov [17] as well as to the article of Bowditch [12].

Instead, we choose to work with models of hyperbolic geometry that are strict generalizations of the two-dimensional models we have developed. These models share some basic properties. First, their base space is an n-dimensional subset of Eulidean space in which we have a reasonable definition of hyperbolic line. Second, parallelism behaves as in the hyperbolic plane, so that for any hyperbolic line ℓ and any point p not on ℓ, there are at least two hyperbolic lines passing through p that are parallel to ℓ (and in fact there are infinitely many hyperbolic lines passing through p that are parallel to ℓ). And finally, the model of hyperbolic geometry looks the same in all directions, so that if p is any point and if r_1 and r_2 are any two hyperbolic rays from p, then there exists an isometry of this hyperbolic space fixing p and taking r_1 to r_2.

We begin with the hyperboloid model. As we saw in Section 6.1, the construction of the hyperboloid model $\mathbb{U}$ began with the quadratic form

$$Q(\mathbf{x}) = -x_0^2 + x_1^2 + x_2^2$$

on $\mathbb{R}^3$. From this quadratic form, we derived the group of reasonable transformations of $\mathbb{U}$, namely, the matrices in $\mathcal{O}^+(Q) = \mathrm{O}(2,1)$ that preserve this quadratic form. The definition of hyperbolic length and the definition of hyperbolic line were both given directly, and they were shown to be invariant under the action of $\mathcal{O}^+(Q)$.

This construction is straightforward to generalize. To build the model space for the hyperboloid model $\mathbb{U}^n$ of n-dimensional hyperbolic space, we take the quadratic form

$$Q_n(\mathbf{x}) = -x_0^2 + \sum_{j=1}^{n} x_j^2$$

on $\mathbb{R}^{n+1}$. The model space $\mathbb{U}^n$ is then the upper sheet of the hyperboloid $\{\mathbf{x} \in \mathbb{R}^{n+1} \mid Q_n(\mathbf{x}) = -1\}$, namely, the set

$$\mathbb{U}^n = \{\mathbf{x} \in \mathbb{R}^{n+1} \mid Q_n(\mathbf{x}) = -1 \text{ and } x_0 > 0\}.$$

Using the same line of argument that we used for $\mathbb{U}$, we can determine the group

$$\mathcal{O}(Q_n) = \{A \in \mathrm{GL}_{n+1}(\mathbb{R}) \mid Q_n(\mathbf{x}) = Q_n(A\mathbf{x}) \text{ for all } \mathbf{x} \in \mathbb{R}^{n+1}\}$$

of invertible $(n+1) \times (n+1)$-matrices that keep this quadratic form invariant, and the group

$$\mathcal{O}^+(Q_n) = \mathrm{O}(n,1) = \{A \in \mathcal{O}(Q_n) \mid A(\mathbb{U}^n) = \mathbb{U}^n\}$$

of invertible $(n+1)\times(n+1)$-matrices that keep this quadratic form invariant and that preserve the upper sheet $\mathbb{U}^n$. The transitivity properties of the action of $\mathcal{O}^+(Q)$ on $\mathbb{U}$ also generalize to transitivity properties of the action of $\mathcal{O}^+(Q_n)$ on $\mathbb{U}^n$. In particular, the proof of Proposition 6.8 immediately generalizes, although the algebra becomes significantly more complicated. A generating set of matrices for $\mathcal{O}^+(Q_n)$ consists of the same three basic types of matrices as for $\mathcal{O}^+(Q)$:

- *Rotation* in the x_jx_k-plane fixing the x_0-axis, where $1 \le j, k \le n$. Rotation by $\alpha \in \mathbb{R}$ in the x_1x_2-plane is given by the matrix
$$\begin{pmatrix} 1 & 0 & 0 & (0) \\ 0 & \cos(\alpha) & -\sin(\alpha) & (0) \\ 0 & \sin(\alpha) & \cos(\alpha) & (0) \\ (0) & (0) & (0) & I_{n-2} \end{pmatrix},$$
where I_{n-2} is the $(n-2)\times(n-2)$ identity matrix and where (0) represents the 0-matrix of the appropriate size. Rotation in the x_jx_k-plane is obtained by permuting the coordinates $x_1, \ldots, x_n$ appropriately.

- *Translation* in the x_0x_j-plane, where $1 \le j \le n$. Translation by $\mu \in \mathbb{R}$ in the x_0x_1-plane is given by the matrix
$$\begin{pmatrix} \cosh(\mu) & \sinh(\mu) & 0 & (0) \\ \sinh(\mu) & \cosh(\mu) & 0 & (0) \\ 0 & 0 & 1 & (0) \\ (0) & (0) & (0) & I_{n-2} \end{pmatrix},$$
where I_{n-2} is the $(n-2)\times(n-2)$ identity matrix and where (0) represents the 0-matrix of the appropriate size. Translation in the x_0x_k-plane is obtained by permuting the coordinates x_1 and x_k.

- *Reflection* in the plane $\{\mathbf{x} \in \mathbb{R}^{n+1} \mid x_j = 0\}$, where $1 \le j \le n$. Reflection in the plane $\{\mathbf{x} \in \mathbb{R}^{n+1} \mid x_1 = 0\}$ is given by the matrix
$$\begin{pmatrix} 1 & 0 & (0) \\ 0 & -1 & (0) \\ (0) & (0) & I_{n-1} \end{pmatrix},$$
where I_{n-1} is the $(n-1)\times(n-1)$ identity matrix and where (0) represents the 0-matrix of the appropriate size. Reflection in the plane $\{\mathbf{x} \in \mathbb{R}^{n+1} \mid x_j = 0\}$ is obtained by permuting the coordinates x_1 and x_j.

As in $\mathbb{U}$, a nonzero tangent vector $\mathbf{t}$ to $\mathbb{U}^n$ satisfies $Q_n(\mathbf{t}) > 0$, and so the hyperbolic length of a piecewise C^1 path $f : [a, b] \to \mathbb{U}^n$ is given by
$$\text{length}_{\mathbb{U}^n}(f) = \int_a^b \sqrt{Q(f'(t))}\mathrm{d}t.$$

The *hyperbolic distance* $\mathrm{d}_{\mathbb{U}^n}(\mathbf{u},\mathbf{v})$ between points $\mathbf{u}$ and $\mathbf{v}$ of $\mathbb{U}^n$ is then defined to be the infimum of the hyperbolic lengths of all piecewise C^1 paths $f:[a,b]\to\mathbb{U}^n$ with $f(a)=\mathbf{u}$ and $f(b)=\mathbf{v}$. The group of isometries of the metric space $(\mathbb{U}^n,\mathrm{d}_{\mathbb{U}^n})$ is then

$$\mathrm{Isom}(\mathbb{U}^n,\mathrm{d}_{\mathbb{U}^n})=\mathcal{O}^+(Q_n).$$

Again, a hyperbolic line is defined to be the intersection of a two-dimensional Euclidean subspace of $\mathbb{R}^{n+1}$ through the origin $\mathbf{0}$ with $\mathbb{U}^n$. Parallelism in $\mathbb{U}^n$ behaves as it does in $\mathbb{U}$, so that given a hyperbolic line ℓ in $\mathbb{U}^n$ and a point $p\in\mathbb{U}^n$ not on ℓ, there exist at least two hyperbolic lines in $\mathbb{U}^n$ passing through p and parallel to ℓ. The argument proving this result in $\mathbb{U}^n$ is essentially the same as the argument given for the analogous fact in $\mathbb{U}$. Also, we have that the distance-realizing paths in $\mathbb{U}^n$ are the almost simple parametrizations of hyperbolic line segments.

We note that this construction of a model $\mathbb{U}^n$ of n-dimensional hyperbolic space also yields that there is a hierarchy of hyperbolic spaces of increasing (or decreasing, depending on the point of view) dimension contained in $\mathbb{U}^n$. We did not remark on this fact when developing our models of the hyperbolic plane, as we never explicitly described a hyperbolic line as a model of one-dimensional hyperbolic space.

So, consider the n-dimensional subspace Y_n of $\mathbb{R}^{n+1}$ given by the condition

$$Y_n=\{\mathbf{x}\in\mathbb{R}^{n+1}\mid x_n=0\}.$$

As $Y_n=\mathbb{R}^n\times\{0\}\subset\mathbb{R}^{n+1}$, we identify Y_n with the n-dimensional Euclidean space $\mathbb{R}^n$ with coordinates $x_0,\ldots,x_{n-1}$. The restriction of Q_n to Y_n is the quadratic form $Q_{n-1}=-x_0^2+\sum_{j=1}^{n-1}x_j^2$ on $\mathbb{R}^n$, and the intersection $Y_n\cap\mathbb{U}^n$ is the set

$$Y_n\cap\mathbb{U}^n=\{\mathbf{x}\in\mathbb{R}^{n+1}\mid x_n=0,\ x_0>0,\ \text{and } Q_n(\mathbf{x})=-1\},$$

which (forgetting the coordinate x_n) is the same as the set

$$\mathbb{U}^{n-1}=\{\mathbf{x}\in\mathbb{R}^n\mid x_0>0\text{ and }Q_{n-1}(\mathbf{x})=-1\}$$

(where we freely make use of this identification of Y_n and $\mathbb{R}^n$). Therefore, the set $Y_n\cap\mathbb{U}^n$ is just the hyperboloid model $\mathbb{U}^{n-1}$ of hyperbolic $(n-1)$-space. In fact, if P is any n-dimensional subspace of $\mathbb{R}^{n+1}$ whose intersection with $\mathbb{U}^n$ is nonempty, then there exists an element of $\mathcal{O}^+(Q_n)$ taking P to Y_n, and so $P\cap\mathbb{U}^n$ is also a copy of the hyperboloid model of $\mathbb{U}^{n-1}$ contained within $\mathbb{U}^n$.

We can continue this process, considering the intersection of the $(n-1)$-dimensional subspace

$$Y_{n-1,n}=\{\mathbf{x}\in\mathbb{R}^{n+1}\mid x_n=0,\ x_{n-1}=0\}$$

of $\mathbb{R}^{n+1}$ with $\mathbb{U}^n$ to get a copy $Y_{n-1,n} \cap \mathbb{U}^n$ of the hyperboloid model $\mathbb{U}^{n-2}$ contained within the copy $Y_n \cap \mathbb{U}^n$ of the hyperboloid model $\mathbb{U}^{n-1}$, which is itself contained in $\mathbb{U}^n$.

One advantage to the hyperboloid model of hyperbolic n-space is that the group of isometries of hyperbolic n-space is immediately realized as a group of matrices, because the group of isometries of the hyperbolic metric on $\mathbb{U}^n$ is just $\mathcal{O}^+(Q)$. Without going into detail, we note that this linearity of the group of isometries has strong consequences.

As a last point regarding the hyperboloid model, let G be the subgroup of $\mathcal{O}^+(Q_n)$ fixing the point $[1, 0, \ldots, 0]^{\mathrm{T}}$ of $\mathbb{U}^n$. For each $r \geq 1$, the action of G preserves the $(n-1)$-dimensional affine subspace $P_r = \{\mathbf{x} \in \mathbb{R}^{n+1} \mid x_0 = r\}$ of $\mathbb{R}^{n+1}$, and hence the subset $Z_r = \{\mathbf{u} \in \mathbb{U}^n \mid u_0 = r\}$ of $\mathbb{U}^n$, which is a Euclidean (and hyperbolic) $(n-1)$-dimensional sphere. The restriction of the quadratic form Q_n to P_r yields a multiple of the standard Euclidean metric on $\{\mathbf{x} \in \mathbb{R}^{n+1} \mid x_0 = r\}$, and G is exactly the group of Euclidean isometries of this $(n-1)$-dimensional sphere with this restricted metric. That is, inside n-dimensional hyperbolic geometry $\mathbb{U}^n$, we have a copy of $(n-1)$-dimensional spherical geometry.

We can also generalize the upper half-plane and Poincaré disc models of the hyperbolic plane to models of hyperbolic n-space. When we were constructing the upper half-plane and Poincaré disc models of the hyperbolic plane, we worked in the complex plane $\mathbb{C}$ and could make use of a great deal of the machinery of complex analysis, which was the content of Chapter 4. However, in general, this complex structure, with its pleasant and useful analytic consequence, is not available.

In the same way that the upper half-plane and the Poincaré disc models of the hyperbolic plane were two manifestations of the same general construction, the *upper half-space* and *Poincaré ball* models of hyperbolic n-space are manifestations of the same general construction. We work in $\overline{\mathbb{R}^n} = \mathbb{R}^n \cup \{\infty\}$, the *one-point compactification* of $\mathbb{R}^n$, where the coordinates on $\mathbb{R}^n$ are $x_1, \ldots, x_n$. As we did in $\overline{\mathbb{C}}$, define a *sphere in* $\overline{\mathbb{R}^n}$ to be either an $(n-1)$-dimensional Euclidean sphere in $\mathbb{R}^n$ or the union of an $(n-1)$-dimensional affine subspace in $\mathbb{R}^n$ with $\{\infty\}$. As before, we can define *reflection* (or *inversion*) in a sphere in $\overline{\mathbb{R}^n}$. There are explicit formulae for reflection in a sphere in $\overline{\mathbb{R}^n}$ similar to the formulae for reflection in a circle in $\overline{\mathbb{C}}$, but as we will not explicitly use them, we do not give them here.

The *general Möbius group* Möb_n of $\overline{\mathbb{R}^n}$ is generated by reflections in spheres in $\overline{\mathbb{R}^n}$. A *Möbius transformation* of $\overline{\mathbb{R}^n}$ is the composition of reflections in an even number of spheres in $\overline{\mathbb{R}^n}$. The group Möb_n^+ of Möbius transformations of $\overline{\mathbb{R}^n}$ has index 2 in the general Möbius group Möb_n.

The group Möb_n^+ of Möbius transformations of $\overline{\mathbb{R}^n}$ shares many basic properties of the group $\text{Möb}^+ = \text{Möb}_2^+$ of Möbius transformations of $\overline{\mathbb{C}}$. Namely, Möb_n^+ acts conformally on $\overline{\mathbb{R}^n}$, so that angles are preserved. Moreover, Möb_n^+ acts transitively on $\overline{\mathbb{R}^n}$ and acts transitively on the collection of all $(n-1)$-dimensional spheres in $\overline{\mathbb{R}^n}$ and on the collection of all n-dimensional balls in $\overline{\mathbb{R}^n}$.

The base space of the *upper half-space model* $\mathbb{H}^n$ of hyperbolic n-space is the upper half-space of $\mathbb{R}^n$, namely

$$\mathbb{H}^n = \{\mathbf{x} \in \mathbb{R}^n \mid x_n > 0\},$$

with its *boundary at infinity*

$$(\mathbb{R}^{n-1} \times \{0\}) \cup \{\infty\} = \overline{\mathbb{R}^{n-1}}.$$

The base space of the *Poincaré ball model* $\mathbb{D}^n$ of hyperbolic n-space is the unit ball in $\mathbb{R}^n$, namely

$$\mathbb{D}^n = \{\mathbf{x} \in \mathbb{R}^n \mid |\mathbf{x}| < 1\},$$

with its *boundary at infinity*

$$\mathbb{S}^{n-1} = \{\mathbf{x} \in \mathbb{R}^n \mid |\mathbf{x}| = 1\}.$$

We note that, by the transitivity of the action of Möb_n^+ on the collection of all n-dimensional balls in $\overline{\mathbb{R}^n}$, there exists an element m of Möb_n^+ taking $\mathbb{H}^n$ to $\mathbb{D}^n$. Hence, we work primarily with $\mathbb{H}^n$, with the understanding that analogous statements hold for $\mathbb{D}^n$.

For any $1 \leq k \leq n-1$, define a *k-dimensional sphere in $\overline{\mathbb{R}^n}$* to be either a k-dimensional Euclidean sphere in $\mathbb{R}^n$ or the union of a k-dimensional affine subspace of $\mathbb{R}^n$ with $\{\infty\}$. Hence, a one-dimensional sphere in $\overline{\mathbb{R}^n}$ is either a Euclidean circle in $\mathbb{R}^n$ or the union of a Euclidean line in $\mathbb{R}^n$ with $\{\infty\}$, whereas an $(n-1)$-dimensional sphere in $\overline{\mathbb{R}^n}$ is the same as a sphere in $\overline{\mathbb{R}^n}$ as defined above. Analogously with the upper half-plane model of the hyperbolic plane, we define a *hyperbolic line* in $\mathbb{H}^n$ to be the intersection of $\mathbb{H}^n$ with a one-dimensional sphere in $\overline{\mathbb{R}^n}$ that is perpendicular to the boundary at infinity $\overline{\mathbb{R}^{n-1}}$ of $\mathbb{H}^n$. Hence, a hyperbolic line in $\mathbb{H}^n$ is either a vertical Euclidean line or is the upper hemi-circle of a Euclidean circle centred on $\overline{\mathbb{R}^{n-1}}$ and perpendicular to $\overline{\mathbb{R}^{n-1}}$.

The hyperbolic element of arc-length on $\mathbb{H}^n$ is $\frac{1}{x_n}|\mathrm{d}\mathbf{x}|$, so that the hyperbolic length of the piecewise C^1 path $f : [a,b] \to \mathbb{H}^n$ given by

$$f(t) = (f_1(t), \ldots, f_n(t))$$

is

$$\text{length}_{\mathbb{H}^n}(f) = \int_f \frac{1}{x_n}|\mathrm{d}\mathbf{x}| = \int_a^b \frac{1}{f_n(t)}|f'(t)|\mathrm{d}t.$$

As we have a means of measuring the hyperbolic lengths of piecewise C^1 paths in $\mathbb{H}^n$, we define the hyperbolic metric $\mathrm{d}_{\mathbb{H}^n}$ on $\mathbb{H}^n$ by defining the hyperbolic distance $\mathrm{d}_{\mathbb{H}^n}(\mathbf{x}, \mathbf{y})$ between points $\mathbf{x}$ and $\mathbf{y}$ of $\mathbb{H}^n$ to be the infimum of the hyperbolic lengths of all piecewise C^1 paths $f : [a, b] \to \mathbb{H}^n$ with $f(a) = \mathbf{x}$ and $f(b) = \mathbf{y}$. This notion of hyperbolic length, and hence of hyperbolic distance on $\mathbb{H}^n$, is invariant under the action of the subgroup $\text{Möb}_n(\mathbb{H}^n)$ of Möb_n preserving $\mathbb{H}^n$. The group of isometries of the metric space $(\mathbb{H}^n, \mathrm{d}_{\mathbb{H}^n})$ is then

$$\text{Isom}(\mathbb{H}^n, \mathrm{d}_{\mathbb{H}^n}) = \text{Möb}_n(\mathbb{H}^n).$$

For $r > 0$, let H_r be the affine $(n-1)$-dimensional subspace

$$H_r = \{\mathbf{x} \in \mathbb{H}^n \mid x_n = r\}$$

of $\mathbb{R}^n$ contained in $\mathbb{R}^n$. The restriction of the hyperbolic metric $\mathrm{d}_{\mathbb{H}^n}$ to H_r is a multiple of the usual Euclidean metric on $\mathbb{R}^{n-1}$. Moreover, the subgroup G of Möb_n keeping, say, H_1 invariant is a subgroup of the subgroup of Möb_n fixing ∞. Note that G keeps H_r invariant for each $r > 0$, and that G acts as the full group of isometries of the Euclidean metric on H_r. Hence, n-dimensional hyperbolic geometry contains a copy of $(n-1)$-dimensional Euclidean geometry. This result, together with the earlier remark that n-dimensional hyperbolic geometry contains a copy of $(n-1)$-dimensional spherical geometry, gives credence to the claim that hyperbolic geometry is in some weak sense universal.

We close this section by noting that, unfortunately, there is in general not a convenient form to express Möbius transformations of $\overline{\mathbb{R}^n}$. We saw that for $\overline{\mathbb{C}} = \overline{\mathbb{R}^2}$, we could express Möbius transformations as quotients of linear polynomials in the coordinate z on $\mathbb{C}$, with a suitable interpretation of the arithmetic of $\{\infty\}$, but no such representation exists for higher dimensions. A good reference for Möbius transformations in higher dimensions is Ahlfors [5]. There is also the paper of Abikoff [1], which discusses another way of realizing the upper half-space model of three-dimensional hyperbolic space $\mathbb{H}^3$, together with its group of isometries.

Solutions to Exercises

Solutions to Chapter 1 exercises:

1.1: Write $x = \mathrm{Re}(z) = \frac{1}{2}(z+\overline{z})$ and $y = \mathrm{Im}(z) = -\frac{i}{2}(z-\overline{z})$, so that

$$ax+by+c = a\frac{1}{2}(z+\overline{z}) - b\frac{i}{2}(z-\overline{z}) + c = \frac{1}{2}(a-ib)z + \frac{1}{2}(a+bi)\overline{z} + c = 0.$$

Note that the slope of this line is $-\frac{a}{b}$, which is the quotient of the imaginary and real parts of the coefficient of z.

Given the circle $(x-h)^2 + (y-k)^2 = r^2$, set $z_0 = h+ik$ and rewrite the equation of the circle as

$$|z-z_0|^2 = z\overline{z} - \overline{z_0}z - z_0\overline{z} + |z_0|^2 = r^2.$$

1.2: A and $\mathbb{S}^1$ are perpendicular if and only if their tangent lines at the point of intersection are perpendicular. Let x be a point of $A \cap \mathbb{S}^1$, and consider the Euclidean triangle T with vertices 0, $re^{i\theta}$, and x. The sides of T joining x to the other two vertices are radii of A and $\mathbb{S}^1$, and so A and $\mathbb{S}^1$ are perpendicular if and only if the interior angle of T at x is $\frac{1}{2}\pi$, which occurs if and only if the Pythagorean theorem holds, which occurs if and only if $s^2 + 1^2 = r^2$.

1.3: Let L_{pq} be the Euclidean line segment joining p and q. The midpoint of L_{pq} is $\frac{1}{2}(p+q)$, and the slope of L_{pq} is $m = \frac{\mathrm{Im}(q)-\mathrm{Im}(p)}{\mathrm{Re}(q)-\mathrm{Re}(p)}$. The perpendicular bisector K of L_{pq} passes through $\frac{1}{2}(p+q)$ and has slope $-\frac{1}{m} = \frac{\mathrm{Re}(p)-\mathrm{Re}(q)}{\mathrm{Im}(q)-\mathrm{Im}(p)}$, and so K has the equation

$$y - \frac{1}{2}\left(\mathrm{Im}(p)+\mathrm{Im}(q)\right) = \left[\frac{\mathrm{Re}(p)-\mathrm{Re}(q)}{\mathrm{Im}(q)-\mathrm{Im}(p)}\right]\left(x - \frac{1}{2}(\mathrm{Re}(p)+\mathrm{Re}(q))\right).$$

The Euclidean centre c of A is the x-intercept of K, which is

$$\begin{aligned} c &= \left[-\frac{1}{2}(\mathrm{Im}(p)+\mathrm{Im}(q))\right]\left[\frac{\mathrm{Im}(q)-\mathrm{Im}(p)}{\mathrm{Re}(p)-\mathrm{Re}(q)}\right] \\ &\quad +\frac{1}{2}(\mathrm{Re}(p)+\mathrm{Re}(q)) \\ &= \frac{1}{2}\left[\frac{(\mathrm{Im}(p))^2-(\mathrm{Im}(q))^2+(\mathrm{Re}(p))^2-(\mathrm{Re}(q))^2}{\mathrm{Re}(p)-\mathrm{Re}(q)}\right] \\ &= \frac{1}{2}\left[\frac{|p|^2-|q|^2}{\mathrm{Re}(p)-\mathrm{Re}(q)}\right]. \end{aligned}$$

The Euclidean radius of A is

$$r = |c-p| = \left|\frac{1}{2}\left[\frac{|p|^2-|q|^2}{\mathrm{Re}(p)-\mathrm{Re}(q)}\right]-p\right|.$$

1.4: One hyperbolic line through i that is parallel to ℓ is the positive imaginary axis $I=\mathbb{H}\cap\{\mathrm{Re}(z)=0\}$. To get a second hyperbolic line through i and parallel to ℓ, take any point x on $\mathbb{R}$ between 0 and 3, say $x=2$, and consider the Euclidean circle A centred on $\mathbb{R}$ through 2 and i.

By Exercise 1.3, the Euclidean centre c of A is $c=\frac{3}{4}$ and the Euclidean radius of A is $r=\frac{5}{4}$. As the real part of every point on A is at most 2, the hyperbolic line $\mathbb{H}\cap A$ is a hyperbolic line passing through i that is parallel to ℓ.

1.5: The Euclidean circle D through i and concentric to A has Euclidean centre -2 and Euclidean radius $\sqrt{5}=|i-(-2)|$, and so one hyperbolic line through i parallel to A is $\mathbb{H}\cap D$.

To construct a second hyperbolic line through i parallel to ℓ, start by taking a point x on $\mathbb{R}$ between A and D, say $x=-4$. Let E be the Euclidean circle centred on $\mathbb{R}$ passing through -4 and i. By Exercise 1.3, the Euclidean centre c of E is $c=-\frac{15}{8}$ and the Euclidean radius is $r=\frac{17}{8}$.

It is an easy calculation that the two Euclidean circles $\{z\in\mathbb{C}\mid |z+2|=1\}$ and $\{z\in\mathbb{C}\mid \left|z+\frac{15}{8}\right|=\frac{17}{8}\}$ are disjoint, and so the hyperbolic line $\mathbb{H}\cap E$ is a hyperbolic line passing through i that is parallel to ℓ.

1.6: If $c=0$, then the Euclidean line L_c passing through i and 0 intersects $\mathbb{S}^1$ at $\pm i$, and so $\xi^{-1}(0)=-i$.

Given a point $c\neq 0$ in $\mathbb{R}$, the equation of the Euclidean line L_c passing through c and i is

$$y=-\frac{1}{c}(x-c)=-\frac{1}{c}x+1.$$

To find where L_c intersects $\mathbb{S}^1$, we find the values of x for which

$$|x + i\,y| = \left| x + i \left(-\frac{1}{c} x + 1 \right) \right| = 1,$$

which simplifies to

$$x \left[\left(1 + \frac{1}{c^2} \right) x - \frac{2}{c} \right] = 0.$$

As $x = 0$ corresponds to i, we have that

$$x = \frac{2c}{c^2 + 1}.$$

So,

$$\xi^{-1}(c) = \frac{2c}{c^2+1} + i\,\frac{c^2-1}{c^2+1}.$$

1.7: Calculating,

$$\xi(1) = 1;\ \xi\left(\exp\left(\frac{2\pi}{3} i\right)\right) = \frac{1}{\sqrt{3}-2};\ \text{ and } \xi\left(\exp\left(\frac{4\pi}{3} i\right)\right) = \frac{-1}{\sqrt{3}+2}.$$

1.8: Let z be a point of $\mathbb{H}$. The Euclidean distance from z to $\mathbb{R}$ is $\text{Im}(z)$. So, $U_{\text{Im}(z)}(z)$ is contained in $\mathbb{H}$, but $U_\varepsilon(z)$ is not contained in $\mathbb{H}$ for any $\varepsilon > \text{Im}(z)$.

1.9: Recall that K is bounded if and only if there exists some $M > 0$ so that K is contained in $U_M(0)$. In particular, we have that $U_M(\infty)$ is contained in X. For any $z \in \mathbb{C} - K$, the Euclidean distance $\delta(z)$ from z to K is positive, because K is closed, and so $U_\varepsilon(z)$ is contained in X for any $0 < \varepsilon < \delta(z)$. Hence, the complement X of K in $\overline{\mathbb{C}}$ is open.

Suppose that W is an open subset of $\overline{\mathbb{C}}$. If $\infty \notin W$, then W is contained in $\mathbb{C}$, and by the definition of $U_\varepsilon(z)$, we have that W is open in $\mathbb{C}$.

If $\infty \in W$, then $U_\varepsilon(\infty)$ is contained in W for some $\varepsilon > 0$. For this same choice of ε, we have that the complement $Y = \overline{\mathbb{C}} - W$ of W is contained in $U_\varepsilon(0)$, and so Y is bounded. The fact that Y is closed follows immediately from the fact that its complement $\overline{\mathbb{C}} - Y = W$ is open in $\overline{\mathbb{C}}$.

1.10: Given $\varepsilon > 0$, we need to find $N > 0$ so that $z_n = \frac{1}{n} \in U_\varepsilon(0)$ for $n > N$. Take $N = \frac{1}{\varepsilon}$. Then, for $n > N$, we have that $z_n = \frac{1}{n} < \varepsilon$, as desired.

Given $\varepsilon > 0$, we need to find $N > 0$ so that $w_n = n \in U_\varepsilon(\infty)$ for $n > N$. Take $N = \varepsilon$. Then, for $n > N$, we have that $w_n = n > \varepsilon$, and so $w_n \in U_\varepsilon(\infty)$, as desired.

1.11: Note that 0 lies in $\overline{X}$, because $\frac{1}{n} \in U_\varepsilon(0) \cap X$ for every $n > \varepsilon$. However, there are no other points of $\overline{X}$ other than 0 and the points of X.

If $z \in \mathbb{C}$ is any point with $\mathrm{Im}(z) \neq 0$, then $U_{\mathrm{Im}(z)}(z) \cap X = \emptyset$. Also, because $|x| \leq 1$ for every $x \in X$, we see that $U_2(\infty) \cap X = \emptyset$.

If $z \in \mathbb{R}$ is any point with $\mathrm{Re}(z) \neq 0$ and $\mathrm{Re}(z) \neq \frac{1}{n}$ for $n \in \mathbb{Z}-\{0\}$, then either z lies between $\frac{1}{m}$ and $\frac{1}{p}$ for some $m, p \in \mathbb{Z}-\{0\}$, or else z lies in one of the intervals $(1, \infty)$ or $(-\infty, -1)$. In the former case, let $\varepsilon = \min\left(\left|z - \frac{1}{m}\right|, \left|z - \frac{1}{p}\right|\right)$, so that $U_\varepsilon(z) \cap X = \emptyset$. In the latter case, let $\varepsilon = |z-1|$ if $z \in (1, \infty)$ or $\varepsilon = |z+1|$ if $z \in (-\infty, -1)$, so that $U_\varepsilon(z) \cap X = \emptyset$.

Hence, $\overline{X} = X \cup \{0\}$.

For Y, take any $z = x + y\,i \in \mathbb{C}$. Given any $\varepsilon > 0$, there exist rational numbers a, b so that $|x - a| < \frac{1}{2}\varepsilon$ and $|y - b| < \frac{1}{2}\varepsilon$, because $\mathbb{Q}$ is dense in $\mathbb{R}$, a fact that can be proven by considering decimal expansions.

Then, $|(x + y\,i) - (a + b\,i)| < \varepsilon$. As for each $\varepsilon > 0$ we can construct a point in $U_\varepsilon(z) \cap Y$, we have that every point of $\mathbb{C}$ lies in $\overline{Y}$.

As for any $\varepsilon > 0$ we have that $n \in U_\varepsilon(\infty)$ for every integer n with $n > \varepsilon$, we also have that $\infty \in \overline{Y}$.

Hence, $\overline{Y} = \overline{\mathbb{C}}$.

1.12: To show that $\overline{X}$ is closed in $\overline{\mathbb{C}}$, we show that $\overline{\mathbb{C}} - \overline{X}$ is open in $\overline{\mathbb{C}}$. Take $z \in \overline{\mathbb{C}} - \overline{X}$.

Suppose that for each $\varepsilon > 0$, the intersection $U_\varepsilon(z) \cap \overline{X} \neq \emptyset$. For each $n \in \mathbb{N}$, choose some $z_n \in U_{1/n}(z) \cap \overline{X}$. As $z_n \in \overline{X}$, there is some $x_n \in X$ so that $x_n \in U_{1/n}(z_n) \cap X$.

Combining these calculations yields that $|x_n - z| \leq |x_n - z_n| + |z_n - z| < \frac{2}{n}$. Hence, for each $n \in \mathbb{N}$, we have that $x_n \in U_{2/n}(z) \cap X$, which implies that $z \in \overline{X}$. This result contradicts our original choice of z.

1.13: The Euclidean line L_{P} can be expressed parametrically as

$$\mathrm{N} + t(\mathrm{P} - \mathrm{N}) = (tp_1, tp_2, tp_3 + 1 - t)$$

for $t \in \mathbb{R}$. L_{P} intersects the x_1x_2-plane when $tp_3 + 1 - t = 0$, that is, when $t = \frac{1}{1-p_3}$. Hence, we see that

$$\xi(\mathrm{P}) = \frac{p_1}{1-p_3} + i\frac{p_2}{1-p_3}.$$

For ξ^{-1}, let $z = x + iy$ be any point of $\mathbb{C}$, and note that z corresponds to the point $Z = (x, y, 0)$ in $\mathbb{R}^3$. Let L be the Euclidean line between N and Z, and note that L is given parametrically by

$$\mathrm{N} + t(Z - \mathrm{N}) = (tx, ty, 1 - t)$$

for $t \in \mathbb{R}$. To find where L intersects $\mathbb{S}^2$, we find the point on L whose distance from the origin is 1, which involves solving

$$(tx)^2 + (ty)^2 + (1-t)^2 = t^2|z|^2 + t^2 - 2t + 1 = 1$$

for t. There are two solutions: $t = 0$, which corresponds to N, and $t = \frac{2}{1+|z|^2}$. The latter value of t yields

$$\xi^{-1}(z) = \left(\frac{2\,\mathrm{Re}(z)}{|z|^2+1}, \frac{2\,\mathrm{Im}(z)}{|z|^2+1}, \frac{|z|^2-1}{|z|^2+1} \right).$$

1.14: Write

$$g(z) = a_n z^n + \cdots + a_1 z + a_0,$$

for $n \geq 1$ with $a_n \neq 0$. We need to quantify the statement that, if $|z|$ is large, then $|g(z)|$ is large. If we wanted to be precise, we could proceed as follows. By the triangle inequality,

$$|g(z)| \geq |\,|a_n z^n| - |a_{n-1}z^{n-1} + \cdots + a_1 z + a_0|\,|.$$

So, set $A = \max\{|a_{n-1}|, \ldots, |a_0|\}$ and note that

$$|a_{n-1}z^{n-1} + \cdots + a_1 z + a_0| \leq A\left(|z^{n-1}| + \cdots + |z| + 1\right) \leq nA|z|^{n-1}$$

for $|z| \geq 1$.

So, given $\varepsilon > 0$, choose $\delta > 0$ so that $\delta > 1$ and so that $|a_n|\delta^n - nA\delta^{n-1} > \varepsilon$. Then, for $|z| > \delta$, we have that

$$\begin{array}{rcl} |g(z)| & \geq & |a_n||z|^n - |a_{n-1}z^{n-1} + \cdots + a_1 z + a_0| \\ & \geq & |a_n||z|^n - nA|z|^{n-1} \\ & \geq & \delta^{n-1}(|a_n|\delta - nA) > \varepsilon, \end{array}$$

as desired.

1.15: If $d = \mathrm{degree}(g) \geq 2$, then f is not a bijection. In fact, the fundamental theorem of algebra gives that there is a point c of $\overline{\mathbb{C}}$ so that there are at least two distinct solutions to $g(z) = c$. If $g(z)$ does not factor as $g(z) = (z-a)^d$, then we may take $c = 0$.

If $g(z) = (z-a)^d$, then we may take $c = 1$, so that the solutions to $g(z) = 1$ are

$$\left\{ z = a + \exp\left(\frac{2\pi k}{d} i\right) \mid 0 \leq k \leq d \right\}.$$

If $d = \mathrm{degree}(g) = 0$, then f is a constant function, which is continuous but is not a bijection.

If $d = \text{degree}(g) = 1$, then $g(z) = az + b$ where $a \neq 0$. We know from Exercise 1.14 that f is continuous. To see that f is a bijection and that f^{-1} is continuous, we write an explicit expression for f^{-1}, namely,

$$f^{-1}(z) = \frac{1}{a}(z - b) \text{ for } z \in \mathbb{C} \text{ and } f^{-1}(\infty) = \infty.$$

1.16: Choose any $z \in \overline{\mathbb{C}}$. As X is dense in $\overline{\mathbb{C}}$, there exists a sequence $\{x_n\}$ in X converging to z.

As f is continuous, we know that $\{f(x_n)\}$ converges to $f(z)$. As $f(x_n) = x_n$, this gives that $\{x_n\}$ converges to both z and $f(z)$, and so $z = f(z)$.

1.17: Let C_k be the circle in $\overline{\mathbb{C}}$ containing ℓ_k. There are several cases to consider, depending on whether C_1 and C_2 are both Euclidean circles, whether both are Euclidean lines, or whether one is a Euclidean line and one is a Euclidean circle.

We give a complete answer in the case in which both C_1 and C_2 are Euclidean circles; the other two cases follow by a similar argument. In essence, we are reproving the fact that if C_1 and C_2 are two Euclidean circles in $\mathbb{C}$, then there exists a circle perpendicular to both C_1 and C_2 if and only if C_1 and C_2 are disjoint.

Let c_k be the Euclidean centre of C_k, and let r_k be its Euclidean radius. Suppose there exists a hyperbolic line ℓ that is perpendicular to both ℓ_1 and ℓ_2. Let A be the circle in $\overline{\mathbb{C}}$ containing ℓ.

It may be that A is a Euclidean line, in which case, C_1 and C_2 are concentric, and so are ultraparallel, because the boundary at infinity of the closed region in $\mathbb{H}$ bounded by ℓ_1 and ℓ_2 is the union of two closed arcs in $\overline{\mathbb{R}}$.

Otherwise, A is a Euclidean circle with Euclidean centre c and Euclidean radius r. Then, A intersects both C_1 and C_2 perpendicularly, and so $|c - c_k|^2 = r^2 + r_k^2$ for $k = 1, 2$ by the Pythagorean theorem. In particular, $|c - c_k| > r_k$. Hence, the boundary at infinity of the closed region in $\mathbb{H}$ bounded by ℓ_1 and ℓ_2 consists of two closed arcs, one containing c and the other containing ∞.

Conversely, suppose that ℓ_1 and ℓ_2 are ultraparallel, let H be the closed region in $\mathbb{H}$ bounded by them, and let a be the closed arc in the boundary at infinity of H not containing ∞.

Let C_k be the Euclidean circle containing ℓ_k, let c_k be the Euclidean centre of C_k, and let r_k be the Euclidean radius of C_k. For each $x \in a$, consider the Euclidean circle A with Euclidean centre x and Euclidean radius r. For A to be perpendicular to C_k, we need that $(c_k - x)^2 = r_k^2 + r^2$ for $k = 1, 2$.

Solving for x gives

$$x = \frac{r_1^2 - r_2^2 + c_2^2 - c_1^2}{2(c_2 - c_1)}.$$

As C_1 and C_2 are disjoint Euclidean circles, we have that $c_2 - c_1 > r_1 + r_2$. Hence,

$$r = \sqrt{(x - c_1)^2 - r_1^2} = \sqrt{(x - c_2)^2 - r_2^2} > 0,$$

and so the hyperbolic line contained in A is perpendicular to both ℓ_1 and ℓ_2.

1.18: Suppose that $q = \infty$. Then, the hyperbolic line contained in the Euclidean line $\{\mathrm{Re}(z) = \mathrm{Re}(p)\}$ is the unique hyperbolic line whose endpoints at infinity are p and q. A similiar argument holds if $p = \infty$.

Suppose that $p \neq \infty$ and that $q \neq \infty$. Then, we may again use the construction from the proof of Proposition 1.2 of the perpendicular bisector of the Euclidean line segment joining p to q to find the unique Euclidean circle centred on the real axis $\mathbb{R}$ that passes through both p and q. Intersecting this circle with $\mathbb{H}$ yields the unique hyperbolic line determined by p and q.

Solutions to Chapter 2 exercises:

2.1: Consider the function $f : \overline{\mathbb{C}} \to \overline{\mathbb{C}}$ given by

$$f(z) = \begin{cases} z & \text{for } \mathrm{Re}(z) \leq 0; \\ z + i\,\mathrm{Re}(z) & \text{for } \mathrm{Re}(z) \geq 0; \\ \infty & \text{for } z = \infty. \end{cases}$$

It is evident that f is continuous, as it is a sum of continuous functions. To see that f is a bijection and that f^{-1} is continuous, we give an explicit formula for f^{-1}, namely,

$$f^{-1}(z) = \begin{cases} z & \text{for } \mathrm{Re}(z) \leq 0; \\ z - i\,\mathrm{Re}(z) & \text{for } \mathrm{Re}(z) \geq 0; \\ \infty & \text{for } z = \infty. \end{cases}$$

Hence, we see that $f \in \mathrm{Homeo}(\overline{\mathbb{C}})$. However, the image of $\overline{\mathbb{R}}$ under f is not a circle in $\overline{\mathbb{C}}$, and so $f \notin \mathrm{Homeo}^{\mathrm{C}}(\overline{\mathbb{C}})$.

2.2: We follow the argument given in the proof of Proposition 2.1. Set $w = az + b$, so that $z = \frac{1}{a}(w - b)$. Substituting this calculation into the equation of a Euclidean circle, namely, $\alpha z\overline{z} + \beta z + \overline{\beta}\overline{z} + \gamma = 0$, yields

$$\begin{aligned} \alpha z\overline{z} + \beta z + \overline{\beta}\overline{z} + \gamma &= \alpha\frac{1}{a}(w-b)\overline{\frac{1}{a}(w-b)} + \beta\frac{1}{a}(w-b) + \overline{\beta}\overline{\frac{1}{a}(w-b)} + \gamma \\ &= \frac{\alpha}{|a|^2}(w-b)\overline{(w-b)} + \frac{\beta}{a}(w-b) + \overline{\left(\frac{\beta}{a}\right)}\overline{w-b} + \gamma \\ &= \frac{\alpha}{|a|^2}\left|w + \frac{\overline{\beta}a}{\alpha} - b\right|^2 + \gamma - \frac{|\beta|^2}{\alpha} = 0, \end{aligned}$$

which is again the equation of a Euclidean circle in $\mathbb{C}$.

2.3: The Euclidean circle A given by the equation $\alpha z\overline{z} + \beta z + \overline{\beta}\overline{z} + \gamma = 0$ has Euclidean centre $-\frac{\overline{\beta}}{\alpha}$ and Euclidean radius $\frac{1}{|\alpha|}\sqrt{|\beta|^2 - \alpha\gamma}$.

As we saw in the solution to Exercise 2.2, the image of A under f is the Euclidean circle given by the equation

$$\frac{\alpha}{|a|^2}\left|w + \frac{\overline{\beta}a}{\alpha} - b\right|^2 + \gamma - \frac{|\beta|^2}{\alpha} = 0,$$

which has Euclidean centre $f\left(-\frac{\overline{\beta}}{\alpha}\right)$ and Euclidean radius $|a|\frac{1}{|\alpha|}\sqrt{|\beta|^2 - \alpha\gamma}$.

2.4: The equation of A expands to

$$z\overline{z} - \overline{z_0}z - z_0\overline{z} + |z_0|^2 - r^2 = 0.$$

The circle $J(A)$ then has the equation

$$(|z_0|^2 - r^2)|w|^2 - z_0 w - \overline{z_0 w} + 1 = 0,$$

which is a Euclidean line if and only if $|z_0|^2 - r^2 = 0$, that is, if and only if $|z_0| = r$, so that A passes through 0.

2.5: The inverse of $m(z) = \frac{az+b}{cz+d}$ is $m^{-1}(z) = \frac{dz-b}{-cz+a}$.

2.6: If the coefficients of p are zero, then p is undefined. So, write $p(z) = \frac{az+b}{cz+d}$ and suppose that $ad - bc = 0$, so that $ad = bc$. Assume that $a \neq 0$. The proof in the cases that one (or several) of the other coefficients is nonzero is similar. Multiply the numerator and denominator of p by a and simplify to get

$$p(z) = \frac{az+b}{cz+d} = \frac{a(az+b)}{a(cz+d)} = \frac{a(az+b)}{acz+bc} = \frac{a(az+b)}{c(az+b)} = \frac{a}{c},$$

and so p is a constant function.

2.7:

1. As $m(\infty) = \frac{2}{3} \neq \infty$, the fixed points lie in $\mathbb{C}$ and are the roots of $3z^2 - 3z - 5 = 0$, which are $\frac{1}{6}[3 \pm \sqrt{69}]$.

2. One fixed point is $z = \infty$; the other lies in $\mathbb{C}$ and is the solution to $z = 7z+6$, which is $z = -1$.

3. As $J(\infty) = 0 \neq \infty$, the fixed points lie in $\mathbb{C}$ and are the roots of $z^2 = 1$, which are $z = \pm 1$.

4. As $m(\infty) = 1 \neq \infty$, the fixed points lie in $\mathbb{C}$ and are the roots of $z^2 = 0$; in particular, m has only one fixed point, at $z = 0$.

2.8: The general form of the Möbius transformation m taking the triple (∞, z_2, z_3) to the triple $(0, 1, \infty)$ is

$$m(z) = \frac{az+b}{cz+d} = \frac{z_2 - z_3}{z - z_3}.$$

2.9: There are six, as there are six permutations of T: $a(z) = z$ takes $(0, 1, \infty)$ to $(0, 1, \infty)$; $b(z) = -(z-1)$ takes $(0, 1, \infty)$ to $(1, 0, \infty)$; $c(z) = \frac{z}{z-1}$ takes $(0, 1, \infty)$ to $(0, \infty, 1)$; $d(z) = \frac{1}{z}$ takes $(0, 1, \infty)$ to $(\infty, 1, 0)$; $e(z) = \frac{-1}{z-1}$ takes $(0, 1, \infty)$ to $(1, \infty, 0)$; and $f(z) = \frac{z-1}{z}$ takes $(0, 1, \infty)$ to $(\infty, 0, 1)$.

2.10: There are many such transformations. One is the Möbius transformation m taking the triple $(i, -1, 1)$ of distinct points on $\mathbb{S}^1 = \partial\mathbb{D}$ to the triple $(0, 1, \infty)$ of distinct points on $\overline{\mathbb{R}} = \partial\mathbb{H}$. Explicitly,

$$m(z) = \frac{z-i}{z-1}\,\frac{-2}{-1-i}.$$

We still need to check that m takes $\mathbb{D}$ to $\mathbb{H}$, which we do by checking, for instance, that the imaginary part of $m(0)$ is positive:

$$\mathrm{Im}(m(0)) = \mathrm{Im}\left(\frac{2i}{1+i}\right) = \mathrm{Im}(1+i) > 0.$$

2.11: $f(z) = z^2$ is invariant under Möb$^+$ if and only if $f(m(z)) = f(z)$ for all $m(z) = \frac{az+b}{cz+d}$ in Möb$^+$ and all $z \in \overline{\mathbb{C}}$. That is, we need to have that

$$f(m(z)) = \left(\frac{az+b}{cz+d}\right)^2 = z^2,$$

and so

$$c^2z^4 + 2cdz^3 + (d^2 - a^2)z^2 - 2abz - b^2 = 0$$

for all z in $\overline{\mathbb{C}}$. In particular, we have that $c = b = 0$ and that $a^2 = d^2$, and so f is not invariant under Möb$^+$.

As $ad - bc = ad = 1$, this gives that either $a = d = \pm 1$ or that $a = -d = i$. In the former case, m is the identity Möbius transformation. In the latter case, $m(z) = -z$. Hence, the only subgroup of Möb$^+$ under which f is invariant is the subgroup $\langle e(z) = z,\ m(z) = -z\rangle$.

2.12: We proceed by direct calculation. Let $m(z) = \frac{az+b}{cz+d}$, where a, b, c, and d lie in $\mathbb{C}$ and $ad - bc = 1$. Then,

$$\begin{aligned}[m(z_1), m(z_2); m(z_3), m(z_4)] &= \left[\frac{az_1+b}{cz_1+d}, \frac{az_2+b}{cz_2+d}; \frac{az_3+b}{cz_3+d}, \frac{az_4+b}{cz_4+d}\right] \\ &= \left[\frac{\frac{az_1+b}{cz_1+d} - \frac{az_4+b}{cz_4+d}}{\frac{az_1+b}{cz_1+d} - \frac{az_2+b}{cz_2+d}}\right]\left[\frac{\frac{az_3+b}{cz_3+d} - \frac{az_2+b}{cz_2+d}}{\frac{az_3+b}{cz_3+d} - \frac{az_4+b}{cz_4+d}}\right]\end{aligned}$$

$$\begin{aligned}
&= \left[\frac{(az_1+b)(cz_4+d)-(az_4+b)(cz_1+d)}{(az_1+b)(cz_2+d)-(az_2+b)(cz_1+d)}\right]\times \\
&\quad \left[\frac{(az_3+b)(cz_2+d)-(az_2+b)(cz_3+d)}{(az_3+b)(cz_4+d)-(az_4+b)(cz_2+d)}\right] \\
&= \left[\frac{z_1-z_4}{z_1-z_2}\right]\left[\frac{z_3-z_2}{z_3-z_4}\right] \\
&= [z_1, z_2; z_3, z_4].
\end{aligned}$$

2.13: If there exists a continuous extension $F:\overline{\mathbb{C}}^4\to\overline{\mathbb{C}}$ of the cross ratio $[z_1,z_2;z_3,z_4]$, then for any two sequences $\{z_n\}$ and $\{w_n\}$ of distinct nonzero points of $\overline{\mathbb{C}}$ with

$$\lim_{n\to\infty} z_n = \lim_{n\to\infty} w_n = 0,$$

we have that

$$\lim_{n\to\infty} F(\infty, 0, w_n, z_n) = \lim_{n\to\infty} F(\infty, 0, z_n, w_n).$$

As 0, ∞, z_n, and w_n are distinct, this equation becomes

$$\lim_{n\to\infty}[\infty, 0; w_n, z_n] = \lim_{n\to\infty}[\infty, 0; z_n, w_n],$$

and so

$$\lim_{n\to\infty}\frac{w_n}{w_n-z_n} = \lim_{n\to\infty}\frac{z_n}{z_n-w_n}.$$

For example, take $z_n=\frac{2}{n}$ and $w_n=\frac{1}{n}$. Hence,

$$-1 = 2,$$

which is a contradiction.

2.14: Calculating, we see that

$$\begin{aligned}
[2+3i, -2i; 1-i, 4] &= \left[\frac{2+3i-4}{2+3i+2i}\right]\left[\frac{1-i+2i}{1-i-4}\right] \\
&= \left[\frac{-2+3i}{2+5i}\right]\left[\frac{1+i}{-3-i}\right] \\
&= \left[\frac{11+16i}{29}\right]\left[\frac{-4-2i}{10}\right] = \frac{-12-86i}{290},
\end{aligned}$$

which is not real, and so $2+3i$, $-2i$, $1-i$, and 4 do not lie on a circle in $\overline{\mathbb{C}}$.

2.15: Calculating, we see that

$$\begin{aligned}
[2+3i, -2i; 1-i, s] &= \left[\frac{2+3i-s}{2+3i+2i}\right]\left[\frac{1-i+2i}{1-i-s}\right] \\
&= \left[\frac{2-s+3i}{2+5i}\right]\left[\frac{1+i}{1-s-i}\right]
\end{aligned}$$

$$\begin{aligned} &= \left[\frac{19-2s+(5s-4)i}{29}\right]\left[\frac{-s+(2-s)i}{(s-1)^2+1}\right] \\ &= \frac{(7s^2-33s+8)+(-3s^2-19s+38)i}{29((s-1)^2+1)}, \end{aligned}$$

which is real if and only if

$$s = \frac{1}{6}\left[-19 \pm \sqrt{817}\right].$$

Hence, there are exactly two real values of s for which $2+3i$, $-2i$, $1-i$, and s lie on a circle in $\overline{\mathbb{C}}$.

2.16: Calculating, we see that

$$[z_1, z_2; z_3, z_4]_2 = \frac{1}{[z_1, z_2; z_3, z_4]}$$

and

$$[z_1, z_2; z_3, z_4]_3 = \frac{1}{1-[z_1, z_2; z_3, z_4]}.$$

2.17: If n fixes a point x of $\overline{\mathbb{C}}$, then $m = p \circ n \circ p^{-1}$ fixes $p(x)$, because

$$m(p(x)) = (p \circ n \circ p^{-1})(p(x)) = p(n(x)) = p(x).$$

As $n = p^{-1} \circ m \circ p$, we see conversely that if m fixes a point y of $\overline{\mathbb{C}}$, then n fixes $p^{-1}(y)$. In particular, m and n have the same number of fixed points.

2.18: As $n_2 \circ n_1^{-1}(0) = 0$ and $n_2 \circ n_1^{-1}(\infty) = \infty$, we can write $n_2 \circ n_1^{-1}(z) = p(z) = cz$ for some $c \in \mathbb{C} - \{0, 1\}$. Hence, $n_2 = p \circ n_1$.

Write $n_k \circ m \circ n_k^{-1}(z) = a_k z$, and note that

$$\begin{aligned} a_2 z = n_2 \circ m \circ n_2^{-1}(z) &= p \circ n_1 \circ m \circ n_1^{-1} \circ p^{-1}(z) \\ &= p \circ (n_1 \circ m \circ n_1^{-1}) \left(\frac{1}{c} z\right) \\ &= p\left(\frac{a_1}{c} z\right) = a_1 z, \end{aligned}$$

and so $a_1 = a_2$, as desired.

2.19: Any Möbius transformation taking x to ∞ and y to 0 can be expressed as $s = J \circ q$, where $J(z) = \frac{1}{z}$ and where q is a Möbius transformation taking x to 0 and y to ∞. Calculating, we see that

$$s \circ m \circ s^{-1}(z) = J \circ (q \circ m \circ q^{-1}) \circ J(z) = \frac{1}{a} z.$$

As by Exercise 2.18 the multiplier of $q \circ m \circ q^{-1}$ is independent of the actual choice of q, subject to the condition that q sends x to 0 and y to ∞, the multiplier of $s \circ m \circ s^{-1}$ is independent of the actual choice of s, subject to the condition that s take x to ∞ and y to 0.

2.20:

1. The fixed points of $m(z) = \frac{2z+5}{3z-1}$ are $z = \frac{1}{6}[3 \pm \sqrt{69}]$. Set

$$q(z) = \frac{z - \frac{1}{6}[3+\sqrt{69}]}{z - \frac{1}{6}[3-\sqrt{69}]},$$

   and calculate that

$$q \circ m \circ q^{-1}(1) = q \circ m(\infty) = q\left(\frac{2}{3}\right) = \frac{\frac{2}{3} - \frac{1}{6}[3+\sqrt{69}]}{\frac{2}{3} - \frac{1}{6}[3-\sqrt{69}]}.$$

   So, m is loxodromic.

2. The fixed points of $m(z) = 7z + 6$ are $z = \infty$ and $z = -1$, and so m is either elliptic or loxodromic. Set $q(z) = z + 1$, and calculate that

$$q \circ m \circ q^{-1}(1) = q \circ m(0) = q(6) = 7.$$

   So, m is loxodromic.

3. $J(z) = \frac{1}{z}$ has fixed points at ± 1, and so it is either elliptic or loxodromic. Instead of conjugating J by a Möbius taking its fixed points to 0 and ∞, we note that $J^2(z) = z$, and so J must be elliptic.

4. The fixed point of $m(z) = \frac{z}{z+1}$ is $z = 0$, and so m is parabolic.

2.21: 1. -34; 2. -1; 3. 2; 4. -4; 5. i; 6. -4.

2.22:

1. $m(z) = \frac{\frac{-2i}{\sqrt{34}}z - \frac{4i}{\sqrt{34}}}{\frac{-5i}{\sqrt{34}}z + \frac{7i}{\sqrt{34}}}$; 2. $m(z) = \frac{i}{iz}$; 3. $m(z) = \frac{\frac{-1}{\sqrt{2}}z - \frac{3}{\sqrt{2}}}{\frac{1}{\sqrt{2}}z + \frac{1}{\sqrt{2}}}$;
4. $m(z) = \frac{\frac{1}{2}z - \frac{i}{2}}{\frac{-i}{2}z + \frac{3}{2}}$; 5. $m(z) = \frac{\frac{i\sqrt{2}}{1+i}z + \frac{\sqrt{2}}{1+i}}{\frac{\sqrt{2}}{1+i}}$; 6. $m(z) = \frac{\frac{i}{2}z}{\frac{-i}{2}z - 2i}$.

2.23: Write $m(z) = \frac{az+b}{cz+d}$ and $n(z) = \frac{\alpha z+\beta}{\gamma z+\delta}$. Then,

$$n \circ m(z) = \frac{(\alpha a + \beta c)z + \alpha b + \beta d}{(\gamma a + \delta c)z + \gamma b + \delta d}$$

and

$$m \circ n(z) = \frac{(a\alpha + b\gamma)z + a\beta + b\delta}{(c\alpha + d\gamma)z + c\beta + d\delta}.$$

Hence,

$$\tau(n \circ m) = (\alpha a + \beta c + \gamma b + \delta d)^2 = \tau(m \circ n),$$

as desired.

2.24: Using Exercise 2.23, we see that

$$\tau((p \circ m) \circ p^{-1}) = \tau(p^{-1} \circ (p \circ m)) = \tau(m).$$

2.25: Calculating, we have that $f'(\rho) = 2\rho - 2\rho^{-3} = 2\rho(1 - \rho^{-4})$, and so $f'(\rho) = 0$ if and only if $\rho = 1$. As

$$\lim_{\rho \to 0^+} f(\rho) = \infty = \lim_{\rho \to \infty} f(\rho),$$

we see that $\rho = 1$ is a global minimum. As $f(1) = 2$, we are done.

2.26:

1. $\tau(m) = \frac{-25}{34}$, and so m is loxodromic with multiplier $\frac{1}{68}\left[-93 - \sqrt{4025}\right]$.

2. $\tau(m) = 0$ and so m is elliptic with multiplier -1.

3. $\tau(m) = 0$, and so m is elliptic with multiplier -1.

4. $\tau(m) = 4$, and so m is parabolic.

5. $\tau(m) = 2$, and so m is elliptic with multiplier i.

6. $\tau(m) = -\frac{9}{4}$, and so m is loxodromic with multiplier -4.

2.27: Instead of calculating, we begin by noting that we can write m as $m(z) = \frac{az+b}{cz+d}$ with $ad - bc = 1$ and $a + d = 2$. Choose p so that $a = 1 + px$ and $d = 1 - px$, and note that this determines p uniquely. As $ad - bc = 1$, we have that $bc = -p^2x^2$.

The fixed points of m satisfy the equation $(1+px)z + b = z(cz + (1-px))$, and so $cz^2 - 2pxz - b = 0$. Completing the square, this becomes $\left(z - \frac{px}{c}\right)^2 = 0$. As $z = x$ is one solution and m is parabolic, we see that $\frac{px}{c} = x$, and so $c = p$. As $bc = -p^2x^2$, this yields that $b = -px^2$, as desired.

2.28: Let $n(z) = az$, and let $p(z)$ be a Möbius transformation taking x to 0 and taking y to ∞. For instance, we may take $p(z) = \frac{z-x}{z-y}$. The determinant of p is $\beta^2 = x - y$, and so normalizing we get that

$$p(z) = \frac{\frac{1}{\beta}z - \frac{x}{\beta}}{\frac{1}{\beta}z - \frac{y}{\beta}}.$$

As m fixes x and y and has multiplier a, we have that $p \circ m \circ p^{-1} = n$, and so $m = p^{-1} \circ n \circ p$. Calculating, we see that

$$p^{-1} \circ n \circ p(z) = \frac{1}{\beta^2} \left[\frac{(x - ay)z + xy(a-1)}{(1-a)z + ax - y} \right] = \frac{\left(\frac{x-ya}{x-y}\right) z + \frac{xy(a-1)}{x-y}}{\left(\frac{1-a}{x-y}\right) z + \frac{xa-y}{x-y}},$$

as desired.

2.29: Obviously, every element $k = \lambda \mathrm{I}$ of K is in $\ker(\mu)$, because $\mu(k)$ is the Möbius transformation $m(z) = \frac{\lambda z}{\lambda} = z$.

Suppose that M is an element of $\mathrm{GL}_2(\mathbb{C})$ so that

$$\mu\left(M = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \right) = \left(m(z) = \frac{az+b}{cz+d} \right)$$

is the identity Möbius transformation. As $m(0) = 0$, we have that $b = 0$; as $m(\infty) = \infty$, we have that $c = 0$. As $m(1) = \frac{a}{d} = 1$, we have that $a = d$, and so

$$M = \begin{pmatrix} a & b \\ c & d \end{pmatrix} = a\,\mathrm{I},$$

as desired.

The conclusion that Möb^+ and $\mathrm{PGL}_2(\mathbb{C})$ are isomorphic follows immediately from the first isomorphism theorem from group theory.

2.30: By definition, $C(z) = \overline{z}$ fixes every point of $\overline{\mathbb{R}}$ and, in particular, fixes 0, 1, and ∞. However, as $C(i) = -i \neq i$, we see that $C(z)$ is not the identity, and so it cannot be an element of Möb^+.

2.31: Let A be the circle in $\overline{\mathbb{C}}$ given by the equation $\alpha z\overline{z} + \beta z + \overline{\beta}\overline{z} + \gamma = 0$. Set $w = C(z) = \overline{z}$, so that $z = \overline{w}$, and note that w then satisfies the equation $\alpha w\overline{w} + \overline{\beta} w + \beta \overline{w} + \gamma = 0$, which is again the equation of a circle in $\overline{\mathbb{C}}$, as desired.

2.32: We check that all possible compositions of pairs again have one of the two desired forms. We already have that the composition of two Möbius transformations is again a Möbius transformation.

We begin by noting that the composition $m \circ C$, where $m(z) = \frac{az+b}{cz+d}$, is

$$(m \circ C)(z) = m(\overline{z}) = \frac{a\overline{z} + b}{c\overline{z} + d}.$$

The composition $m \circ n$, where $n(z) = \frac{\alpha\overline{z}+\beta}{\gamma\overline{z}+\delta}$, is

$$(m \circ n)(z) = \frac{(a\alpha + b\gamma)\overline{z} + a\beta + b\delta}{(c\alpha + d\gamma)\overline{z} + c\beta + d\delta},$$

and so it has the desired form. Similarly, the composition $n \circ m$ has the desired form.

The composition $p \circ n$, where $p(z) = \frac{a\overline{z}+b}{c\overline{z}+d}$, is

$$(p \circ n)(z) = \frac{(a\overline{\alpha} + b\overline{\gamma})z + a\overline{\beta} + b\overline{\delta}}{(c\overline{\alpha} + d\overline{\gamma})z + c\overline{\beta} + d\overline{\delta}},$$

and so it has the desired form.

2.33: One is

$$p(z) = m(z) = \frac{\frac{1}{\sqrt{2}}z + \frac{i}{\sqrt{2}}}{\frac{i}{\sqrt{2}}z + \frac{1}{\sqrt{2}}},$$

for which we have already seen that $p \circ C \circ p^{-1}(z) = \frac{1}{z}$.

Consider also the Möbius transformation n taking $(0, 1, \infty)$ to $(i, 1, -1)$, namely,

$$n(z) = \frac{\frac{1-i}{2}z + i}{\frac{-1+i}{2}z + 1},$$

and so

$$(n \circ C \circ n^{-1})(z) = n\left(\frac{\overline{z} + i}{\frac{1+i}{2}\overline{z} + \frac{1+i}{2}}\right) = \frac{1}{\overline{z}},$$

as desired.

2.34: As $f(z) = az + b$ is the composition of $L(z) = az$ and $P(z) = z + b$, it suffices to check that Proposition 2.20 holds for the transformations L and P.

For P, write $b = \beta e^{i\varphi}$, so that the Euclidean line ℓ passing through 0 and b makes angle φ with $\mathbb{R}$. We express translation along ℓ as the reflection in two lines A and B perpendicular to ℓ, with A passing through 0 and B passing through $\frac{1}{2}b$.

Set $\theta = \varphi - \frac{1}{2}\pi$. Reflection in A is given as

$$C_A(z) = e^{2i\theta}\overline{z} = -e^{2i\varphi}\overline{z},$$

and reflection in B is given as

$$C_B(z) = -e^{2i\varphi}\left(\overline{z} - \frac{1}{2}\overline{b}\right) + \frac{1}{2}b.$$

Calculating,

$$(C_B \circ C_A)(z) = C_B(-e^{2i\varphi}\overline{z}) = -e^{2i\varphi}\left(-e^{-2i\varphi}z - \frac{1}{2}\overline{b}\right) + \frac{1}{2}b = z + b.$$

For L, write $a = \alpha^2 e^{2i\theta}$, and note that L is the composition of $D(z) = \alpha^2 z$ and $E(z) = e^{2i\theta}z$.

We can express D as the composition of the reflection $c(z) = \frac{1}{\overline{z}}$ in $\mathbb{S}^1$ and the reflection $c_2(z) = \frac{\alpha^2}{\overline{z}}$ in the Euclidean circle with Euclidean centre 0 and Euclidean radius α.

We can express E as the composition of the reflection $C(z) = \overline{z}$ in $\mathbb{R}$ and the reflection $C_2(z) = e^{i\theta}\overline{z}$ in the Euclidean line through 0 making angle θ with $\mathbb{R}$.

2.35: We use the notation of the proof of Theorem 2.23. As X_k passes through z_0 and z_k, we have that $C(X_k)$ passes through $C(z_0) = \overline{z_0}$ and $C(z_k) = \overline{z_k}$, and so $C(X_k)$ has slope

$$S_k = \frac{\mathrm{Im}(\overline{z_k} - \overline{z_0})}{\mathrm{Re}(\overline{z_k} - \overline{z_0})} = -\frac{\mathrm{Im}(z_k - z_0)}{\mathrm{Re}(z_k - z_0)} = -s_k.$$

The angle angle$(C(X_1), C(X_2))$ between $C(X_1)$ and $C(X_2)$ is then

$$\begin{aligned} \mathrm{angle}(C(X_1), C(X_2)) &= \arctan(S_2) - \arctan(S_1) \\ &= -\arctan(s_2) + \arctan(s_1) = -\mathrm{angle}(X_1, X_2). \end{aligned}$$

Hence, C is conformal, as it preserves the absolute value of the angle between Euclidean lines.

2.36: In the case in which $a = 0$, the condition that $ad - bc = 1$ yields that $c \neq 0$. Consider the two points $m(1) = \frac{b}{c+d}$ and $m^{-1}(\infty) = -\frac{d}{c}$.

Solving for d and b in terms of c, we get

$$d = -m^{-1}(\infty)c \text{ and } b = m(1)(c+d) = (m(1) - m^{-1}(\infty))c.$$

Hence,

$$1 = ad - bc = (m^{-1}(\infty) - m(1))c^2,$$

and so again b, c, and d are either all real or all purely imaginary.

In the case in which $c = 0$, we have that $a \neq 0$ and $d \neq 0$. In this case, we can write $m(z) = \frac{a}{d}z + \frac{b}{d}$, and so both $m(0) = \frac{b}{d}$ and $m(1) = \frac{a+b}{d}$ are real, which gives

$$b = m(0)d \text{ and } a = (m(1) - m(0))d.$$

Hence,

$$1 = ad - bc = (m(1) - m(0))d^2,$$

and so again a, b, and d are either all real or all purely imaginary.

2.37: Start by taking an element p of Möb taking $\overline{\mathbb{R}}$ to $\mathbb{S}^1$, such as $p(z) = \frac{z-i}{-iz+1}$. Set $m(z) = \frac{az+b}{cz+d}$, and calculate that

$$p \circ m \circ p^{-1}(z) = \frac{(a+d+(b-c)i)z + b + c + (a-d)i}{(b+c-(a-d)i)z + a + d - (b-c)i}.$$

Set $\alpha = a + d + (b - c)i$ and $\beta = b + c + (a - d)i$.

If a, b, c, and d are all real, then with α and β as above, we can rewrite $p \circ m \circ p^{-1}$ as

$$p \circ m \circ p^{-1}(z) = \frac{\alpha z + \beta}{\overline{\beta} z + \overline{\alpha}}.$$

If a, b, c, and d are all purely imaginary, then with α and β as above, we can rewrite $p \circ m \circ p^{-1}$ as

$$p \circ m \circ p^{-1}(z) = \frac{\alpha z + \beta}{-\overline{\beta} z - \overline{\alpha}}.$$

If $n(z) = \frac{a\overline{z}+b}{c\overline{z}+d}$, then

$$p \circ n \circ p^{-1}(z) = \frac{(a - d - (b + c)i)\overline{z} + b - c - (a + d)i}{(-b + c - (a + d)i)\overline{z} - a + d - (b + c)i}.$$

Set $\delta = a - d - (b + c)i$ and $\gamma = b - c - (a + d)i$.

If a, b, c, and d are all real, then with δ and γ as above, we can rewrite $p \circ n \circ p^{-1}$ as

$$p \circ n \circ p^{-1}(z) = \frac{\delta \overline{z} + \gamma}{-\overline{\gamma} z - \overline{\delta}}.$$

If a, b, c, and d are all purely imaginary, then with δ and γ as above, we can rewrite $p \circ n \circ p^{-1}$ as

$$p \circ n \circ p^{-1}(z) = \frac{\delta \overline{z} + \gamma}{\overline{\gamma} z + \overline{\delta}}.$$

2.38: This is similar to the proof of Theorem 2.4. First, note that the elements listed as generators are all elements of Möb($\mathbb{H}$). Consider the element $m(z) = \frac{az+b}{cz+d}$ of Möb($\mathbb{H}$), where a, b, c, $d \in \mathbb{R}$ and $ad - bc = 1$.

If $c = 0$, then $m(z) = \frac{a}{d} z + \frac{b}{d}$. As $1 = ad - bc = ad$, we have that $\frac{a}{d} = a^2 > 0$.

If $c \neq 0$, then $m(z) = f(K(g(z)))$, where $g(z) = c^2 z + cd$ and $f(z) = z + \frac{a}{c}$.

For $n(z) = \frac{a\overline{z}+b}{c\overline{z}+d}$, where a, b, c, and d are purely imaginary and $ad - bc = 1$, note that $B \circ n = m$, where $m(z) = \frac{\alpha z+\beta}{\gamma z+\delta}$ is an element of Möb($\mathbb{H}$). Hence, we can write $n = B^{-1} \circ m = B \circ m$.

2.39: We know from Theorem 2.26 that every element of Möb($\mathbb{H}$) either has the form $m(z) = \frac{az+b}{cz+d}$ where a, b, c, $d \in \mathbb{R}$ and $ad - bc = 1$ or has the form $n(z) = \frac{a\overline{z}+b}{c\overline{z}+d}$ where a, b, c, and d are purely imaginary and $ad - bc = 1$.

The Möbius transformation $p(z) = \frac{z-i}{-iz+1}$ takes $\overline{\mathbb{R}}$ to $\mathbb{S}^1$ and takes $\mathbb{H}$ to $\mathbb{D}$ because $p(i) = 0$.

For m, we calculate

$$p \circ m \circ p^{-1}(z) = \frac{(a+d+(b-c)i)z+b+c+(a-d)i}{(b+c-(a-d)i)z+a+d-(b-c)i} = \frac{\alpha z+\beta}{\overline{\beta}z+\overline{\alpha}},$$

where $\alpha = a+d+(b-c)i$ and $\beta = b+c+(a-d)i$.

For n, we calculate

$$p \circ n \circ p^{-1}(z) = \frac{(a-d-(b+c)i)\overline{z}+b-c-(a+d)i}{(-b+c-(a+d)i)\overline{z}-a+d-(b+c)i} = \frac{\delta\overline{z}+\gamma}{\overline{\gamma z}+\overline{\delta}},$$

with $\delta = a-d-(b+c)i$ and $\gamma = b-c-(a+d)i$.

2.40: Let ℓ be a hyperbolic line in $\mathbb{H}$. Using Lemma 2.8, it suffices to construct an element of Möb($\mathbb{H}$) that takes ℓ to the positive imaginary axis I in $\mathbb{H}$. One approach is to construct an element of Möb($\mathbb{H}$) taking the endpoints at infinity of ℓ to 0 and ∞, as is done in the solution to Exercise 2.41. We take another approach here.

Choose a point z on ℓ. By Proposition 2.28, there exists an element m of Möb($\mathbb{H}$) with $m(z) = i$. Let φ be the angle between the two hyperbolic lines I and $m(\ell)$, measured from I to $m(\ell)$.

For each θ, the Möbius transformation

$$n_\theta(z) = \frac{\cos(\theta)z - \sin(\theta)}{\sin(\theta)z + \cos(\theta)}$$

lies in Möb($\mathbb{H}$) and fixes i. Also, the angle between I and $n_\theta(I)$ at i, measured from I to $n_\theta(I)$, is 2θ.

So, if we take $\theta = \frac{1}{2}\varphi$, we have that $n_\theta(I)$ and $m(\ell)$ are both hyperbolic lines through i that make angle φ with I. Hence, $m(\ell) = n_\theta(I)$, and so $I = n_\theta^{-1} \circ m(\ell)$.

2.41: For any two points $y < x$ in $\mathbb{R}$, the Möbius transformation $m(z) = \frac{z-x}{z-y}$ satisfies $m(x) = 0$ and $m(y) = \infty$. Also, the determinant of m is $x - y > 0$, and so m lies in Möb$^+$($\mathbb{H}$).

For $y = -2$ and $x = 1$, we get $m(z) = \frac{z-1}{z+2}$ as an element of Möb($\mathbb{H}$) taking ℓ to I.

2.42: Let I be the positive imaginary axis, and let r_I be the closed hyperbolic ray consisting of the portion of I above and including i. Using Lemma 2.8, it suffices to show that there exists an element m of Möb$^+$($\mathbb{H}$) taking any element (ℓ, r, z_0) in X to (I, r_I, i).

Start with an element n of Möb$^+$($\mathbb{H}$) taking the endpoints at infinity of ℓ to 0 and ∞, the endpoints at infinity of I. (If the endpoints at infinity of ℓ, call them p and q, lie in $\mathbb{R}$ with $p < q$, then $n(z) = \frac{z-q}{z-p}$. We leave the case of one of p or q being ∞ for the reader.) We then have that $n(z_0) = \lambda i$ for some $\lambda > 0$. Let $s(z) = \frac{1}{\lambda} z$. Then, $s \circ n$ satisfies $s \circ n(\ell) = I$ and $s \circ n(z_0) = i$. If $s \circ n(r) = r_I$, then set $m = s \circ n$. If $s \circ n(r) \neq r_I$, then set $m = K \circ s \circ n$, where $K(z) = -\frac{1}{z}$. As all of n, s, K are elements of $\text{Möb}^+(\mathbb{H})$, we have that $m \in \text{Möb}^+(\mathbb{H})$.

2.43: Let H be the open half-plane $H = \{z \in \mathbb{H} \mid \text{Re}(z) > 0\}$ determined by the positive imaginary axis I. Yet again using Lemma 2.8, given any open half-plane L in $\mathbb{H}$, it suffices to construct an element of $\text{Möb}(\mathbb{H})$ taking L to H.

Let ℓ be the bounding line for L. By Exercise 2.40, there is an element m of $\text{Möb}(\mathbb{H})$ satisfying $m(\ell) = I$. In particular, m takes the two open half-planes determined by ℓ to the two open half-planes determined by I.

If $m(L) = H$, we are done. If $m(L) \neq H$, then $B \circ m(L) = H$, where $B(z) = -\overline{z}$ is reflection in I, and we are done.

2.44: As the hyperbolic line ℓ determined by $\sqrt{2}$ and $-\sqrt{2}$ lies in the Euclidean circle with Euclidean centre 0 and Euclidean radius $\sqrt{2}$, reflection in ℓ is given by $r(z) = \frac{2}{\overline{z}}$.

Composing, we see that

$$r \circ q(z) = \frac{2z+2}{z+2} = \frac{\sqrt{2}z + \sqrt{2}}{\frac{1}{\sqrt{2}}z + \sqrt{2}} = m(z),$$

which is loxodromic fixing $\sqrt{2}$ and $-\sqrt{2}$.

As $r(z) = r^{-1}(z)$, we then have that $q(z) = r \circ m(z)$, as desired.

2.45: Setting $q(z) = z$, we get $z = -\overline{z} + 1$, which we can rewrite as $\text{Re}(z) = \frac{1}{2}$. Hence, the fixed points of q are exactly the points on the hyperbolic line in $\mathbb{H}$ contained in the Euclidean line $\{\text{Re}(z) = \frac{1}{2}\}$.

2.46: The fixed points in $\overline{\mathbb{C}}$ of

$$q(z) = \frac{2i\overline{z} - i}{3i\overline{z} - 2i}$$

are the solutions in $\overline{\mathbb{C}}$ of $q(z) = z$, which are the points z in $\overline{\mathbb{C}}$ satisfying

$$3i|z|^2 - 2i(z + \overline{z}) + i = 0.$$

Writing $z = x + iy$, we see that the fixed points of q in $\overline{\mathbb{C}}$ are the points on the Euclidean circle

$$x^2 - \frac{4}{3}x + \frac{1}{3} + y^2 = 0,$$

which is the Euclidean circle

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

Hence, the fixed points of q are exactly the points on the hyperbolic line in $\mathbb{H}$ contained in the Euclidean circle with Euclidean centre $\frac{2}{3}$ and Euclidean radius $\frac{1}{3}$.

2.47: There are two cases. Suppose that n is reflection in the line $\{x = a\}$. From our work in Section 2.6, we can write $n(z) = -\overline{z} + 2a$. Calculating, we see that

$$p \circ n(z) = p(-\overline{z} + 2a) = -\overline{z} + 1 - 2a.$$

In fact, $p \circ n$ fixes every point z in $\mathbb{H}$ with $2\,\mathrm{Re}(z) = 1 - 2a$, and so $p \circ n$ is reflection in the hyperbolic line $\{x = \frac{1}{2}(1 - 2a)\}$ in $\mathbb{H}$.

Suppose now that n is reflection in the line contained in the Euclidean circle with Euclidean centre c and Euclidean radius r. By our work in Section 2.6, we can write $n(z) = c + \frac{r^2}{\overline{z} - c}$. Calculating, we see that

$$p \circ n(z) = c + 1 + \frac{r^2}{\overline{z} - c} = \frac{(c+1)\overline{z} + r^2 - c(c+1)}{\overline{z} - c}.$$

Setting $p \circ n(z) = z$, we see that the fixed points of $p \circ n$ are those points in $\overline{\mathbb{C}}$ satisfying

$$x^2 - (2c+1)x + y^2 + c(c+1) - r^2 + iy = 0.$$

Setting imaginary parts equal, we see that $y = 0$. In particular, we see that $p \circ n$ has no fixed points in $\mathbb{H}$, and the fixed points in $\overline{\mathbb{R}}$ are the two solutions to

$$\left(x - \frac{1}{2}(2c+1))\right)^2 = r^2 + \frac{1}{4}.$$

Solutions to Chapter 3 exercises:

3.1: As the Euclidean distance from z to $\mathbb{S}^1$ is $1 - |z|$, we see that $\delta(z) = \frac{1}{1-|z|}$. Parametrize C_r by the path $f : [0, 2\pi] \to \mathbb{D}$ given by $f(t) = re^{it}$, so that $|f(t)| = r$ and $|f'(t)| = |ire^{it}| = r$. Calculating, we see that

$$\begin{aligned} \mathrm{length}(C_r) = \mathrm{length}(f) &= \int_f \frac{1}{1-|z|}|\mathrm{d}z| \\ &= \int_0^{2\pi} \frac{1}{1-|f(t)|}|f'(t)|\mathrm{d}t \\ &= \int_0^{2\pi} \frac{r}{1-r}\mathrm{d}t = \frac{2\pi r}{1-r}. \end{aligned}$$

3.2: On $[0,1]$, we have that $|f(t)| = |t+ti| = \sqrt{2}\,t$ and $|f'(t)| = \sqrt{2}$, whereas on $[-1,0]$, we have that $|f(t)| = |t-ti| = \sqrt{2t^2} = -\sqrt{2}\,t$ and $|f'(t)| = \sqrt{2}$. So,

$$\begin{aligned}\text{length}(f) = \int_f \frac{1}{1+|z|^2}|\mathrm{d}z| &= \int_{-1}^{1} \frac{1}{1+2t^2}\,\sqrt{2}\mathrm{d}t \\ &= 2\arctan(\sqrt{2}).\end{aligned}$$

3.3: Parametrize A_λ by the path $f : [-1,1] \to \mathbb{H}$ given by $f(t) = t + i\lambda$. As $\text{Im}(f(t)) = \lambda$ and $|f'(t)| = 1$, we see that

$$\text{length}(f) = \int_{-1}^{1} \frac{c}{\lambda}\mathrm{d}t = \frac{2c}{\lambda}.$$

B_λ lies on the Euclidean circle with Euclidean centre 0 and Euclidean radius $\sqrt{1+\lambda^2}$. The Euclidean line segment between 0 and $1+i\lambda$ makes angle θ with the positive real axis, where $\cos(\theta) = \frac{1}{\sqrt{1+\lambda^2}}$. So, we can parametrize B_λ by the path $g : [\theta, \pi-\theta] \to \mathbb{H}$ given by $g(t) = \sqrt{1+\lambda^2}\,e^{i\theta}$.

As $\text{Im}(g(t)) = \sqrt{1+\lambda^2}\,\sin(\theta)$ and $|g'(t)| = \sqrt{1+\lambda^2}$, we see that

$$\text{length}(g) = \int_\theta^{\pi-\theta} c\csc(t)\,\mathrm{d}t = c\,\ln\left[\frac{\sqrt{1+\lambda^2}+1}{\sqrt{1+\lambda^2}-1}\right].$$

3.4: As $K'(z) = \frac{1}{z^2}$, the condition imposed on $\rho(z)$ is that

$$0 = \mu_K(z) = \rho(z) - \rho(K(z))|K'(z)| = \rho(z) - \rho\left(-\frac{1}{z}\right)\frac{1}{|z|^2}.$$

Substituting $\rho(z) = \frac{c}{\text{Im}(z)}$ and using that

$$\rho\left(-\frac{1}{z}\right) = \rho\left(\frac{-\overline{z}}{|z|^2}\right) = \frac{c|z|^2}{\text{Im}(-\overline{z})} = \frac{c|z|^2}{\text{Im}(z)},$$

we obtain

$$\rho(z) - \rho\left(-\frac{1}{z}\right)\frac{1}{|z|^2} = \frac{c}{\text{Im}(z)} - \frac{c|z|^2}{\text{Im}(z)}\frac{1}{|z|^2} = \frac{c}{\text{Im}(z)} - \frac{c}{\text{Im}(z)} = 0,$$

as desired.

For a piecewise C^1 path $f : [a,b] \to \mathbb{H}$ given by $f(t) = x(t) + iy(t)$, we have that $B \circ f(t) = -x(t) + iy(t)$.

Hence, we have that $|(B\circ f)'(t)| = |f'(t)|$ and $\text{Im}(B\circ f)(t) = y(t) = \text{Im}(f(t))$, and so

$$\begin{aligned}\text{length}(B\circ f) &= \int_a^b \frac{c}{\text{Im}((B\circ f)(t))}|(B\circ f)'(t)|\mathrm{d}t \\ &= \int_a^b \frac{c}{\text{Im}(f(t))}|f'(t)|\mathrm{d}t = \text{length}(f),\end{aligned}$$

as desired.

3.5: As $|f_n'(t)| = |1 + int^{n-1}| = \sqrt{1 + n^2 t^{2n-2}}$ and as $\mathrm{Im}(f_n(t)) = t^n + 1$, we have that

$$\mathrm{length}_{\mathbb{H}}(f_n) = \int_{f_n} \frac{1}{\mathrm{Im}(z)} |\mathrm{d}z| = \int_0^1 \frac{\sqrt{1 + n^2 t^{2n-2}}}{1 + t^n} \mathrm{d}t.$$

For $n = 1$, this gives that

$$\mathrm{length}_{\mathbb{H}}(f_1) = \int_0^1 \frac{\sqrt{2}}{1+t} \mathrm{d}t = \sqrt{2}\, \ln(2).$$

For $n \geq 2$, this integral is more difficult to evaluate explicitly.

3.6: As $n \to \infty$, the curves $\gamma_n = f_n([0,1])$ seem to converge to the curve γ that is the union of the horizontal Euclidean line segment ℓ_1 joining i and $1+i$ and the vertical Euclidean line segment ℓ_2 joining $1+i$ and $1+2i$.

Consequently, we might expect that $\mathrm{length}_{\mathbb{H}}(\gamma_n) \to \mathrm{length}_{\mathbb{H}}(\gamma)$ as $n \to \infty$.

Parametrizing ℓ_1 by $f : [0,1] \to \mathbb{H}$ given by $f(t) = t + i$, we see that

$$\mathrm{length}_{\mathbb{H}}(\ell_1) = \mathrm{length}_{\mathbb{H}}(f) = 1.$$

Parametrizing ℓ_2 by $g : [1,2] \to \mathbb{H}$ given by $g(t) = 1 + ti$, we see that

$$\mathrm{length}_{\mathbb{H}}(\ell_2) = \mathrm{length}_{\mathbb{H}}(g) = \ln(2).$$

Hence, we expect that $\mathrm{length}_{\mathbb{H}}(\gamma) = 1 + \ln(2)$.

3.7: Let $G = \{p_b(z) = z + b \mid b \in \mathbb{C}\}$ be the subgroup of $\mathrm{M\ddot{o}b}^+$ generated by all parabolic Möbius transformations fixing ∞; note that every element of G is either parabolic or is the identity. The element of arc-length $\lambda(z)|\mathrm{d}z|$ on $\mathbb{C}$ is invariant under G, so that

$$\int_f \lambda(z)|\mathrm{d}z| = \int_{p\circ f} \lambda(z)|\mathrm{d}z| = \int_f \lambda(p(z))|p'(z)||\mathrm{d}z|$$

for every $p \in G$ and every piecewise C^1 path $f : [a,b] \to \mathbb{C}$, and so by Lemma 3.10, we see that

$$\lambda(z) = \lambda(p(z))|p'(z)|$$

for every $z \in \mathbb{C}$ and every $p \in G$. If $p(z) = p_b(z) = z + b$ for some $b \in \mathbb{C}$, we have that $\lambda(p(z)) = \lambda(z + b)$ and $|p'(z)| = 1$. Therefore, if $\lambda(z)|\mathrm{d}z|$ is invariant under G, then $\lambda(z + b) = \lambda(z)$ for every $z \in \mathbb{C}$ and every $b \in \mathbb{C}$. Setting $z = 0$, we see that $\lambda(0) = \lambda(b)$ for every $b \in \mathbb{C}$, and so $\lambda(z)$ is constant.

3.8: Let $H = \{m \in \text{Möb}^+ \mid m(\infty) = \infty\}$ be the subgroup of Möb^+ consisting of all Möbius transformations fixing ∞. Note that H contains G. As invariance of $\lambda(z)|\mathrm{d}z|$ under H implies invariance under G, we see by Exercise 3.7 that $\lambda(z)$ is constant. We also have that $\lambda(z)|\mathrm{d}z|$ is invariant under the subgroup $\{m_a(z) = az \mid a \in \mathbb{C}, a \neq 0\}$ of H, and so

$$\lambda(z)|\mathrm{d}z| = \lambda(m_a(z))|m_a'(z)||\mathrm{d}z| = |a|\lambda(az)|\mathrm{d}z|$$

for all $z \in \mathbb{C}$ and all $a \in \mathbb{C}$, $a \neq 0$. Again, by Lemma 3.10, we have that $\lambda(z) = a\lambda(az)$ for all $z \in \mathbb{C}$ and all $a \in \mathbb{C}$, $a \neq 0$. Setting $z = 1$, we see that $\lambda(1) = |a|\lambda(a)$. Hence, $\frac{1}{|a|}\lambda(1) = \lambda(a)$ for all $a \in \mathbb{C}$, $a \neq 0$. As $\lambda(z)$ is constant, we let $a \to \infty$ to see that $\lambda(a) = 0$.

3.9: As $g(-1) = g(1) = 2i$, and as g achieves its minimum at $t = 0$, the image of $[-1, 1]$ under g is the hyperbolic line segment joining i to $2i$, covered twice.

The hyperbolic length of g is

$$\text{length}_{\mathbb{H}}(g) = \int_g \frac{1}{\text{Im}(z)}|\mathrm{d}z| = \int_{-1}^{1} \frac{|2t|}{t^2+1}\mathrm{d}t = 2\ln(2).$$

3.10: For each $n \geq 2$, define the numbers

$$2 = \lambda_0 < \lambda_1 < \ldots < \lambda_n = 10$$

by setting

$$\mathrm{d}_{\mathbb{H}}(\lambda_k i, \lambda_{k+1} i) = \frac{1}{n}\mathrm{d}_{\mathbb{H}}(2i, 10i)$$

for $0 \leq k \leq n-1$. As

$$\mathrm{d}_{\mathbb{H}}(\lambda_k i, \lambda_{k+1} i) = \ln\left[\frac{\lambda_{k+1}}{\lambda_k}\right],$$

we see that

$$\mathrm{d}_{\mathbb{H}}(\lambda_0 i, \lambda_k i) = \ln\left[\frac{\lambda_k}{\lambda_0}\right] = \frac{k}{n}\mathrm{d}_{\mathbb{H}}(2i, 10i).$$

Hence,

$$\ln(\lambda_k) = \frac{k}{n}\ln(5) + \ln(2),$$

and so

$$\lambda_k = 2 \cdot 5^{\frac{k}{n}} i.$$

For example, for $n = 2$, we get that the midpoint of the hyperbolic line segment between $2i$ and $10i$ is $2\sqrt{5}i$.

3.11: By Exercise 1.3, the Eulidean centre of the Euclidean circle containing the hyperbolic line ℓ passing through z_1 and z_2 is

$$c = \frac{1}{2}\left[\frac{|z_1|^2 - |z_2|^2}{\mathrm{Re}(z_1) - \mathrm{Re}(z_2)}\right] = \frac{1}{2}\left[\frac{|z_1|^2 - |z_2|^2}{x_1 - x_2}\right].$$

Setting the Euclidean radius of the Euclidean circle to be $r = |z_1 - c|$, the endpoints at infinity of ℓ are $c - r$ and $c + r$.

Set $m(z) = \frac{z-(c+r)}{z-(c-r)}$. As the determinant of m is $c + r - (c - r) = 2r > 0$, we have that m lies in $\text{Möb}^+(\mathbb{H})$. Calculating, we see that

$$m(z_1) = \frac{z_1 - (c+r)}{z_1 - (c-r)} \text{ and } m(z_2) = \frac{z_2 - (c+r)}{z_2 - (c-r)},$$

both of which lie on the positive imaginary axis by construction.

Hence,

$$\begin{aligned} \mathrm{d}_{\mathbb{H}}(z_1, z_2) = \mathrm{d}_{\mathbb{H}}(m(z_1), m(z_2)) &= \left|\ln\left[\frac{m(z_2)}{m(z_1)}\right]\right| \\ &= \left|\ln\left[\frac{(z_2 - (c+r))}{(z_2 - (c-r))}\frac{(z_1 - (c-r))}{(z_1 - (c+r))}\right]\right|. \end{aligned}$$

3.12: We use the notation and formula from Exercise 3.11. For $A = i$ and $B = 1 + 2i$, we have $c = 2$ and $r = \sqrt{5}$, and so

$$\mathrm{d}_{\mathbb{H}}(A, B) = \ln\left[\frac{\sqrt{5}+1}{\sqrt{5}-1}\right].$$

For $A = i$ and $C = -1 + 2i$, we have $c = -2$ and $r = \sqrt{5}$, and so

$$\mathrm{d}_{\mathbb{H}}(A, C) = \ln\left[\frac{\sqrt{5}+1}{\sqrt{5}-1}\right].$$

Note that we expect this, because $A = K(A)$ and $C = K(B)$, where $K(z) = -\overline{z}$ is an element of $\text{Möb}(\mathbb{H})$ and hence preserves hyperbolic distance.

For $A = i$ and $D = 7i$, we have that $\mathrm{d}_{\mathbb{H}}(A, D) = \ln(7)$.

For $B = 1 + 2i$ and $C = -1 + 2i$, we have $c = 0$ and $r = \sqrt{5}$, and so

$$\mathrm{d}_{\mathbb{H}}(B, C) = \ln\left[\frac{\sqrt{5}+1}{\sqrt{5}-1}\right].$$

For $B = 1 + 2i$ and $D = 7i$, we have $c = -22$ and $r = \sqrt{533}$, and so

$$\mathrm{d}_{\mathbb{H}}(B, D) = \ln\left[\frac{41+\sqrt{533}}{41-\sqrt{533}}\right].$$

As $C = K(B)$ and $D = K(D)$ for $K(z) = -\overline{z}$, we have that

$$\begin{aligned} \mathrm{d}_{\mathbb{H}}(C, D) &= \mathrm{d}_{\mathbb{H}}(K(B), K(D)) \\ &= \mathrm{d}_{\mathbb{H}}(B, D) \\ &= \ln\left[\frac{41 + \sqrt{533}}{41 - \sqrt{533}}\right]. \end{aligned}$$

3.13: If there exists an element q of Möb($\mathbb{H}$) taking (z_1, z_2) to (w_1, w_2), then

$$\mathrm{d}_{\mathbb{H}}(w_1, w_2) = \mathrm{d}_{\mathbb{H}}(q(z_1), q(z_2)) = \mathrm{d}_{\mathbb{H}}(z_1, z_2).$$

If, on the other hand, we have that $\mathrm{d}_{\mathbb{H}}(w_1, w_2) = \mathrm{d}_{\mathbb{H}}(z_1, z_2)$, we proceed as follows. There exists an element m of Möb($\mathbb{H}$) taking z_1 to i and taking z_2 to $e^{\mathrm{d}_{\mathbb{H}}(z_1,z_2)}i$; there also exists an element n of Möb($\mathbb{H}$) taking w_1 to i and taking w_2 to $e^{\mathrm{d}_{\mathbb{H}}(w_1,w_2)}i$. Note that $m(z_1) = n(w_1) = i$. As $\mathrm{d}_{\mathbb{H}}(w_1, w_2) = \mathrm{d}_{\mathbb{H}}(z_1, z_2)$, we have that $m(z_2) = n(w_2)$, and so $q = n^{-1} \circ m$ takes (z_1, z_2) to (w_1, w_2).

3.14: If $f(x) = f(y)$, then $\mathrm{d}(f(x), f(y)) = 0$. Hence, $\mathrm{d}(x, y) = 0$, and so $x = y$ by the first of the three conditions describing a metric. Hence, f is injective.

To show that f is continuous at x, take some $\varepsilon > 0$. We need to find $\delta > 0$ so that $f(U_\delta(x)) \subset U_\varepsilon(f(z))$. However, because $\mathrm{d}(x, y) = \mathrm{d}(f(x), f(y))$, we see that if $y \in U_\delta(x)$, then $\mathrm{d}(x, y) < \delta$, and so $\mathrm{d}(f(x), f(y)) < \delta$, and so $f(y) \in U_\delta(f(x))$. Hence, take $\delta = \varepsilon$.

3.15: We know from our work in Section 2.1 that $f(z) = az$ is a homeomorphism of $\mathbb{C}$ for every $a \in \mathbb{C} - \{0\}$. As

$$|f(z) - f(w)| = |az - aw| = |a|\,|z - w|,$$

we see that f is an isometry if and only if $|a| = 1$.

3.16: This follows from Proposition 3.20. Suppose that y lies in the hyperbolic line segment ℓ_{xz} joining x to z. Then,

$$\mathrm{d}_{\mathbb{H}}(x, y) + \mathrm{d}_{\mathbb{H}}(y, z) = \mathrm{d}_{\mathbb{H}}(x, z).$$

As f is a hyperbolic isometry, it preserves hyperbolic distance, and so

$$\mathrm{d}_{\mathbb{H}}(f(x), f(y)) + \mathrm{d}_{\mathbb{H}}(f(y), f(z)) = \mathrm{d}_{\mathbb{H}}(f(x), f(z)).$$

In particular, $f(y)$ lies in the hyperbolic line segment $\ell_{f(x)\,f(z)}$ joining $f(x)$ to $f(z)$, and so

$$f\left(\ell_{xz}\right) = \ell_{f(x)\,f(z)}.$$

As a hyperbolic line can be expressed as a nested union of hyperbolic line segments, we have that hyperbolic isometries take hyperbolic lines to hyperbolic lines.

3.17: Without loss of generality, we suppose that $\lambda < \mu$. Write $y = \xi i$, and consider $\mathrm{d}_{\mathbb{H}}(x, y) = |\ln(\lambda) - \ln(\xi)|$. As a function of ξ, $g(\xi) = \ln(\lambda) - \ln(\xi)$ is strictly decreasing, because $g'(\xi) = -\frac{1}{\xi}$. Hence, for any number c, there is at most one solution to $g(\xi) = c$.

Hence, for any $c > 0$, there are two solutions y to $\mathrm{d}_{\mathbb{H}}(x, y) = c$. One is $y = e^{\ln(\lambda)-c}i$, and the other is $y = e^{\ln(\lambda)+c}i$. Geometrically, one is above $x = \lambda i$ on I and the other is below.

Similarly, there are two solutions y to $\mathrm{d}_{\mathbb{H}}(y, z) = c'$, one above z on I and one below. Hence, there can be only one solution to the two equations $\mathrm{d}_{\mathbb{H}}(x, y) = c$ and $\mathrm{d}_{\mathbb{H}}(y, z) = c'$.

3.18: If X and Y do not have disjoint closures, then there exists a point x in $\overline{X} \cap \overline{Y}$. As $x \in \overline{X}$, there exists a sequence $\{x_n\}$ of points of X converging to X, and as $x \in \overline{Y}$, there exists a sequence $\{y_n\}$ of points of Y converging to Y. As $\mathrm{d}_{\mathbb{H}}$ is continuous, we have that

$$\lim_{n\to\infty} \mathrm{d}_{\mathbb{H}}(x_n, y_n) = \mathrm{d}_{\mathbb{H}}(x, x) = 0.$$

As $\mathrm{d}_{\mathbb{H}}(X, Y) \leq \mathrm{d}_{\mathbb{H}}(x_n, y_n)$ for all n, we have shown that $\mathrm{d}_{\mathbb{H}}(X, Y) = 0$.

Suppose now that $\mathrm{d}_{\mathbb{H}}(X, Y) = 0$. As

$$\mathrm{d}_{\mathbb{H}}(X, Y) = \inf\{\mathrm{d}_{\mathbb{H}}(x, y) \mid x \in X,\ y \in Y\},$$

there exists a sequence of points $\{x_n\}$ of X and a sequence of points $\{y_n\}$ of Y so that $\lim_{n\to\infty} \mathrm{d}_{\mathbb{H}}(x_n, y_n) = 0$.

As X is compact, there exists a subsequence $\{x_{n_k}\}$ of $\{x_n\}$ that converges to a point x of X. As $\lim_{n\to\infty} \mathrm{d}_{\mathbb{H}}(x_n, y_n) = 0$, we have that $\lim_{k\to\infty} \mathrm{d}_{\mathbb{H}}(x_{n_k}, y_{n_k}) = 0$, and so $\{y_{n_k}\}$ converges to x as well. Hence, x is a point of $\overline{X} \cap \overline{Y}$, and so X and Y do not have disjoint closures.

3.19: Begin by applying an element of Möb($\mathbb{H}$) so that ℓ lies in the Euclidean circle with Euclidean centre 0 and Euclidean radius 1 and so that $p = \lambda i$ for some $\lambda > 1$. In this case, the unique hyperbolic line through p that is perpendicular to ℓ is the positive imaginary axis I, which intersects ℓ at i.

Using the formula for $\mathrm{d}_{\mathbb{H}}(z_1, z_2)$ given in Section 3.5, and a lot of algebraic massage, we derive that the hyperbolic distance

$$\mathrm{d}_{\mathbb{H}}(e^{i\theta}, \lambda i) = \ln\left[\cosh(\ln(\lambda))\csc(\theta) + \sqrt{\cosh^2(\ln(\lambda))\csc^2(\theta) - 1}\right].$$

The derivative of this function is negative for $0 < \theta \leq \frac{\pi}{2}$, and so $\mathrm{d}_{\mathbb{H}}(e^{i\theta}, \lambda i)$ achieves its unique minumum at $\theta = \frac{\pi}{2}$.

In particular, note that this result shows the following. Let ℓ be a hyperbolic line, let p be a point in $\mathbb{H}$ not on ℓ, and let a be the point on ℓ satisfying $\mathrm{d}_{\mathbb{H}}(p,a) = \mathrm{d}_{\mathbb{H}}(p,\ell)$. Then, for a point z in ℓ, the hyperbolic distance $\mathrm{d}_{\mathbb{H}}(p,z)$ is monotonically increasing as a function of $\mathrm{d}_{\mathbb{H}}(a,z)$.

3.20: The distance from the point $\rho e^{i\varphi}$ to the positive imaginary axis I is equal to the hyperbolic length of the hyperbolic line segment from $\rho e^{i\varphi}$ to I that meets I perpendicularly, which is the hyperbolic line segment joining $\rho e^{i\varphi}$ and ρi.

To calculate $\mathrm{d}_{\mathbb{H}}(\rho e^{i\varphi}, \rho i)$, we may use, for instance, Exercise 3.11. The hyperbolic line passing through $\rho e^{i\varphi}$ and ρi lies on the Euclidean circle with Euclidean centre 0 and Euclidean radius ρ, and so

$$\mathrm{d}_{\mathbb{H}}(\rho e^{i\varphi}, \rho i) = \left| \ln \left[\frac{\sin(\varphi)}{1+\cos(\varphi)} \right] \right|.$$

On $(0, \frac{\pi}{2}]$, we have $1+\cos(\varphi) \geq 1$ and $\sin(\varphi) \leq 1$, and so

$$\mathrm{d}_{\mathbb{H}}(\rho e^{i\varphi}, \rho i) = \ln \left[\frac{1+\cos(\varphi)}{\sin(\varphi)} \right].$$

Hence, W_ε is the set of points of $\mathbb{H}$ for which $\mathrm{d}_{\mathbb{H}}(\rho e^{i\varphi}, \rho i) = \varepsilon$ is constant, and so φ is constant. This is a Euclidean ray from the origin, where we take $\varphi = \frac{\pi}{2} - \theta$.

As $\mathrm{d}_{\mathbb{H}}(\rho e^{i\varphi}, \rho i) = \mathrm{d}_{\mathbb{H}}(\rho e^{i(\pi-\varphi)}, \rho i)$, we see that W_ε has a second component, namely, the Euclidean ray from the origin making angle $\frac{\pi}{2} + \theta = \pi - \varphi$ with the positive real axis.

3.21: Use the triple transitivity of the action of Möb($\mathbb{H}$) on $\overline{\mathbb{R}}$ to assume that that the endpoints at infinity of ℓ_0 are 0 and ∞, and that the endpoints at infinity of ℓ_1 are 1 and ∞.

For each $r > 1$, let c_r be the hyperbolic line contained in the Euclidean circle with Euclidean centre 0 and Euclidean radius r, and note that c_r is the unique hyperbolic line through ri that is perpendicular to ℓ_0. Hence, if there is a hyperbolic line perpendicular to both ℓ_0 and ℓ_1, then it will be one of the c_r.

Write the point of intersection of c_r and ℓ_1 as $re^{i\theta}$, and note that $\cos(\theta) = \frac{\sqrt{r^2-1}}{r}$ is nonzero for all $r > 1$. However, θ is also the angle between ℓ_1 and c_r, measured from ℓ_1 to c_r, and so no c_r intersects ℓ_1 perpendicularly.

3.22: By the ordering of the points around $\overline{\mathbb{R}}$, there exists an element of Möb($\mathbb{H}$) taking z_0 to 0, taking z_1 to ∞, taking w_0 to 1, and taking w_1 to $x > 1$.

From our work in Section 2.3, we know that

$$[z_0, w_0; w_1, z_1] = [0, 1; x, \infty] = \frac{x-1}{0-1} = 1 - x,$$

and so

$$1 - [z_0, w_0; w_1, z_1] = x.$$

Calculating, we see that

$$\tanh^2\left[\frac{1}{2}\mathrm{d}_{\mathbb{H}}(\ell_0, \ell_1)\right] = \tanh^2\left[\frac{1}{2}\ln\left[\frac{\sqrt{x}+1}{\sqrt{x}-1}\right]\right] = \frac{1}{x},$$

as desired.

3.23: We calculate this proportion by determining the proportion of hyperbolic lines passing through p that intersect ℓ in terms of the angle of their tangent lines at p.

Apply an element of Möb($\mathbb{H}$) so that ℓ lies in the Euclidean circle with Euclidean centre 0 and Euclidean radius 1 and so that $p = \lambda i$ for some $\lambda > 1$. Let ℓ_0 be the hyperbolic ray from p to 1, and let φ be the angle between ℓ_0 and the positive imaginary axis I.

We can calculate φ as follows. Note that ℓ_0 lies on the Euclidean circle A with Euclidean centre $c = \frac{1}{2}(1 - \lambda^2)$ and Euclidean radius $r = |c - 1|$, and so the equation of A in $\mathbb{C}$ is $(x - c)^2 + y^2 = r^2$. Differentiating implicitly with respect to x, we get that

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{-x + c}{y}.$$

Hence, at $\lambda i = (0, \lambda)$, we have that the slope of the tangent line to A at $(0, \lambda)$ is

$$\frac{\mathrm{d}y}{\mathrm{d}x}(0, \lambda) = \frac{c}{\lambda} = \frac{1 - \lambda^2}{2\lambda}.$$

Hence,

$$\varphi = \arctan\left(\frac{\lambda^2 - 1}{2\lambda}\right) - \frac{\pi}{2},$$

and so the proportion of hyperbolic rays from $p = \lambda i$ intersecting ℓ is

$$\frac{2\varphi}{2\pi} = \frac{1}{\pi}\arctan\left(\frac{\lambda^2 - 1}{2\lambda}\right) - \frac{\pi}{2}.$$

Solutions to Chapter 4 exercises:

4.1: As m^{-1} takes $\mathbb{D}$ to $\mathbb{H}$, we have that

$$\mathrm{length}_{\mathbb{D}}(m \circ f) = \mathrm{length}_{\mathbb{H}}(m^{-1} \circ m \circ f) = \mathrm{length}_{\mathbb{H}}(f),$$

as desired.

4.2: Parametrize the hyperbolic line segment between 0 and r by the path $f : [0, r] \to \mathbb{D}$ given by $f(t) = t$. As the image of f is the hyperbolic line segment in $\mathbb{D}$ joining 0 and r, we have that $\mathrm{d}_{\mathbb{D}}(0, r) = \mathrm{length}_{\mathbb{D}}(f)$. We have already calculated that

$$\mathrm{d}_{\mathbb{D}}(0, r) = \mathrm{length}_{\mathbb{D}}(f) = \ln\left[\frac{1+r}{1-r}\right].$$

Solving for r as a function of $\mathrm{d}_{\mathbb{D}}(0, r)$, we get that

$$r = \tanh\left[\frac{1}{2}\mathrm{d}_{\mathbb{D}}(0, r)\right].$$

4.3: Apply an element of $\mathrm{Möb}(\mathbb{D})$ so that ℓ_1 lies in the real axis and so that the point of intersection of ℓ_1 and ℓ_2 is 0.

The endpoints at infinity of ℓ_1 are then $z_1 = 1$ and $z_2 = -1$, and the endpoints at infinity of ℓ_2 are $w_1 = e^{i\theta}$ and $w_2 = -e^{i\theta}$, where θ is the angle between ℓ_1 and ℓ_2.

Hence,

$$[z_1, w_1; z_2, w_2] = [1, e^{i\theta}; -1, -e^{i\theta}] = \frac{-\sin^2(\theta)}{(1-\cos(\theta))^2} = -\cot^2\left(\frac{\theta}{2}\right),$$

and so

$$[z_1, w_1; z_2, w_2]\tan^2\left(\frac{\theta}{2}\right) = -1,$$

as desired.

4.4: As the hyperbolic radius of S_s is s, the Euclidean radius of S_s is $r = \tanh(\frac{1}{2}s)$, by Exercise 4.2. Parametrize S_s by $f : [0, 2\pi] \to \mathbb{D}$, where $f(t) = r\exp(it)$. Then,

$$\mathrm{length}(f) = \int_f \frac{2}{1-|z|^2}\,|\mathrm{d}z| = \int_0^{2\pi} \frac{2}{1-r^2}\,r\,\mathrm{d}t = \frac{4\pi r}{1-r^2} = 2\pi\sinh(s).$$

4.5: Note that if $0 < r < s < 1$, then

$$\mathrm{d}_{\mathbb{D}}(re^{i\theta}, se^{i\theta}) = \mathrm{d}_{\mathbb{D}}(0, se^{i\theta}) - \mathrm{d}_{\mathbb{D}}(0, re^{i\theta}) = \ln\left(\frac{(1+s)(1-r)}{(1-s)(1+r)}\right).$$

Writing the Euclidean centre of A as $c = \frac{1}{5} - \frac{1}{4}i = \frac{\sqrt{41}}{20}e^{i\theta}$, and noting that $|c| > \frac{1}{10}$, we see that A intersects the hyperbolic ray $R = \{w \in \mathbb{D} | \arg(w) = \theta\}$ in two points. As reflection in R takes A to itself and is a hyperbolic isometry, we see that the hyperbolic centre of A also lies on R. The two points of intersection of A and R are $\frac{\sqrt{41}-2}{20}e^{i\theta}$ and $\frac{\sqrt{41}+2}{20}e^{i\theta}$. The hyperbolic distance between these two points is

$$\begin{aligned}
&\mathrm{d}_{\mathbb{D}}\left(\frac{\sqrt{41}-2}{20}e^{i\theta}, \frac{\sqrt{41}+2}{20}e^{i\theta}\right) \\
&\quad = \mathrm{d}_{\mathbb{D}}\left(0, \frac{\sqrt{41}+2}{20}e^{i\theta}\right) - \mathrm{d}_{\mathbb{D}}\left(0, \frac{\sqrt{41}-2}{20}e^{i\theta}\right) \\
&\quad = \ln\left(\frac{\left(1+\frac{\sqrt{41}+2}{20}\right)}{\left(1-\frac{\sqrt{41}+2}{20}\right)}\frac{\left(1-\frac{\sqrt{41}-2}{20}\right)}{\left(1+\frac{\sqrt{41}-2}{20}\right)}\right) \\
&\quad = \ln\left(\frac{22+\sqrt{41}}{18-\sqrt{41}}\,\frac{22-\sqrt{41}}{18+\sqrt{41}}\right) \\
&\quad = \ln\left(\frac{443}{283}\right) = Z.
\end{aligned}$$

Hence, the hyperbolic radius of A is $\frac{1}{2}Z$.

The hyperbolic centre of A is the point $\alpha e^{i\theta}$ for which $\frac{\sqrt{41}-2}{20} < \alpha < \frac{\sqrt{41}+2}{20}$ and $\mathrm{d}_{\mathbb{D}}\left(\frac{\sqrt{41}-2}{20}e^{i\theta}, \alpha e^{i\theta}\right) = \mathrm{d}_{\mathbb{D}}\left(\alpha e^{i\theta}, \frac{\sqrt{41}+2}{20}e^{i\theta}\right) = \frac{1}{2}Z$.

As

$$\mathrm{d}_{\mathbb{D}}\left(\frac{\sqrt{41}-2}{20}e^{i\theta}, \alpha e^{i\theta}\right) = \ln\left(\frac{(1+\alpha)(22-\sqrt{41})}{(1-\alpha)(18+\sqrt{41})}\right)$$

and

$$\mathrm{d}_{\mathbb{D}}\left(\frac{\sqrt{41}+2}{20}e^{i\theta}, \alpha e^{i\theta}\right) = \ln\left(\frac{(22+\sqrt{41})(1-\alpha)}{(18-\sqrt{41})(1+\alpha)}\right),$$

we can set

$$\ln\left(\frac{(1+\alpha)(22-\sqrt{41})}{(1-\alpha)(18+\sqrt{41})}\right) = \ln\left(\frac{(22+\sqrt{41})(1-\alpha)}{(18-\sqrt{41})(1+\alpha)}\right) = K$$

and solve for α. Hence, the hyperbolic center is $\alpha e^{i\theta}$.

4.6: As vertical Euclidean lines in $\mathbb{H}$ are both Euclidean lines and hyperbolic lines, and as reflection in the line $\ell = \{\mathrm{Re}(z) = 1\}$ takes A to itself, the hyperbolic centre of A lies on ℓ. As $A \cap \ell = \{1+2i, 1+4i\}$, the hyperbolic radius of A is

$$\frac{1}{2}\mathrm{d}_{\mathbb{H}}(1+2i, 1+4i) = \frac{1}{2}\mathrm{d}_{\mathbb{H}}(2i, 4i) = \frac{1}{2}\ln(2).$$

The hyperbolic centre is the point $1+si$, where

$$\mathrm{d}_{\mathbb{H}}(1+2i, 1+si) = \mathrm{d}_{\mathbb{H}}(1+si, 1+4i),$$

and so s satisfies $\ln\left(\frac{s}{2}\right) = \ln\left(\frac{4}{s}\right)$. Hence, $s^2 = 8$, and so $s = 2\sqrt{2}$.

4.7: We proceed as in the solution to Exercise 4.6. The hyperbolic radius is

$$R = \frac{1}{2}\mathrm{d}_{\mathbb{H}}(a + (b + r)i, a + (b - r)i) = \frac{1}{2}\ln\left(\frac{b+r}{b-r}\right).$$

Calculating, we see that $\tanh(R) = \frac{r}{b}$, and so $b\tanh(R) = r$.

The hyperbolic centre is the point $a + si$ where

$$\mathrm{d}_{\mathbb{H}}(a + si, a + (b + r)i) = \mathrm{d}_{\mathbb{H}}(a + si, a + (b - r)i).$$

Hence,

$$\ln\left(\frac{b+r}{s}\right) = \ln\left(\frac{s}{b-r}\right),$$

and so $s^2 = b^2 - r^2$. That is, the hyperbolic centre is $a + i\sqrt{b^2 - r^2}$.

4.8: As $\mathrm{Im}(\xi(z)) = \mathrm{Im}(iz) = \mathrm{Re}(z)$ and $|\xi'(z)| = |i| = 1$, we see that

$$\mathrm{ds}_{\mathrm{X}} = \frac{1}{\mathrm{Im}(z))}|\xi'(z)||\mathrm{d}z| = \frac{1}{\mathrm{Re}(z)}|\mathrm{d}z|.$$

4.9: First, note that the image of the horizontal Euclidean line $\{z \in \mathbb{C} \,|\, \mathrm{Im}(z) = c\}$ in X under ξ is the Euclidean ray $\{z \in \mathbb{H} \mid \arg(w) = c\}$, and so ξ is indeed a holomorphic homeomorphism between X and $\mathbb{H}$.

Let $z = x + iy$. Then, the pullback of $\frac{1}{\mathrm{Im}(z)}|\mathrm{d}z|$ by ξ is

$$\mathrm{ds}_X = \frac{1}{\mathrm{Im}(\xi(z))}|\xi'(z)||\mathrm{d}z| = \frac{1}{e^x\sin(y)}|e^z||\mathrm{d}z| = \frac{1}{\sin(y)}|\mathrm{d}z|.$$

4.10: We mimic the pullback argument for holomorphic functions $\varphi : X \to \mathbf{H}$. Consider the function $\varphi_{a,r} : D_{a,r} \to \mathbb{D}$ given by $\varphi_{a,r}(z) = \frac{1}{r}(z - a)$. Calculating the hyperbolic length of a path $f : [0, 1] \to D_{a,r}$, we see that

$$\begin{aligned}
\int_f \lambda_{a,r}(z)|\mathrm{d}z| &= \int_{\varphi_{a,r}\circ f} \frac{2}{1 - |z|^2}|\mathrm{d}z| \\
&= \int_0^1 \frac{2}{1 - |\varphi_{a,r}\circ f(t)|^2}|(\varphi_{a,r}\circ f)'(t)|\,\mathrm{d}t \\
&= \int_0^1 \frac{2}{1 - |\varphi_{a,r}(f(t))|^2}|\varphi'_{a,r}(f(t))|\,|f'(t)|\,\mathrm{d}t \\
&= \int_f \frac{2}{1 - |\varphi_{a,r}(z)|^2}|\varphi'_{a,r}(z)|\,|\mathrm{d}z|,
\end{aligned}$$

and so

$$\lambda_{a,r}(z) = \frac{2}{1 - |\varphi_{a,r}(z)|^2}|\varphi'_{a,r}(z)| = \frac{2r}{r^2 - |z - a|^2}.$$

Note that $\lambda_{a,r}(z) \geq \lambda(z)$ if and only if

$$\frac{2r}{r^2 - |z-a|^2} \geq \frac{2}{1-|z|^2}$$

if and only if

$$r(1-|z|^2) \geq r^2 - |z-a|^2,$$

which (by multiplying out, collecting terms, and refactoring) occurs if and only if

$$\left| z - \frac{a}{1-r} \right|^2 + \frac{r((1-r)^2 - |a|^2)}{(1-r)^2} = \left| z - \frac{a}{1-r} \right|^2 + \frac{r(1-r-|a|)(1-r+|a|)}{(1-r)^2} \geq 0.$$

As $0 < r < 1 - |a|$, we see that $1 - r - |a| > 0$ and that $1 - r + |a| > 0$. As the above equation is always satisfied (and the left-hand side is equal to 0 if and only if $z = a$ and $r = 0$), we see that $\lambda_{a,r}(z) \geq \lambda(z)$ for all $z \in D_{a,r}$, as desired.

4.11: Using the same argument as for holomorphic homeomorphisms, the pullback of $\frac{2}{1-|z|^2}|\mathrm{d}z|$ by $f(z) = z^2$ is

$$\frac{2|f'(z)|}{1-|f(z)|^2}|\mathrm{d}z| = \frac{4|z|}{1-|z|^4}|\mathrm{d}z|.$$

Then,

$$\frac{4|z|}{1-|z|^4} \leq \frac{2}{1-|z|^2}$$

if and only if

$$2|z|(1-|z|^2) \leq 1 - |z|^4,$$

which occurs if and only if

$$0 \leq 1-|z|^4-2|z|(1-|z|^2) = (1+|z|^2)(1-|z|^2)-2|z|(1-|z|^2) = (1-|z|)^2(1-|z|^2),$$

and this inequality always holds true for $|z| < 1$.

4.13: Calculating, we see that the curvature is identically -1.

4.14: Calculating, we see that the curvature at $z \in \mathbb{C}$ is identically 16.

Solutions to Chapter 5 exercises:

5.1: Let z_0 and z_1 be two points of $X = \cap_{\alpha \in A} X_\alpha$, and let $\ell_{z_0 z_1}$ be the hyperbolic line segment joining z_0 to z_1.

As each X_α is convex, we have that $\ell_{z_0 z_1}$ is contained in each X_α, and so $\ell_{z_0 z_1}$ is then contained in their intersection $X = \cap_{\alpha \in A} X_\alpha$.

5.2: The Euclidean radius of D_s is $r = \tanh(\frac{1}{2}s)$. For each θ, let ℓ_θ be the hyperbolic line contained in the Euclidean circle with Euclidean centre on the

Euclidean line $\{te^{i\theta} | t > 0\}$ and passing through $re^{i\theta}$. Let H_θ be the closed half-plane determined by ℓ_θ containing 0. As we may express D_s as the intersection $D_s = \cap_\theta H_\theta$ and as each H_θ is convex, we see that D_s is convex.

As any open hyperbolic disc can be taken by an element of Möb($\mathbb{D}$) to some D_s for some $s > 0$ and as Möb($\mathbb{D}$) preserves convexity, we see that all open hyperbolic discs are convex.

Repeating this argument with a closed hyperbolic disc and open half-planes H_θ, we see that all closed hyperbolic discs are convex as well.

5.3: By definition, $X \subset \text{conv}(X)$. Conversely, because X is a convex set in the hyperbolic plane containing X, we have that $\text{conv}(X)$ is the intersection of X and other sets, and so $\text{conv}(X) \subset X$. Hence, $X = \text{conv}(X)$.

5.4: Let ℓ_1 and ℓ_2 be the two hyperbolic lines, and let the endpoints at infinity of ℓ_k be x_k and y_k. Set $Z = \{x_1, y_1, x_2, y_2\}$. In the case in which x_1, y_1, x_2, y_2 are distinct points, the convex hull $\text{conv}(\ell_1 \cup \ell_2)$ of the union $\ell_1 \cup \ell_2$ is equal to the convex hull $\text{conv}(Z)$ of Z. This is the region in the hyperbolic plane bounded by four of the six hyperbolic lines determined by these four points.

Note, in the degenerate case, that ℓ_1 and ℓ_2 share an endpoint at infinity, the convex hull $\text{conv}(\ell_1 \cup \ell_2)$ is the region bounded by the three hyperbolic lines determined by the three points in Z.

5.5: Let ℓ_{xy} be the closed hyperbolic line segment joining x to y, and let ℓ be the hyperbolic line containing ℓ_{xy}. We can express ℓ as the intersection $\ell = \cap_{\alpha \in A} H_\alpha$ of a collection $\{H_\alpha\}_{\alpha \in A}$ of (two) closed half-planes.

Now let ℓ_x be any hyperbolic line passing through x other than ℓ, and let H_x be the closed half-plane determined by ℓ_x that contains ℓ_{xy}. Similarly, take ℓ_y to be any hyperbolic line other than ℓ passing through y and let H_y be the closed half-plane determined by ℓ_y that contains ℓ_{xy}. Then, we may express ℓ_{xy} as the intersection

$$\ell_{xy} = H_x \cap H_y \cap \ell = H_x \cap H_y \cap (\cap_{\alpha \in A} H_\alpha)$$

of a collection of closed half-planes.

As each closed half-plane can be expressed as the intersection of a collection of open half-planes, we can also express ℓ_{xy} as the intersection of a collection of open half-planes.

Now, let ℓ_{xz} be the hyperbolic ray determined by $x \in \mathbb{H}$ and $z \in \overline{\mathbb{R}}$. Let ℓ and ℓ_x be as defined above, and note that

$$\ell_{xz} = H_x \cap \ell = H_x \cap (\cap_{\alpha \in A} H_\alpha).$$

Again, because each closed half-plane can be expressed as the intersection of a collection of open half-planes, we can express ℓ_{xz} as the intersection of a collection of open half-planes.

5.6: Let $\mathcal{H} = \{H_\alpha\}_{\alpha\in A}$ be an uncountable collection of half-planes. Let ℓ_α be the bounding line for H_α. We work in the upper half-plane $\mathbb{H}$ for the sake of concreteness. Let $\mathbb{Q}^+ = \mathbb{Q} \cap (0, \infty)$ denote the set of positive rational numbers.

For each $q \in \mathbb{Q}^+$, consider the hyperbolic disc $U_q(i)$ with hyperbolic centre i and hyperbolic radius q. As the union $\cup_{q\in\mathbb{Q}^+} U_q(i)$ is equal to $\mathbb{H}$, there is some $q \in \mathbb{Q}^+$ so that $U = U_q(i)$ intersects infinitely many bounding lines ℓ_α.

In particular, there is a sequence $\{\ell_{\alpha_n}\}$ of bounding lines, each of which intersects U. For each n, choose a point $x_n \in U \cap \ell_{\alpha_n}$. As the closure $\overline{U}$ of U is closed and bounded, it is compact, and so there exists a subsequence of $\{x_{\alpha_n}\}$, which we again call $\{x_{\alpha_n}\}$ to avoid proliferation of subscripts, so that $\{x_{\alpha_n}\}$ converges to some point x of U.

By the definition of convergence, for each $\varepsilon > 0$, the hyperbolic disc $U_\varepsilon(x)$ contains infinitely many x_{α_n}. Hence, for each $\varepsilon > 0$, the hyperbolic disc $U_\varepsilon(x)$ intersects infinitely many bounding lines ℓ_α, and so $\{H_\alpha\}_{\alpha\in A}$ is not locally finite.

5.7: Let P be a hyperbolic polygon, and suppose that P contains three points x, y, and z that do not lie on the same hyperbolic line. Given two points p and q in the hyperbolic plane, let ℓ_{pq} be the closed hyperbolic line segment joining them. Then, the set

$$X = \cup\{\ell_{zp} \mid p \in \ell_{xy}\}$$

has nonempty interior.

In fact, let p be the midpoint of ℓ_{xy} and let q be the midpoint of ℓ_{zp}. Then, the three numbers $\mathrm{d}_{\mathbb{H}}(q, \ell_{xy})$, $\mathrm{d}_{\mathbb{H}}(q, \ell_{xz})$, and $\mathrm{d}_{\mathbb{H}}(q, \ell_{yz})$ are all positive. If we set

$$\varepsilon = \min\{\mathrm{d}_{\mathbb{H}}(q, \ell_{xy}),\ \mathrm{d}_{\mathbb{H}}(q, \ell_{xz}),\ \mathrm{d}_{\mathbb{H}}(q, \ell_{yz})\},$$

then $U_\varepsilon(q)$ is contained in X.

Hence, the only degenerate hyperbolic polygons are closed convex subsets of hyperbolic lines, which are exactly the hyperbolic lines, closed hyperbolic rays, closed hyperbolic line segments, and points.

5.8: As P has only finitely many sides, and as each side contains exactly two vertices, because it is a closed hyperbolic line segment, we see that P has exactly as many vertices as it has sides. Let $v_1, \ldots, v_n$ be the vertices of P, and let $V = \{v_1, \ldots, v_n\}$.

By definition, P is a convex set containing V, and so $\text{conv}(V) \subset P$.

Conversely, note that because $\text{conv}(V)$ contains the vertices of P, we have that $\text{conv}(V)$ contains all sides of P, because each side of P is the closed hyperbolic line segment joining two of the vertices. That is, we have just shown that ∂P is contained in $\text{conv}(V)$.

Now let x be any point in the interior of P, and let ℓ be any hyperbolic line through x. The intersection of P with ℓ is a hyperbolic line segment ℓ_0 in ℓ whose endpoints are in ∂P.

Hence, because $\text{conv}(V)$ is convex and contains the endpoints of ℓ_0, we have that $\text{conv}(V)$ contains ℓ_0. In particular, x is a point of $\text{conv}(V)$, and so $P \subset \text{conv}(V)$. Hence, $\text{conv}(V) = P$.

5.9: For notational ease, let ℓ_{jk} be the hyperbolic line passing through x_j and x_k. Note that, no matter the value of s, the hyperbolic lines ℓ_{12} and ℓ_{34} are parallel, because they are contained in parallel Euclidean lines.

The hyperbolic line ℓ_{13} is contained in the Euclidean circle C_{13} with Euclidean centre 0 and Euclidean radius $\sqrt{2}$, whereas the hyperbolic line ℓ_{24} is contained in the Euclidean circle C_{24} with Euclidean centre $\frac{1}{4}s^2 - 1$ and Euclidean radius $\frac{1}{4}\sqrt{s^4+64}$. Note that C_{13} and C_{24} intersect at a point of $\mathbb{R}$ exactly when $\pm\sqrt{2}$ lies on C_{24}.

Calculating, we have that $-\sqrt{2}$ lies on C_{24} precisely when

$$\left|\left(\frac{1}{4}s^2 - 1\right) + \sqrt{2}\right| = \frac{1}{4}\sqrt{s^4+64},$$

namely,

$$s = \sqrt{10 + 6\sqrt{2}}.$$

Similarly, we have that $\sqrt{2}$ lies on C_{24} precisely when

$$\left|\left(\frac{1}{4}s^2 - 1\right) - \sqrt{2}\right| = \frac{1}{4}\sqrt{s^4+64},$$

namely,

$$s = \sqrt{10 - 6\sqrt{2}}.$$

Hence, Q_s is a hyperbolic parallelogram if and only if

$$\sqrt{10 - 6\sqrt{2}} \le s \le \sqrt{10 + 6\sqrt{2}}.$$

5.10: This proof is the same as the proof of Exercise 5.8.

5.11: We first note that every hyperbolic triangle is contained in an ideal hyperbolic triangle. To see this, let T be a hyperbolic triangle with vertices v_1, v_2, and v_3, and let x be any point in the interior of T.

Let y_k to be the endpoint at infinity of the hyperbolic ray determined by x and v_k, and let P be the ideal triangle with ideal vertices y_1, y_2, and y_3. Then, P contains T. Hence, it suffices to work in the case in which T is an ideal triangle.

To make the calculation easier, let m be an element of $\text{Möb}(\mathbb{H})$ taking T to the ideal triangle with ideal vertices at $y_1 = 0$, $y_2 = 1$, and $y_3 = \infty$. Let ℓ_{jk} be the hyperbolic line determined by y_j and y_k.

For each $r > 0$, let C_r be the Euclidean circle with Euclidean centre 0 and Euclidean radius r. Note that C_r intersects ℓ_{12} at the point $re^{i\theta}$, where $\cos(\theta) = r$. (This relationship between r and θ is obtained by noticing that $re^{i\theta}$ also lies in the Euclidean circle $\left(x - \frac{1}{2}\right)^2 + y^2 = \frac{1}{4}$, which contains the hyperbolic line ℓ_{12}.) For each point ri on ℓ_{13}, the hyperbolic distance between ri and ℓ_{12} is equal to the hyperbolic distance between ri and $re^{i\theta}$

By the solution to Exercise 3.20, the hyperbolic distance between ri and $re^{i\theta}$ is

$$\mathrm{d}_{\mathbb{H}}(re^{i\theta}, ri) = \ln\left[\frac{1 + \cos(\theta)}{\sin(\theta)}\right].$$

By symmetry, we need only consider θ in the range $[\frac{\pi}{4}, \frac{\pi}{2}]$. On $[\frac{\pi}{4}, \frac{\pi}{2}]$, the function $\ln\left[\frac{1+\cos(\theta)}{\sin(\theta)}\right]$ is decreasing, and so $\mathrm{d}_{\mathbb{H}}(re^{i\theta}, ri)$ is maximized at $\theta = \frac{\pi}{4}$. Hence,

$$\mathrm{d}_{\mathbb{H}}(x, \ell_{12}) \leq \ln\left[\frac{1 + \frac{1}{\sqrt{2}}}{\frac{1}{\sqrt{2}}}\right] = \ln(\sqrt{2} + 1).$$

5.12: Let z_0 be the hyperbolic centre of A, let H be the group of all elliptic Möbius transformations in $\text{Möb}^+(\mathbb{D})$ fixing z_0, and note that $h(A) = A$ for every element $h \in H$. Moreover, note that H acts transitively on $\mathbb{S}^1$. (Both of these statements can easily be seen to be true by taking A with its hyperbolic centre at 0; in which case, H becomes the group of all rotations fixing 0.)

Also, note that for every $h \in H$, $h(T)$ is again a hyperbolic ideal triangle circumscribing $h(A) = A$. (If ℓ is a hyperbolic line that is a side of T, then because T circumscribes A, ℓ is tangent to A at some point ξ, and then $h(\ell)$ is a hyperbolic line tangent to $h(A) = A$ at $h(\xi)$.) Let w be one of the vertices of T. As H acts transitively on $\mathbb{S}^1$, if we are given any $z \in \mathbb{S}^1$, then there exists some $h_z \in H$ for which $h_z(w) = z$, and so $h_z(T)$ is then a hyperbolic ideal triangle circumscribing A with one vertex at z, as desired.

5.13: The hyperbolic area of X_s is

$$\mathrm{area}_{\mathbb{H}}(X_s) = \int_{X_s} \frac{1}{y^2}\, \mathrm{d}x\, \mathrm{d}y = \int_{-1}^{1} \int_{s}^{\infty} \frac{1}{y^2}\, \mathrm{d}y\, \mathrm{d}x = \frac{2}{s}.$$

5.14: Rewriting in terms of x and y, we see that $B(x,y) = (-x,y)$. Hence,

$$DB(x,y) = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix},$$

and so

$$\det(DB(x,y)) = -1.$$

As

$$h \circ B(x,y) = \frac{1}{y^2},$$

the change of variables theorem yields that

$$\mathrm{area}_{\mathbb{H}}(B(X)) = \int_{B(X)} \frac{1}{y^2}\, \mathrm{d}x\, \mathrm{d}y = \int_{X} \frac{1}{y^2}\, \mathrm{d}x\, \mathrm{d}y = \mathrm{area}_{\mathbb{H}}(X),$$

as desired.

5.15: Rewriting in terms of x and y, we see that $f(x,y) = (x+y,y)$. Hence,

$$Df(x,y) = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix},$$

and so

$$\det(Df(x,y)) = 1.$$

As

$$h \circ f(x,y) = \frac{1}{y^2},$$

the change of variables theorem yields that

$$\mathrm{area}_{\mathbb{H}}(f(X)) = \int_{f(X)} \frac{1}{y^2}\, \mathrm{d}x\, \mathrm{d}y = \int_{X} \frac{1}{y^2}\, \mathrm{d}x\, \mathrm{d}y = \mathrm{area}_{\mathbb{H}}(X).$$

This completes the proof that f preserves hyperbolic area.

To see that f is not an element of $\mathrm{Möb}(\mathbb{H})$, note that f takes the Euclidean line $\{\mathrm{Re}(z) = 1\}$, which contains a hyperbolic line, to the Euclidean line $\{\mathrm{Re}(z) = 1 + \mathrm{Im}(z)\}$, which does not intersect $\mathbb{R}$ perpendicularly and so does not contain a hyperbolic line.

5.16: Using the change of variables theorem and the appropriate reformulation of Lemma 3.10, it suffices to consider the equation

$$\frac{1}{y^2} = \frac{1}{g(x,y)^2} \left| \frac{\partial g}{\partial y} \right|.$$

First note that because the left-hand side of this equation is never 0 in $\mathbb{H}$ (because $y > 0$ for points (x, y) in $\mathbb{H}$) and that $g(x, y)$ is never 0 (for the same reason), we have that $\left|\frac{\partial g}{\partial y}\right|$ is never 0. Hence, there are two cases to consider.

First, suppose that $\left|\frac{\partial g}{\partial y}\right| > 0$ on $\mathbb{H}$. The equation above then becomes

$$\frac{1}{y^2} = \frac{1}{g(x,y)^2}\frac{\partial g}{\partial y},$$

and integrating with respect to y, we see that

$$-\frac{1}{y} = -\frac{1}{g(x,y)} + f(x)$$

where $f(x)$ is any function of x. Solving for $g(x, y)$, we see that

$$g(x,y) = \frac{y}{1 + yf(x)}.$$

We now see what conditions are imposed on $f(x)$: As both $y > 0$ and $g(x, y) > 0$, we see that $1 + yf(x) > 0$ for all $(x, y) \in \mathbb{H}$, and so $\frac{1}{y} > -f(x)$ for all $(x, y) \in \mathbb{H}$. That is, $f(x) > -\frac{1}{y}$ for all $x \in \mathbb{R}$ and all $y > 0$, and so $f(x) \geq 0$ for all $x \in \mathbb{R}$. As a particular example, $f(x) = x^2$ works.

Now suppose that $\left|\frac{\partial g}{\partial y}\right| < 0$ on $\mathbb{H}$. The equation above then becomes

$$\frac{1}{y^2} = -\frac{1}{g(x,y)^2}\frac{\partial g}{\partial y},$$

and integrating with respect to y, we see that

$$-\frac{1}{y} = \frac{1}{g(x,y)} + f(x)$$

where $f(x)$ is any function of x. Solving for $g(x, y)$, we see that

$$g(x,y) = -\frac{y}{1 + yf(x)}.$$

We now see what conditions are imposed on $f(x)$: As both $y > 0$ and $g(x, y) > 0$, we see that $1 + yf(x) < 0$ for all $(x, y) \in \mathbb{H}$, and so $\frac{1}{y} < -f(x)$ for all $(x, y) \in \mathbb{H}$. That is, $f(x) < -\frac{1}{y}$ for all $x \in \mathbb{R}$ and all $y > 0$, and so $f(x) \leq 0$ for all $x \in \mathbb{R}$. As a particular example, $f(x) = -x^2$ works.

Geometrically, this function preserves each vertical Euclidean (or hyperbolic) line in $\mathbb{H}$ and shifts vertically along each such line by $g(x, y)$.

5.17: The hyperbolic radius s of D_s is related to the Euclidean radius R by $R = \tanh(\frac{1}{2}s)$. The hyperbolic area of D_s is then

$$\text{area}_{\mathbb{D}}(D_s) \quad = \quad \int_{D_s} \frac{4r}{(1-r^2)^2}\,\mathrm{d}r\,\mathrm{d}\theta$$

$$\begin{aligned} &= \int_0^R \int_0^{2\pi} \frac{4r}{(1-r^2)^2}\,\mathrm{d}r\,\mathrm{d}\theta \\ &= 2\pi \int_0^R \frac{4r}{(1-r^2)^2}\,\mathrm{d}r\,\mathrm{d}\theta = \frac{4\pi R^2}{1-R^2} = 4\pi\sinh^2\left(\frac{1}{2}s\right). \end{aligned}$$

5.18: As $\text{length}_{\mathbb{D}}(S_s) = 2\pi\sinh(s)$ and as $\text{area}_{\mathbb{D}}(D_s) = 4\pi\sinh^2\left(\frac{1}{2}s\right)$, we have that

$$q_{\mathbb{D}}(s) = \frac{\text{length}_{\mathbb{D}}(S_s)}{\text{area}_{\mathbb{D}}(D_s)} = \frac{2\pi\sinh(s)}{4\pi\sinh^2\left(\frac{1}{2}s\right)} = \coth\left(\frac{1}{2}s\right).$$

In particular, note that $q_{\mathbb{D}}(s) \to 1$ as $s \to \infty$ and $q_{\mathbb{D}}(s) \to \infty$ as $s \to 0$.

The corresponding Euclidean quantity $q_{\mathbb{C}}(r) = \frac{2}{r}$ behaves much differently as the radius of the Euclidean circle and Euclidean disc get large. Namely, $q_{\mathbb{C}}(r) \to 0$ as $r \to \infty$, whereas we again have that $q_{\mathbb{C}}(r) \to \infty$ as $r \to 0$.

5.19: We begin with the fact that if C_1 and C_2 are intersecting Euclidean circles, where C_k has Euclidean centre c_k and Euclidean radius r_k, then using the law of cosines, the angle θ between C_1 and C_2 satisfies

$$|c_1 - c_2|^2 = r_1^2 + r_2^2 - 2r_1r_2\cos(\theta),$$

and so

$$\cos(\theta) = \frac{r_1^2 + r_2^2 - |c_1 - c_2|^2}{2r_1r_2}.$$

Set $v_1 = i$, $v_2 = 2 + 2i$, and $v_3 = 4 + i$. Let s_{jk} be the side of P joining v_j to v_k, let ℓ_{jk} be the hyperbolic line containing s_{jk}, and let C_{jk} be the Euclidean circle containing ℓ_{jk}. Calculating, we see that C_{12} has Euclidean centre $\frac{7}{4}$ and Euclidean radius $\frac{\sqrt{65}}{4}$, that C_{23} has Euclidean centre $\frac{9}{4}$ and Euclidean radius $\frac{\sqrt{65}}{4}$, and that C_{13} has Euclidean centre 2 and Euclidean radius $\sqrt{5}$.

The angle α between C_{12} and C_{13} is given by

$$\cos(\alpha) = \frac{\frac{65}{16} + 5 - |\frac{7}{4} - 2|^2}{2\frac{\sqrt{65}}{4}\sqrt{5}} = \frac{18}{\sqrt{325}},$$

namely,

$$\alpha \sim 0.0555.$$

The angle β between C_{23} and C_{13} is given by

$$\cos(\beta) = \frac{\frac{65}{16} + 5 - |\frac{9}{4} - 2|^2}{2\frac{\sqrt{65}}{4}\sqrt{5}} = \frac{18}{\sqrt{325}},$$

namely,

$$\beta \sim 0.0555.$$

The angle γ between C_{12} and C_{23} is given by

$$\cos(\gamma) = \frac{\frac{65}{16} + \frac{65}{16} - |\frac{7}{4} - \frac{9}{4}|^2}{2\frac{\sqrt{65}}{4}\frac{\sqrt{65}}{4}} = \frac{126}{130},$$

namely,

$$\gamma \sim 0.2487.$$

Hence, we see by Theorem 5.16 that

$$\mathrm{area}_{\mathbb{H}}(P) = \pi - (\alpha + \beta + \gamma) \sim 2.7819.$$

5.20: Let C_0 be the Euclidean circle in $\mathbb{C}$ containing the hyperbolic line ℓ_0 passing through $rp_0 = r$ and $rp_1 = r\exp\left(\frac{2\pi i}{n}\right)$. The Euclidean centre of C_0 is then of the form $s\exp\left(\frac{\pi i}{n}\right)$ for some $s > 1$. As C_0 must intersect $\mathbb{S}^1$ perpendicularly, we have from Exercise 1.2 that the Euclidean radius of C_0 is $\sqrt{s^2-1}$.

For C_0 to pass through r, we must have that

$$\left| s\exp\left(\frac{\pi i}{n}\right) - r \right| = \sqrt{s^2-1},$$

and so

$$s = \frac{r^2+1}{2r\cos\left(\frac{\pi}{n}\right)}.$$

In particular, the Euclidean centre of C_0 is

$$s\exp\left(\frac{\pi i}{n}\right) = \frac{r^2+1}{2r\cos\left(\frac{\pi}{n}\right)}\exp\left(\frac{\pi i}{n}\right),$$

and the Euclidean radius of C_0 is

$$\sqrt{s^2-1} = \sqrt{\frac{(r^2+1)^2}{4r^2\cos^2\left(\frac{\pi}{n}\right)} - 1}.$$

We can repeat this calculation for the Euclidean circle C_{n-1} containing the hyperbolic line ℓ_{n-1} passing through $rp_0 = r$ and $rp_{n-1} = r\exp\left(\frac{2\pi(n-1)i}{n}\right)$. The Euclidean centre of C_{n-1} is

$$s\exp\left(\frac{-\pi i}{n}\right) = \frac{r^2+1}{2r\cos\left(\frac{\pi}{n}\right)}\exp\left(\frac{-\pi i}{n}\right),$$

and the Euclidean radius of C_{n-1} is

$$\sqrt{s^2-1} = \sqrt{\frac{(r^2+1)^2}{4r^2\cos^2\left(\frac{\pi}{n}\right)} - 1}.$$

The interior angle $\alpha(r)$ of $P_n(r)$ at $r = rp_0$ is equal to the angle between C_0 and C_{n-1}, and hence it satisfies

$$\begin{aligned}\cos(\alpha(r)) &= \frac{2(s^2-1) - \left|s\exp\left(\frac{\pi i}{n}\right) - s\exp\left(\frac{-\pi i}{n}\right)\right|^2}{2(s^2-1)} \\ &= \frac{2(s^2-1) - 4s^2\sin^2\left(\frac{\pi}{n}\right)}{2(s^2-1)} \\ &= 1 - \frac{2(r^2+1)^2\sin^2\left(\frac{\pi}{n}\right)}{(r^2+1)^2 - 4r^2\cos^2\left(\frac{\pi}{n}\right)}.\end{aligned}$$

As $(r^2+1)^2 - 4r^2\cos^2\left(\frac{\pi}{n}\right) = (r^2-1)^2 + 4r^2\sin^2\left(\frac{\pi}{n}\right)$, the denominator is never zero for $0 < r < 1$. The continuity of $\alpha(r)$ then follows immediately from the continuity of the right-hand side of this expression and the continuity of arccos.

5.21: For $n \geq 5$, the interval of possible angles of a regular hyperbolic n-gon is $(0, \frac{n-2}{n}\pi)$. As $\frac{n-2}{n} > \frac{1}{2}$ for $n \geq 5$, this interval contains $\frac{1}{2}\pi$, and so there exists a regular hyperbolic n-gon with all right angles.

5.22: The hyperbolic length of the side of $P_n(r)$ joining $rp_0 = r$ to $rp_1 = r\exp\left(\frac{2\pi i}{n}\right)$ is equal to $d_{\mathbb{D}}\left(r, r\exp\left(\frac{2\pi i}{n}\right)\right)$.

Set $\theta = \frac{2\pi}{n}$. To calculate $d_{\mathbb{D}}(r, re^{i\theta})$, we first choose an element m of Möb($\mathbb{D}$) satisfying $m(r) = 0$. Write $m(z) = \frac{\alpha z + \beta}{\overline{\beta}z + \overline{\alpha}}$, where $|\alpha|^2 - |\beta|^2 = 1$.

As $m(r) = \frac{\alpha r + \beta}{\overline{\beta}r + \overline{\alpha}} = 0$, we have that $\beta = -\alpha r$, and so

$$m(z) = \frac{\alpha z - \alpha r}{-\overline{\alpha}rz + \overline{\alpha}} = \frac{\alpha(z-r)}{\overline{\alpha}(-rz+1)}.$$

Thus,

$$m(re^{i\theta}) = \frac{\alpha r(e^{i\theta}-1)}{\overline{\alpha}(-r^2e^{i\theta}+1)}$$

and

$$|m(re^{i\theta})| = \left|\frac{\alpha r(e^{i\theta}-1)}{\overline{\alpha}(-r^2e^{i\theta}+1)}\right| = \left|\frac{r(e^{i\theta}-1)}{-r^2e^{i\theta}+1}\right|.$$

Hence,

$$\begin{aligned}&d_{\mathbb{D}}(r, re^{i\theta}) \\ &\quad = d_{\mathbb{D}}(m(r), m(re^{i\theta})) \\ &\quad = d_{\mathbb{D}}(0, m(re^{i\theta})) \\ &\quad = \ln\left[\frac{1 + |m(re^{i\theta})|}{1 - |m(re^{i\theta})|}\right] \\ &\quad = \ln\left[\frac{|-r^2e^{i\theta}+1| + |r(e^{i\theta}-1)|}{|-r^2e^{i\theta}+1| - |r(e^{i\theta}-1)|}\right] \\ &\quad = \ln\left[\frac{(1+r^2)^2 - 4r^2\cos(\theta) + 2r\sqrt{2(1-2r^2\cos(\theta)+r^4)(1-\cos(\theta))}}{(1-r^2)^2}\right].\end{aligned}$$

5.23: These follow directly, with some algebraic massage, from the definition of $\cosh(x)$ and $\sinh(x)$ in terms of e^x, namely,

$$\cosh(x) = \frac{1}{2}\left(e^x + e^{-x}\right) \text{ and } \sinh(x) = \frac{1}{2}\left(e^x - e^{-x}\right).$$

5.24: As both $\sinh(c)$ and $\sin(\gamma)$ are positive, we consider instead the quantity $\frac{\sinh^2(c)}{\sin^2(\gamma)}$.

Write $A = \cosh(a)$, $B = \cosh(b)$, and $C = \cosh(c)$. By the law of cosines I, we have that

$$\sin^2(\gamma) = 1 - \cos^2(\gamma) = 1 - \left(\frac{AB - C}{\sinh(a)\sinh(b)}\right)^2.$$

Multiplying through, we get

$$\begin{aligned}
\sin^2(\gamma)\sinh^2(a)\sinh^2(b) &= \sinh^2(a)\sinh^2(b) - (AB - C)^2 \\
&= \sinh^2(a)\sinh^2(b) - A^2B^2 - C^2 + 2ABC \\
&= (A^2 - 1)(B^2 - 1) - A^2B^2 - C^2 + 2ABC \\
&= A^2B^2 - A^2 - B^2 + 1 - A^2B^2 - C^2 + 2ABC \\
&= 1 - A^2 - B^2 - C^2 + 2ABC.
\end{aligned}$$

Hence, we have that

$$\frac{\sin^2(\gamma)}{\sinh^2(c)} = \frac{1 - A^2 - B^2 - C^2 + 2ABC}{\sinh^2(a)\sinh^2(b)\sinh^2(c)}.$$

As the right-hand side remains unchanged after permuting a, b, and c, and simultaneously permuting α, β, and γ, the left-hand side must be unchanged as well, and so we see that

$$\frac{\sinh^2(c)}{\sin^2(\gamma)} = \frac{\sinh^2(b)}{\sin^2(\beta)} = \frac{\sinh^2(a)}{\sin^2(\alpha)}.$$

Taking square roots, we obtain the law of sines.

For the law of cosines II, start with the law of cosines I, which gives that

$$\cos(\gamma) = \frac{AB - C}{\sinh(a)\sinh(b)} = \frac{AB - C}{\sqrt{(A^2 - 1)(B^2 - 1)}}.$$

Applying the law of cosines II to the other two vertices gives

$$\cos(\alpha) = \frac{BC - A}{\sqrt{(B^2 - 1)(C^2 - 1)}} \text{ and } \sin(\alpha) = \frac{\sqrt{1 + 2ABC - A^2 - B^2 - C^2}}{\sqrt{(B^2 - 1)(C^2 - 1))}},$$

and that

$$\cos(\beta) = \frac{AC - B}{\sqrt{(A^2 - 1)(C^2 - 1)}} \text{ and } \sin(\beta) = \frac{\sqrt{1 + 2ABC - A^2 - B^2 - C^2}}{\sqrt{(A^2 - 1)(C^2 - 1))}}.$$

Hence,

$$\begin{aligned}\frac{\cos(\gamma)+\cos(\alpha)\cos(\beta)}{\sin(\alpha)\sin(\beta)} &= \frac{(BC-A)(AC-B)+(AB-C)(C^2-1)}{1+2ABC-A^2-B^2-C^2} \\ &= C=\cosh(c),\end{aligned}$$

as desired.

5.25: Consider the hyperbolic law of cosines I with $\alpha=\frac{\pi}{2}$. Let a be the hyperbolic length of the side of T opposite the vertex with angle α, and let b and c be the hyperbolic lengths of the sides adjacent to the vertex with angle α. Then,

$$\cosh(a)=\cosh(b)\cosh(c).$$

5.26: Every point a on A has the form $a=\alpha e^{i\theta}$. Hence, $m(a)=\lambda a=\lambda\alpha e^{i\theta}$.

Calculating, we see that

$$\begin{aligned}\mathrm{d}_{\mathbb{H}}\left(\alpha e^{i\theta},\lambda\alpha e^{i\theta}\right) &= \mathrm{d}_{\mathbb{H}}\left(e^{i\theta},\lambda e^{i\theta}\right) \\ &= \mathrm{d}_{\mathbb{H}}\left(e^{i\theta}-\cos(\theta),\lambda e^{i\theta}-\cos(\theta)\right) \\ &= \mathrm{d}_{\mathbb{H}}\left(i\sin(\theta),(\lambda-1)\cos(\theta)+i\lambda\sin(\theta)\right) \\ &= \mathrm{d}_{\mathbb{H}}\left(i,(\lambda-1)\cot(\theta)+\lambda i\right).\end{aligned}$$

Write $(\lambda-1)\cot(\theta)+\lambda i=\rho^{i\varphi}$. Calculating, we see that

$$\rho=\sqrt{(\lambda-1)^2\cot^2(\theta)+\lambda^2}$$

and

$$\csc(\varphi)=\frac{\rho}{\lambda}.$$

By Exercise 5.25, we have that $\mathrm{d}_{\mathbb{H}}\left(i,(\lambda-1)\cot(\theta)+i\lambda\right)$ satisfies

$$\cosh(\mathrm{d}_{\mathbb{H}}(i,(\lambda-1)\cot(\theta)+\lambda i))=\cosh(\mathrm{d}_{\mathbb{H}}(i,\rho i))\cosh(\mathrm{d}_{\mathbb{H}}(\rho i,\rho e^{i\varphi})).$$

By the solution to Exercise 3.20, we have that

$$\mathrm{d}_{\mathbb{H}}(\rho i,\rho e^{i\varphi})=\ln\left[\frac{1+\cos(\varphi)}{\sin(\varphi)}\right],$$

and so

$$\cosh(\mathrm{d}_{\mathbb{H}}(\rho i,\rho e^{i\varphi}))=\csc(\varphi).$$

Hence,

$$\begin{aligned}\cosh\left(\mathrm{d}_{\mathbb{H}}\left(\alpha e^{i\theta},\lambda\alpha e^{i\theta}\right)\right) &= \cosh(\ln(\rho))\csc(\varphi) \\ &= \frac{1}{2\lambda}(\rho^2+1) \\ &= \frac{1}{2}\left(\frac{(\lambda-1)^2}{\lambda}\cot^2(\theta)+\frac{\lambda^2+1}{\lambda}\right) \\ &= (\cosh(\ln(\lambda))-1)\cot^2(\theta)+\cosh(\ln(\lambda)).\end{aligned}$$

5.27: Write $z = t\exp(i\theta)$. Let T_1 be the hyperbolic triangle in $\mathbb{D}$ with vertices at 0, r, and z. Note that

$$\cosh(\mathrm{d}_{\mathbb{D}}(0,r)) = \cosh\left(\ln\left(\frac{1+r}{1-r}\right)\right) = \frac{1+r^2}{1-r^2}$$

and

$$\sinh(\mathrm{d}_{\mathbb{D}}(0,r)) = \sinh\left(\ln\left(\frac{1+r}{1-r}\right)\right) = \frac{2r}{1-r^2}.$$

Applying the law of cosines I to T_1, we see that

$$\begin{aligned}\cosh(\mathrm{d}_{\mathbb{D}}(z,r)) &= \cosh(\mathrm{d}_{\mathbb{D}}(0,r))\cosh(\mathrm{d}_{\mathbb{D}}(0,z)) \\ &\quad - \sinh(\mathrm{d}_{\mathbb{D}}(0,r))\sinh(\mathrm{d}_{\mathbb{D}}(0,z))\cos(\theta) \\ &= \left(\frac{1+r^2}{1-r^2}\right)\left(\frac{1+t^2}{1-t^2}\right) - \left(\frac{2r}{1-r^2}\right)\left(\frac{2t}{1-t^2}\right)\cos(\theta)\end{aligned}$$

and

$$\begin{aligned}\cosh(\mathrm{d}_{\mathbb{D}}(z,-r)) &= \cosh(\mathrm{d}_{\mathbb{D}}(0,-r))\cosh(\mathrm{d}_{\mathbb{D}}(0,z)) \\ &\quad - \sinh(\mathrm{d}_{\mathbb{D}}(0,-r))\sinh(\mathrm{d}_{\mathbb{D}}(0,z))\cos(\pi-\theta) \\ &= \left(\frac{1+r^2}{1-r^2}\right)\left(\frac{1+t^2}{1-t^2}\right) - \left(\frac{2r}{1-r^2}\right)\left(\frac{2t}{1-t^2}\right)\cos(\pi-\theta) \\ &= \left(\frac{1+r^2}{1-r^2}\right)\left(\frac{1+t^2}{1-t^2}\right) + \left(\frac{2r}{1-r^2}\right)\left(\frac{2t}{1-t^2}\right)\cos(\theta).\end{aligned}$$

Adding these two equations, we see that

$$\cosh(\mathrm{d}_{\mathbb{D}}(z,r)) + \cosh(\mathrm{d}_{\mathbb{D}}(z,-r)) = 2\left(\frac{1+r^2}{1-r^2}\right)\left(\frac{1+t^2}{1-t^2}\right).$$

Recall that r is fixed and we wish for $\cosh(\mathrm{d}_{\mathbb{D}}(z,r)) + \cosh(\mathrm{d}_{\mathbb{D}}(z,-r))$ to be constant, which implies that t is constant, by the above equation. As this equation is independent of θ, we see that the set for which $\cosh(\mathrm{d}_{\mathbb{D}}(t\exp(i\theta),r)) + \cosh(\mathrm{d}_{\mathbb{D}}(t\exp(i\theta),-r))$ is constant is thus a Euclidean (or hyperbolic) circle centred at 0, namely, the points $t\exp(i\theta)$ for t constant and θ arbitrary.

5.28: The fact that the three interior angles are equal follows immediately from the hyperbolic law of cosines I, namely, that

$$\cos(\alpha) = \frac{\cosh^2(a) - \cosh(a)}{\sinh^2(a)}.$$

Consider the hyperbolic triangle T' formed by bisecting the hyperbolic triangle T, so that T' has angles α, $\frac{1}{2}\alpha$, and $\frac{1}{2}\pi$, and has the corresponding hyperbolic lengths of the opposite sides being b, $\frac{1}{2}a$, and a, where b is as yet undetermined.

Applying the hyperbolic law of cosines I to T', we obtain

$$\begin{aligned}\cos\left(\frac{1}{2}\alpha\right) &= -\cos\left(\frac{1}{2}\pi\right)\cos(\alpha) + \sin\left(\frac{1}{2}\pi\right)\sin(\alpha)\cosh\left(\frac{1}{2}a\right) \\ &= \sin(\alpha)\cosh\left(\frac{1}{2}a\right) \\ &= 2\sin\left(\frac{1}{2}\alpha\right)\cos\left(\frac{1}{2}\alpha\right)\cosh\left(\frac{1}{2}a\right).\end{aligned}$$

Dividing through by $\cos\left(\frac{1}{2}\alpha\right)$, we obtain

$$1 = 2\sin\left(\frac{1}{2}\alpha\right)\cosh\left(\frac{1}{2}a\right),$$

as desired.

5.29: We work in $\mathbb{H}$. Let v_α, v_β, and v_γ be the vertices of the hyperbolic triangle T with interior angles α, β, and γ, respectively. As each interior angle is positive, there is a unique hyperbolic ray from a vertex of T into T that bisects that angle at that vertex; this hyperbolic ray is the angle bisector. Draw the angle bisectors from v_α and v_β; these intersect at a point p inside T. Let $v = \mathrm{d}_{\mathbb{H}}(v_\alpha, p)$ and $w = \mathrm{d}_{\mathbb{H}}(v_\beta, p)$. Draw the hyperbolic line segment from p to v_γ, and let $m = \mathrm{d}_{\mathbb{H}}(v_\gamma, p)$.

This hyperbolic line segment divides the angle γ at v_γ into subangles γ_1 and γ_2; note that $\gamma = \gamma_1 + \gamma_2$. We wish to show that $\gamma_1 = \gamma_2$.

These three hyperbolic line segments from the vertices of T to p divide T up into three smaller hyperbolic triangles. We apply the hyperbolic law of sines to these hyperbolic triangles.

For the hyperbolic triangle with vertices v_α, v_γ, and p, we see that

$$\frac{\sinh(m)}{\sin(\alpha/2)} = \frac{\sinh(v)}{\sin(\gamma_2)};$$

for the hyperbolic triangle with vertices v_β, v_γ, and p, we see that

$$\frac{\sinh(m)}{\sin(\beta/2)} = \frac{\sinh(w)}{\sin(\gamma_1)};$$

for the hyperbolic triangle with vertices v_α, v_β, and p, we see that

$$\frac{\sinh(w)}{\sin(\alpha/2)} = \frac{\sinh(v)}{\sin(\beta/2)}.$$

So,

$$\begin{aligned}\frac{\sinh(m)}{\sin(\alpha/2)} &= \frac{\sinh(v)}{\sin(\gamma_2)} \\ &= \frac{1}{\sin(\gamma_2)}\,\frac{\sinh(w)}{\sin(\alpha/2)}\,\sin(\beta/2) \text{ (using the third equation)} \\ &= \frac{1}{\sin(\gamma_2)}\,\frac{\sin(\beta/2)}{\sin(\alpha/2)}\,\frac{\sin(\gamma_1)}{\sin(\beta/2)}\,\sinh(m) \text{ (using the second equation)} \\ &= \frac{\sin(\gamma_1)}{\sin(\gamma_2)}\,\frac{\sinh(m)}{\sin(\alpha/2)},\end{aligned}$$

and so $\frac{\sin(\gamma_1)}{\sin(\gamma_2)} = 1$ (because $\sinh(m) \neq 0$). Hence, $\sin(\gamma_1) = \sin(\gamma_2)$. As $0 < \gamma_1, \gamma_2 < \pi$, either $\gamma_1 = \gamma_2$ or $\gamma_1 = \pi - \gamma_2$; the latter case is excluded by the Gauss–Bonnet formula, and so $\gamma_1 = \gamma_2$ as desired.

5.30: Let v be the vertex at which the sides (of lengths) B and A meet, and let w be the vertex at which the sides (of lengths) C and D meet. (We adopt the notational convention that we refer to a side of R and the length of that side with the same label.) Draw the hyperbolic line segment c between v and w. This line segment splits R into two hyperbolic triangles: One hyperbolic triangle, T_1, has angles δ_1 at v, φ, and γ_1 at w, whereas the other hyperbolic triangle T_2 has angles δ_2 at v, $\frac{\pi}{2}$, and γ_2 at w. Note that $\delta_1 + \delta_2 = \frac{\pi}{2} = \gamma_1 + \gamma_2$.

As $\delta_1 + \delta_2 = \frac{\pi}{2} = \gamma_1 + \gamma_2$, we have that

$$\cos(\delta_1) = \cos\left(\frac{\pi}{2} - \delta_2\right) = \sin(\delta_2),$$
$$\sin(\delta_1) = \sin\left(\frac{\pi}{2} - \delta_2\right) = \cos(\delta_2),$$

and

$$\cos(\gamma_1) = \cos\left(\frac{\pi}{2} - \gamma_2\right) = \sin(\gamma_2),$$
$$\sin(\gamma_1) = \sin\left(\frac{\pi}{2} - \gamma_2\right) = \cos(\gamma_2).$$

Applying the law of sines to T_2, we see that

$$\frac{\sin(\delta_2)}{\sinh(C)} = \frac{1}{\sinh(c)} = \frac{\sin(\gamma_2)}{\sinh(B)}.$$

Applying the law of cosines II to T_2, we see that

$$\cos(\delta_2) = \sin(\gamma_2)\cosh(C)$$

and

$$\cos(\gamma_2) = \sin(\delta_2)\cosh(B).$$

Finally, applying the hyperbolic Pythagorean theorem to T_2, we see that

$$\cosh(c) = \cosh(C)\cosh(B).$$

Applying the law of cosines II to T_1, making use of the relationships derived above, and simplifying, we see that

$$\begin{aligned}
\cos(\varphi) &= -\cos(\delta_1)\cos(\gamma_1)+\sin(\delta_1)\sin(\gamma_1)\cosh(c)\\
&= -\sin(\delta_2)\sin(\gamma_2)+\cos(\delta_2)\cos(\gamma_2)\cosh(c)\\
&= -\frac{\sinh(C)\sinh(B)}{\sinh^2(c)}+\sin(\gamma_2)\cosh(C)\sin(\delta_2)\cosh(B)\cosh(c)\\
&= -\frac{\sinh(C)\sinh(B)}{\sinh^2(c)}+\frac{\sinh(B)\cosh(C)\sinh(C)\cosh(B)\cosh(c)}{\sinh^2(c)}\\
&= \frac{-\sinh(C)\sinh(B)+\sinh(B)\sinh(C)\cosh^2(c)}{\sinh^2(c)}\\
&= \frac{(-1+\cosh^2(c))\sinh(C)\sinh(B)}{\sinh^2(c)}=\sinh(C)\sinh(B),
\end{aligned}$$

as desired.

For the second identity, apply the law of cosines II to T_1 to obtain

$$\cos(\gamma_1)=-\cos(\varphi)\cos(\delta_1)+\sin(\varphi)\sin(\delta_1)\cosh(A).$$

Making use of the relationships derived above, and simplifying, we see that

$$\begin{aligned}
\cosh(A)\sin(\varphi) &= \frac{\cos(\gamma_1)+\cos(\varphi)\cos(\delta_1)}{\sin(\delta_1)}\\
&= \frac{\sin(\gamma_2)+\sinh(C)\sinh(B)\sin(\delta_2)}{\cos(\delta_2)}\\
&= \frac{\sinh(B)+\sinh^2(C)\sinh(B)}{\sinh(c)\cos(\delta_2)}\\
&= \frac{\sinh(B)\cosh^2(C)}{\sinh(c)\cos(\delta_2)}\\
&= \frac{\sinh(B)\cosh^2(C)}{\sinh(c)\cosh(C)\sin(\gamma_2)}=\cosh(C),
\end{aligned}$$

as desired.

Solutions to Chapter 6.1 exercises:

6.1: We can write $q(\mathbf{x})=\mathbf{x}^{\mathrm{T}}B\mathbf{x}$, where

$$B=\begin{pmatrix} -4 & 7 & -1\\ 7 & 2 & -8\\ -1 & -8 & 10\end{pmatrix}.$$

The eigenvalues of B, and hence of q, are (approximately) 16.0537, −9.4049, and 1.3511.

6.2: Note that because $\mathbf{x}^{\mathrm{T}}A\mathbf{x}$ is a 1×1 matrix, it is necessarily symmetric and so

$$\mathbf{x}^{\mathrm{T}}A\mathbf{x} = (\mathbf{x}^{\mathrm{T}}A\mathbf{x})^{\mathrm{T}} = \mathbf{x}^{\mathrm{T}}A^{\mathrm{T}}\mathbf{x}.$$

Hence,

$$f(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}A\mathbf{x} = \frac{1}{2}\left(\mathbf{x}^{\mathrm{T}}A\mathbf{x} + \mathbf{x}^{\mathrm{T}}A^{\mathrm{T}}\mathbf{x}\right) = \mathbf{x}^{\mathrm{T}}\left(\frac{1}{2}\left(A + A^{\mathrm{T}}\right)\right)\mathbf{x}.$$

As $\frac{1}{2}(A + A^{\mathrm{T}})$ is symmetric, we have that $B = \frac{1}{2}\left(A + A^{\mathrm{T}}\right)$.

6.3: Let $\mathbf{x} \in \mathbb{R}^3$ be arbitrary. As $AA^{-1} = I_3$, we have that

$$q(\mathbf{x}) = q(I_3\mathbf{x}) = q(AA^{-1}\mathbf{x}) = q(A^{-1}\mathbf{x}),$$

where the third equality follows from $A \in \mathcal{O}(q)$.

6.4: Write $q(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}B\mathbf{x}$, where B is a symmetric 3×3 matrix. As q is assumed to be degenerate, we know that B is not invertible. Hence, 0 is an eigenvalue of B. Let α_1 and α_2 be the other two eigenvalues of B.

As B is symmetric, it is diagonalizable, and so there exists a matrix C so that

$$C^{-1}BC = E = \begin{pmatrix} 0 & 0 & 0 \\ 0 & \alpha_1 & 0 \\ 0 & 0 & \alpha_2 \end{pmatrix}.$$

We choose C to be the matrix whose columns are the eigenvectors of B normalized to have unit length, so that $C^{-1} = C^{\mathrm{T}}$. Then, $B = CEC^{\mathrm{T}}$. Therefore, we can rewrite q as

$$q(\mathbf{x}) = x^{\mathrm{T}}Bx = x^{\mathrm{T}}CEC^{\mathrm{T}}\mathbf{x} = (C^{\mathrm{T}}\mathbf{x})^{\mathrm{T}}EC^{\mathrm{T}}\mathbf{x}.$$

For

$$\mathbf{x} = C\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix},$$

we see that

$$\begin{aligned} q(\mathbf{x}) &= (C^{\mathrm{T}}\mathbf{x})^{\mathrm{T}}EC^{\mathrm{T}}\mathbf{x} \\ &= \mathbf{x}^{\mathrm{T}}CE\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = (0), \end{aligned}$$

as desired.

References

[1] W. Abikoff, 'The bounded model for hyperbolic 3-space and a quaternionic uniformization theorem', *Math. Scand.* **54** (1984), 5–16.

[2] W. Abikoff, 'The uniformization theorem', *Amer. Math. Monthly* **88** (1981), 574–592.

[3] L. V. Ahlfors, *Complex Analysis*, McGraw Hill, New York, 1979.

[4] L. V. Ahlfors, *Conformal Invariants*, McGraw-Hill, New York, 1973.

[5] L. V. Ahlfors, *Möbius Transformations in Several Dimensions*, Ordway Professorship Lectures in Mathematics, University of Minnesota, School of Mathematics, Minneapolis, Minn., 1981.

[6] H. Anton and R. C. Busby, *Contemporary Linear Algebra*, John Wiley and Sons, Inc., 2003.

[7] A. F. Beardon, *The Geometry of Discrete Groups*, Graduate Texts in Mathematics, Springer-Verlag, New York, 1983.

[8] A. F. Beardon, 'The hyperbolic metric of a rectangle', *Ann. Acad. Sci. Fenn.* **26** (2001), 401–407.

[9] A. F. Beardon, 'The hyperbolic metric in a rectangle II', *Ann. Acad. Sci. Fenn.* **28** (2003), 143–152.

[10] A. F. Beardon, 'An introduction to hyperbolic geometry', in *Ergodic Theory, Symbolic Dynamics, and Hyperbolic Spaces*, edited by T. Bedford, M. Keane, and C. Series, Oxford University Press, Oxford, 1991, 1–34.

[11] R. Bonola, *Non-Euclidean Geometry*, Dover Publications Inc., New York, 1955.

[12] B. H. Bowditch, 'Notes of Gromov's hyperbolicity criterion for path-metric spaces', in *Group Theory from a Geometrical Viewpoint* (Proceedings of the workshop held in Trieste, March 26–April 6, 1990), edited by É. Ghys, A. Haefliger, and A. Verjovsky, World Scientific Publishing Co., Inc., River Edge, NJ, 1991, 64–167.

[13] Yu. D. Burago and V. A. Zalgaller, *Geometric Inequalities*, *Grundlehren der Mathematischen Wissenschaften* 285, Springer-Verlag, New York, 1988.

[14] H. S. M. Coxeter, *Non-Euclidean geometry*, *Mathematical Expositions* 2, University of Toronto Press, Toronto, 1978.

[15] W. Fenchel, *Elementary Geometry in Hyperbolic Space*, *de Gruyter Studies in Mathematics* 11, Walter de Gruyter, New York, 1989.

[16] M. J. Greenberg, *Euclidean and Non-Euclidean Geometries*, W. H. Freeman and Co., New York, 1993.

[17] M. Gromov, 'Hyperbolic groups', in *Essays in Group Theory*, Math. Sci. Res. Inst. Publ. **8**, Springer, New York, 1987, 75–263.

[18] V. Guillemin and A. Pollack, *Differential Topology*, Prentice-Hall, Inc., Englewood Cliffs, NJ, 1974.

[19] H. von Helmholz, 'On the origin and significance of the geometrical axioms', in *The World of Mathematics*, volume 1, edited by J. R. Newman, Simon and Schuster, New York, 1956, 647–668.

[20] I. N. Herstein, *Abstract Algebra*, Prentice Hall Inc., Upper Saddle River, NJ, 1996.

[21] E. Hille, *Analytic Function Theory*, Chelsea, New York, 1977.

[22] B. Iversen, *Hyperbolic geometry*, *London Mathematical Society Student Texts* 25, Cambridge University Press, Cambridge, 1992.

[23] G. A. Jones and D. Singerman, *Complex Functions, an Algebraic and Geometric Viewpoint*, Cambridge University Press, Cambridge, 1987.

[24] P. J. Kelly and G. Matthews, *The Non-Euclidean, Hyperbolic Plane*, *Universitext*, Springer-Verlag, New York, 1981.

[25] J. L. Locher, editor, *M. C. Escher, His Life and Complete Graphic Work*, Abradale Press, New York, 1992.

[26] J. Munkres, *Topology, a First Course*, Prentice-Hall, Inc., Englewood Cliffs, NJ, 1975.

[27] D. Pedoe, *Geometry – a Comprehensive Course*, Dover Publications, New York, 1988.

[28] L. Rédei, *Foundation of Euclidean and Non-Euclidean Geometries According to F. Klein*, *International Series of Monographs in Pure and Applied Mathematics* 97, Pergamon Press, Oxford, 1968.

[29] B. A. Rosenfeld, *A History of Non-Euclidean Geometry*, Springer-Verlag, New York, 1988.

[30] D. Schattschneider, *Visions of Symmetry*, W. H. Freeman and Company, New York, 1990.

[31] S. Stahl, *The Poincaré Half-Plane*, Jones and Bartlett, Boston, 1993.

[32] J. Stillwell, 'Poincaré, geometry and topology', in *Henri Poincaré: Science et Philosophie*, Akademie Verlag, Berlin, 1996, 231–240.

[33] J. Stillwell, *Sources of Hyperbolic Geometry*, History of Mathematics, volume 10, American Mathematical Society, Providence, RI, 1996.

[34] G. Strang, *Linear Algebra and its Applications*, Academic Press, New York, 1980.

[35] W. P. Thurston, *Three-Dimensional Geometry and Topology*, Princeton University Press, Princeton, NJ, 1997.

[36] R. J. Trudeau, *The Non-Euclidean Revolution*, Birkhäuser Boston, Boston, MA, 1987.

[37] C. R. Wylie, Jr., *Foundations of Geometry*, McGraw-Hill Book Co., New York, 1964.

List of Notation

The purpose of this section is to provide a list of the various bits of notation that appear throughout the book. The chapter or section in which the notation first appears is given in brackets.

$\mathbb{R}$	real numbers [1.1]
$\mathbb{C}$	complex numbers [1.1]
$\mathrm{Re}(z)$	$= x$, the real part of the complex number $z = x + iy$ [1.1]
$\mathrm{Im}(z)$	$= y$, the imaginary part of $z = x + iy$ [1.1]
$\lvert z\rvert$	$= \sqrt{(\mathrm{Re}(z))^2 + (\mathrm{Im}(z))^2}$, the norm or modulus of z [1.1]
$\mathbb{H}$	$= \{z \in \mathbb{C} \mid \mathrm{Im}(z) > 0\}$, upper half plane in $\mathbb{C}$ [1.1]
$\mathbb{S}^1$	$= \{z \in \mathbb{C} \mid \lvert z\rvert = 1\}$, unit circle in $\mathbb{C}$ [1.1]
$\mathbb{N}$	$= \{1,\ 2\ ,3, \ldots\}$, natural numbers [1.2]
$\mathbb{Q}$	rational numbers [1.2]
$\min(x, y)$	minimum of two real numbers x and y [1.2]
$\overline{\mathbb{C}}$	$= \mathbb{C} \cup \{\infty\}$, Riemann sphere [1.2]
$\mathbb{Z}$	$= \{0,\ \pm 1,\ \pm 2, \ldots\}$, integers [1.2]
$\mathbb{S}^2$	unit sphere in $\mathbb{R}^3$ [1.2]
∂X	topological boundary of X in $\overline{\mathbb{C}}$ [1.2, 3.1, 4.2]
$U_\varepsilon(z)$	open disc of radius ε and centre z [1.2, 5.2, 3.3]
$U_\varepsilon(\infty)$	open disc in $\overline{\mathbb{C}}$ of radius ε and centre ∞ [1.2]
$\overline{\mathbb{R}}$	$= \mathbb{R} \cup \{\infty\}$, the extended real line [1.2]
$\overline{X}$	closure of a set X in $\overline{\mathbb{C}}$ [1.2]
$\mathrm{Homeo}(\overline{\mathbb{C}})$	group of homeomorphisms of $\overline{\mathbb{C}}$ [1.2]

$\mathbb{D}$	$= \{z \in \mathbb{C} \mid \lvert z \rvert < 1\}$, open unit disc in $\mathbb{C}$ [1.3, 4.1]
$\mathrm{Homeo}^{\mathrm{C}}(\overline{\mathbb{C}})$	set of circle preserving homeomorphisms of $\overline{\mathbb{C}}$ [2.1]
$\mathrm{Möb}^+$	group of Möbius transformations [2.1]
$[z_1, z_2; z_3, z_4]$	cross ratio [2.3]
$\mathrm{Möb}$	general Möbius group [2.6]
$\mathrm{angle}(C_1, C_2)$	angle between curves [2.7]
$\mathrm{Möb}(\mathbb{H})$	subgroup of Möb preserving $\mathbb{H}$ [2.8]
$\mathrm{Möb}(\overline{\mathbb{R}})$	subgroup of Möb preserving $\overline{\mathbb{R}}$ [2.8]
$\mathrm{Möb}(\mathbb{S}^1)$	subgroup of Möb preserving $\mathbb{S}^1$ [2.8]
$\mathrm{Möb}(A)$	subgroup of Möb preserving the circle A [2.8]
$\mathrm{Möb}(\mathbb{D})$	subgroup of Möb preserving $\mathbb{D}$ [2.8, 4.1]
$\mathrm{Möb}^+(\mathbb{H})$	subgroup of $\mathrm{Möb}^+$ preserving $\mathbb{H}$ [2.8]
$\mathrm{bij}(X)$	group of bijections from X to X [2.2]
$\det(m)$	determinant of the Möbius transformation m [2.5]
$\tau(m)$	square of the trace of the Möbius transformation m [2.5]
$\mathrm{GL}_2(\mathbb{C})$	group of 2×2 invertible matrices over $\mathbb{C}$ [2.5]
$\mathrm{PGL}_2(\mathbb{C})$	$= \mathrm{GL}_2(\mathbb{C})/\{\alpha I \mid \alpha \in \mathbb{C},\ \alpha \neq 0\}$ [2.5]
$\mathrm{SL}_2(\mathbb{C})$	group of 2×2 matrices with determinant 1 over $\mathbb{C}$ [2.5]
$\rho(z)\,\lvert\mathrm{d}z\rvert$	general element of arc-length [3.1]
$\mathrm{length}_\rho(f)$	length of piecewise C^1 path f with respect to $\rho(z)\,\lvert\mathrm{d}z\rvert$ [3.1]
$\mathrm{length}_{\mathbb{H}}(f)$	hyperbolic length in $\mathbb{H}$ of piecewise C^1 path f [3.2]
(X, d)	a metric space [3.3]
$\mathit{\Gamma}[x, y]$	paths in $\mathbb{H}$ from x to y [3.3]
$\mathrm{d}_{\mathbb{H}}(x, y)$	hyperbolic distance between points x, $y \in \mathbb{H}$ [3.4]
$(\mathbb{H}, \mathrm{d}_{\mathbb{H}})$	upper half-plane as a metric space [3.4]
$\mathrm{Isom}(\mathbb{H}, \mathrm{d}_{\mathbb{H}})$	group of isometries of $(\mathbb{H}, \mathrm{d}_{\mathbb{H}})$ [3.6]
$\mathrm{d}_{\mathbb{H}}(X, Y)$	hyperbolic distance between sets X, Y in $\mathbb{H}$ [3.7]
$\mathrm{Möb}^+(\mathbb{D})$	subgroup of $\mathrm{Möb}^+$ preserving $\mathbb{D}$ [4.1]
$\mathrm{length}_{\mathbb{D}}(f)$	hyperbolic length in $\mathbb{D}$ of piecewise C^1 path f [4.1]
$\mathit{\Theta}[x, y]$	paths in $\mathbb{D}$ from x to y [4.1]
$\mathrm{d}_{\mathbb{D}}(x, y)$	hyperbolic distance between points x, $y \in \mathbb{D}$ [4.1]
$(\mathbb{D}, \mathrm{d}_{\mathbb{D}})$	Poincare disc as a metric space [4.1]
$\mathrm{Isom}(\mathbb{D}, \mathrm{d}_{\mathbb{D}})$	group of isometries of $(\mathbb{D}, \mathrm{d}_{\mathbb{D}})$ [4.1]
ds_X	hyperbolic element of arc-length on holomorphic disc X [4.2]

$\mathrm{length}_X(f)$	hyperbolic length on holomorphic disc X of path f [4.2]
$\mathrm{Isom}(X, \mathrm{d}_X)$	group of isometries of (X, d_X) [4.2]
$\mathrm{curv}(z)$	curvature [4.2]
$\mathrm{conv}(X)$	convex hull of X [5.1]
$\mathrm{area}_{\mathbb{H}}(X)$	hyperbolic area of X in $\mathbb{H}$ [5.3]
$\mathrm{area}_{\mathbb{D}}(X)$	hyperbolic area of X in $\mathbb{D}$ [5.3]
$Q(\mathbf{x})$	$= -x_0^2 + x_1^2 + x_2^2$, quadratic form on $\mathbb{R}^3$ [6.1]
$\mathcal{O}(Q)$	group of matrices perserving quadratic form $Q(\mathbf{x})$ [6.1]
$\mathrm{GL}_3(\mathbb{R})$	group of 3×3 invertible matrices over $\mathbb{R}$ [6.1]
S_c	$= \{\mathbf{x} \in \mathbb{R}^3 \mid Q(\mathbf{x}) = c\}$, level set of $Q(\mathbf{x})$ [6.1]
$\mathbb{U}$	upper sheet of hyperboloid $Q(\mathbf{x}) = -1$ in $\mathbb{R}^3$ [6.1]
$\mathbb{L}$	lower sheet of hyperboloid $Q(\mathbf{x}) = -1$ in $\mathbb{R}^3$ [6.1]
$\mathcal{O}^+(Q)$	group of matrices perserving both $Q(\mathbf{x})$ and $\mathbb{U}$ [6.1]
$\nabla Q(\mathbf{x})$	gradient of $Q(\mathbf{x})$ [6.1]
$\mathrm{length}_{\mathbb{U}}(f)$	hyperbolic length of piecewise C^1 path f [6.1]
$\mathrm{d}_{\mathbb{U}}(\mathbf{u}, \mathbf{v})$	hyperbolic distance between points $\mathbf{u}$, $\mathbf{v} \in \mathbb{U}$ [6.1]
$\mathrm{O}(2,1)$	$= \mathcal{O}^+(Q)$ [6.1]
$\mathrm{SO}(2,1)$	subgroup of $\mathrm{O}(2,1)$ of elements of determinant 1 [6.1]
$\mathrm{Isom}(\mathbb{U}, \mathrm{d}_{\mathbb{U}})$	$= \mathcal{O}^+(Q)$, group of isometries of $(\mathbb{U}, \mathrm{d}_{\mathbb{U}})$ [6.1]
$Q_n(\mathbf{x})$	$= -x_0^2 + \sum_{j=1}^n x_j^2$, quadratic form on $\mathbb{R}^{n+1}$ [6.2]
$\mathbb{U}^n$	upper sheet of hyperboloid $Q_n(\mathbf{x}) = -1$ in $\mathbb{R}^{n+1}$ [6.2]
$\mathcal{O}(Q_n)$	group of matrices perserving quadratic form $Q_n(\mathbf{x})$ [6.2]
$\mathcal{O}^+(Q_n)$	group of matrices perserving both $Q_n(\mathbf{x})$ and $\mathbb{U}^n$ [6.2]
$\mathrm{O}(n,1)$	$= \mathcal{O}^+(Q_n)$ [6.2]
$\mathrm{length}_{\mathbb{U}^n}(f)$	hyperbolic length of piecewise C^1 path f [6.2]
$\mathrm{d}_{\mathbb{U}^n}(\mathbf{u}, \mathbf{v})$	hyperbolic distance between points $\mathbf{u}$, $\mathbf{v} \in \mathbb{U}^n$ [6.2]
$\mathrm{Isom}(\mathbb{U}^n, \mathrm{d}_{\mathbb{U}^n})$	$= \mathcal{O}^+(Q_n)$, group of isometries of $(\mathbb{U}^n, \mathrm{d}_{\mathbb{U}^n})$ [6.2]
$\overline{\mathbb{R}^n}$	one-point compactification of $\mathbb{R}^n$ [6.2]
$\mathrm{M\ddot{o}b}_n$	general Möbius group acting on $\overline{\mathbb{R}^n}$ [6.2]
$\mathrm{M\ddot{o}b}_n^+$	group of Möbius transformations acting on $\overline{\mathbb{R}^n}$ [6.2]
$\mathbb{H}^n$	upper half-space model of hyperbolic n-space [6.2]
$\mathbb{D}^n$	Poincare ball model of hyperbolic n-space [6.2]
$\mathrm{M\ddot{o}b}_n(\mathbb{H}^n)$	subgroup of $\mathrm{M\ddot{o}b}_n$ preserving $\mathbb{H}^n$ [6.2]
$\mathrm{length}_{\mathbb{H}^n}(f)$	hyperbolic length of piecewise C^1 path f [6.2]

$\mathrm{d}_{\mathbb{H}^n}(\mathbf{x}, \mathbf{y})$	hyperbolic distance between points $\mathbf{x}$, $\mathbf{y} \in \mathbb{H}^n$ [6.2]
$\mathrm{Isom}(\mathbb{H}^n, \mathrm{d}_{\mathbb{H}^n})$	$= \mathrm{M\ddot{o}b}_n(\mathbb{H}^n)$, group of isometries of $(\mathbb{H}^n, \mathrm{d}_{\mathbb{H}^n})$ [6.2]

Index