Geometrie 2021

Tom Ilmanen

December 10, 2021

Contents

I Beginning **6**

1 Preliminaries **7**
1 Metadata 8
2 References 9
3 Set theory 10
4 Complex numbers 17

2 Introduction **21**
5 First off 22
6 Euclidean space 26

3 Metric spaces **28**
7 Definition of metric spaces 29
8 Isometries 33
9 Examples of metric spaces 36
10 More examples of metric spaces 39

II Spheres **43**

4 Area and circumference of an intrinsic disk in S^2 **44**
11 Length and distance in S^2 45
12 Area and circumference of an intrinsic disk 49

5 Angle excess **56**
13 Angle excess 57
14 Proof of the angle excess formula 61

6 Stereographic projection **63**
15 The map problem 64
16 Stereographic projection 67
17 The spherical metric on $\mathbb{R}^2$ 71

7 Spherical arclength on S^2 **73**
18 Similarities of $\mathbb{R}^2$ 74
19 Spherical arclength 78

20 Stereographic projection is conformal 81

III Hyperbolic space 84

8 The hyperbolic metric 85
21 The hyperbolic metric 86
22 Sizes near the boundary 88

9 Geodesics 90
23 Minimizing curves and geodesics 91
24 The x-axis is a minimizing geodesic 93
25 Length along the x-axis 95
26 Geodesics in $\mathbb{H}^2$ 97

10 Circumference and area of a hyperbolic disk 101
27 Other expressions for the length 102
28 Circumference and area of a hyperbolic disk 103
29 Visualization and resources 105

11 The extended complex plane and clines 111
30 The extended complex plane and the Riemann sphere 112
31 Clines 113

12 Inversion 115
32 Inversion in a cline 116
33 Transferring operations between $\hat{\mathbb{C}}$ and S^2 118
34 Inversion takes clines to clines and is conformal 120
35 The stretch factor of $1/z$ 122

13 Möbius transformations 124
36 Möbius transformations 125
37 Handling ∞ correctly 126
38 Möbius transformations are invertible 128

14 The group of Möbius transformations 131
39 Transformation groups 132
40 The Möbius transformations form a group 135
41 Matrix multiplication and $\mathrm{Möb}_+$ 137

15 Factoring Möbius transformations 140
42 Factoring Möbius transformations 141
43 Möbius transformations are conformal and preserve clines 143
44 Relation of Möb to $\mathrm{Conf}(S^2)$ 144

16 Properties of Möbius transformations 145
45 Möbius transformations are 3-transitive 146

46 The cross ratio and its symmetries 149
47 The cross ratio is preserved 152
48 When the cross ratio is real 154

17 The elements of $\mathrm{Möb}_+(B_1)$ 156
49 Some elements of $\mathrm{Möb}_+(B_1)$ 157
50 Factoring elements of $\mathrm{Möb}_+(B_1)$ 162

18 Isometries, geodesics, and distances in $\mathbb{H}^2$ 163
51 Isometries of $\mathbb{H}^2$ 164
52 Geodesics of $\mathbb{H}^2$ 165
53 Cross ratio formula for distance 167
54 Arccosh formula for distance 168

IV End 169

19 Bibliography 170
55 Books 171
56 Articles, blogs, and references 173
57 Software, visualization, and activities 174

List of figures 175

Part I

Beginning

1

Preliminaries

§1 Metadata

Tom Ilmanen, lecturer

Raphael Appenzeller, organizer

Lectures Friday 14-16 weekly in HG F5:

24.09.; 01.10.; 08.10.; 15.10.; 22.10.; 29.10.; 05.11.; 12.11.; 19.11.; 26.11.; 03.12.; 10.12.; 17.12. (exam)

Exercise sections Monday 16-18 biweekly:

27.09.; 11.10.; 25.10.; 08.11.; 22.11.; 06.12.; 20.12.

Exercises are issued Friday week n, discussed in section Monday week $n+1$, due Monday week $n+2$, returned in section Monday week $n+3$, where n is odd.

Website: `https://metaphor.ethz.ch/x/2021/hs/401-1511-00L`

Exercises: `https://metaphor.ethz.ch/x/2021/hs/401-1511-00L`

Script: `https://metaphor.ethz.ch/x/2021/hs/401-1511-00L/literatur/script.pdf`

Forum: `https://forum.math.ethz.ch/t/geometrie-herbst-2021/277`

Exam: 17.12.20 in class.

§2 References

For more detail and additional sources, see §55, §56, §57.

Last year's script:

- T. Ilmanen, Geometrie 2020, `https://metaphor.ethz.ch/x/2020/hs/401-1511-00L/literatur/script.pdf`.
  The topics were different, but the older script has more about group theory.
  It also has many pictures and audiovisuals.

Very accessible:

- J. R. Weeks, *The Shape of Space*, recommended.
- M. Hitchman, *Geometry with an Introduction to Cosmic Topology*, `https://mphitchman.com/geometry/frontmatter.html`, recommended.
- E. A. Abbott, *Flatland*.

For fractals:

- Falconer, *The geometry of fractal sets*.

For group theory:

- D. Saracino, *Abstract Algebra: A First Course*.

For linear algebra:

- K. Jänich, *Lineare Algebra*.
- G. Fischer, *Lineare Algebra: Eine Einführung für Studienanfänger*.

For complex analysis:

- L. Ahlfors, *Complex Analysis*. Looking for a more available reference.

For hyperbolic geometry:

- J. W. Anderson, *Hyperbolic Geometry*, recommended.
- W. P. Thurston, *Three-dimensional Geometry and Topology*.
- B. Loustau, *Hyperbolic geometry*, online notes.
- A. F. Beardon, *The Geometry of Discrete Groups*.

Mathematical symbols:

- *Liste mathematischer Symbole*,
  `https://de.wikipedia.org/wiki/Liste_mathematischer_Symbole`

Mathematical dictionaries:

- G. Eisenreich, R. Sube, *Dictionary of Mathematics; Wörterbuch Mathematik*, Verlag Harry Deutsch, 1987.

Images:

- Details of the picture credits are in the List of Figures after Part IV.

§3 Set theory

References

- Saracino 1-3 (sets), 59-65 (functions), 80-82 (equivalence relations).
- Rotman, Appendix II (equivalence relations), Appendix III (functions).

In this section we cover

- Sets, elements, subsets
- Products of sets
- Functions, graphs
- Injective, surjective, bijective
- Images and preimages
- Composition
- Equivalence relations
- Divisibility

Sets

A *set* is a collection of elements. The elements can be anything, including other sets. The order in which the elements are listed is not important. Nor do repetitions count. We use curly brackets for sets.

The empty set (the set with no elements) is written $\varnothing$ or $\{\}$.

Examples

- $\{1, 2, 3\} = \{3, 1, 2\} = \{2, 2, 1, 3\}$
- $\{a, b\} = \{c, d\}$ iff ($a = c$ and $b = d$) or ($a = d$ and $b = c$)
- $\{a, b\} = \{c\}$ iff $a = b = c$
- $\mathbb{N}_+ = \{1, 2, 3, 4, 5, \dots\}$
- $\mathbb{N}_0 = \{0, 1, 2, 3, 4, 5, \dots\}$
- $\mathbb{Z}, \mathbb{Q}, \mathbb{R}, \mathbb{C}$
- The quaternions $\mathcal{H}$

We can also define sets by giving a domain A and a condition $P(x)$ in the form

$$\{x \in A : P(x)\},$$

where $P(x)$ is a proposition about x, that is, a function of x that takes the values "true" or "false".

Examples

- $\{x \in \mathbb{R} : x^{10} > 100\}$
- $\{n \in \mathbb{N}_+ : n^2 < -2\} = \varnothing$
- $\{n \geq 2 : (q > 0 \ \& \ q|n) \Rightarrow (q = 1 \text{ or } q = n)\}$ (the prime numbers)

We write

$$x \in M$$

to mean "x is an element of A".

Examples

- $1 \in \{1, 2\}$
- $\pi \notin \mathbb{Q}$.

We write

$$A \subseteq B$$

to mean "A is a subset of B", that is,

$$A \subseteq B \qquad \Longleftrightarrow \qquad (\forall a : a \in A \implies a \in B) \tag{3.1}$$

Note that $\forall$ means "for all" and $\exists$ means "there exists".

Examples

- $\{1, 2\} \subseteq \{1, 2, 3\}$
- $\{(a, b, c) : a, b, c \in \mathbb{N}_+, a^{17} + b^{17} = c^{17}\} \subseteq \varnothing$.

Let $n \geq 0$. An n-tuple is an ordered list of mathematical objects with n entries. Differently from a set, the order matters. Repetitions are allowed and they matter. We use parentheses to indicate an n-tuple. A 2-tuple is also called an *ordered pair*.

Examples

- $(2, 3, 5, 3)$ 4-tuple
- $(5, \sin(x), \{3, 5\})$ 3-tuple
- $()$ 0-tuple
- $(2, 3) \neq (3, 2)$

Definition 3.1 Let A and B be sets. The *Cartesian product* of A and B is the set of ordered pairs (a, b) with $a \in A$ and $b \in B$, i.e.

$$A \times B \coloneqq \{(a, b) \mid a \in A \text{ and } b \in B\} \tag{3.2}$$

Similarly, we define

$$A_1 \times \ldots \times A_n$$

as the set of all n-tuples $(x_1, \ldots, x_n\}$, where $x_i \in A_i$, $i = 1, \ldots, n$. We write A^n for

$$\underbrace{A \times \ldots \times A}_{n \text{ times}}.$$

We identify

$$A \times B \times C$$

with

$$(A \times B) \times C,$$

etc. This involves dropping some internal parentheses.

In particular, the set of all n-tuples of real numbers is written $\mathbb{R}^n$. In this lecture, we'll mostly be interested in $\mathbb{R}^2$ and $\mathbb{R}^3$.

Functions

Definition 3.2 A function f from a set X to a set Y is a rule that assigns to each x in X exactly one element y in Y. We write it as follows:

$$\begin{aligned} f \colon X &\to Y \\ x &\mapsto y = f(x) \end{aligned}$$

We also write

$$X \xrightarrow{f} Y.$$

If X is finite, a function can be defined by a table. The *graph* of f is the set of all ordered pairs $(x, f(x))$ such that $x \in X$:

$$(x, y) \in \text{graph}(f) \qquad \Longleftrightarrow \qquad y = f(x).$$

The graph of f is a subset of $X \times Y$.

The set X is called the *domain* of f, written $\text{dom}(f)$. The set Y is called the *target space*, written $\text{target}(f)$. The *image* of f is the set

$$\text{im}(f) \coloneqq \{f(x) \big| x \in X\}.$$

More generally, the image of a subset $A \subseteq X$ is defined by

$$f(A) \coloneqq \{f(x) \big| x \in A\}.$$

f is called *surjective* if

$$\text{im}(f) = Y,$$

that is, every $y \in Y$ gets hit by some $x \in X$. In symbols:

$$f \text{ is surjective} \iff \forall y \in Y \; \exists x \in X \; f(x) = y. \tag{3.3}$$

Figure 3.1: Surjective

Note that the definition of surjective depends on our choice of the target space Y. Therefore, strictly speaking, the definition of a function must include a specification of its target space, and two functions are not equal unless they have the same target space. Usually, but not always, we can overlook this.

Question Is the function $f : \mathbb{R} \to \mathbb{R}$, $x \mapsto x^2$ surjective? What about $g : \mathbb{R} \to [0, \infty)$, $x \mapsto x^2$? Is f "equal" to g?

The *preimage* of an element $y \in Y$ is the subset

$$\{x \in X : f(x) = y\}$$

of X. We write

$$f^{-1}(a)$$

for it. If $f^{-1}(y)$ consists of a single point x, we sometimes use $f^{-1}(y)$ to mean the element x rather than the set $\{x\}$.

Similarly, the preimage of a subset $B \subseteq Y$ is the subset

$$f^{-1}(B) := \{x \in X : f(x) \in B\}$$

of X.

The function f is called *injective* if for each $y \in Y$, $f^{-1}(y)$ consists of at most one element. That is, y gets hit by at most one element of X. In symbols:

$$f \text{ is injective} \iff (\forall x, x' \in X : f(x) = f(x') \implies x = x') \tag{3.4}$$

Figure 3.2: Injective

A function f is called *bijective* if f is both injective and surjective. Such a function is also said to be a *one-to-one correspondence*.

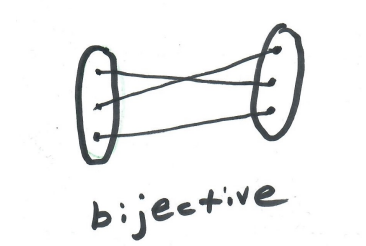

Figure 3.3: Bijective

If f is bijective, there exists a function (the inverse of f)

$$f^{-1} : Y \to X$$

which takes each element $y \in Y$ to its (unique) preimage x in X.

Let $f : X \to Y$ be a function. If $A \subseteq X$, we define the *restriction*

$$f|A : A \to Y,$$

of f to A by

$$(f|A)(x) := f(x) \qquad \text{for all } x \in A.$$

If $A \neq X$, then $f|A$ has a different domain from f, so it's a different function.

Occasionally (!) we might want to explicitly redefine the target space as well. If $B \supseteq \mathrm{im}(f)$, we define the *co-restriction*

$$f \restriction B : X \to B,$$

of f to A by

$$(f \restriction B)(x) := f(x) \qquad \text{for all } x \in X.$$

It's the same "rule", but it has a different target space, so it's a different function. Officially, $f \restriction B \neq f$ unless $B = Y$. Usually this will not matter.

Let X, Y and Z be sets. Let

$$f\colon X \to Y, \qquad g\colon Y \to Z$$

be functions. The function

$$\begin{aligned} g \circ f\colon X &\to Z, \\ x &\mapsto g(f(x)), \end{aligned}$$

that assigns $g(f(x))$ to x is called the *composition* of f and g. To make the order precise, we say "f followed by g". In symbols

$$(g \circ f)(x) := g(f(x)).$$

We can also write the *commutative diagram*

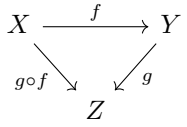

Figure 3.4: Commutative diagram

Equivalence relations

A *relation* P is a function $P(x, y)$ with two arguments and values in {true, false}. Usually it is written with the relation sign in the middle. So

$$xPy$$

means x has the relation P to y. An example of a relation is

$$x \text{ is a sister of } y.$$

An *equivalence relation* is a relation $\cong$ such that:

$$\begin{array}{lcll} x \cong x & & & \text{(reflexive)} \\ x \cong y & \iff & y \cong x & \text{(symmetric)} \\ x \cong y, y \cong z & \implies & x \cong z & \text{(transitive).} \end{array}$$

An example of an equivalence relation is

$$x \text{ and } y \text{ have the same parents.}$$

A *partition* or *decomposition* of a set X is a subdivision of X into disjoint subsets $(A_i)_{i\in I}$ whose union is X:

$$X = \bigcup\{A_i \,\big|\, i \in I\}, \qquad A_i \cap A_j = \varnothing \quad \text{for } i \neq j \in I.$$

The main fact about equivalence relations is that they induce a partition of the set on which they are defined, characterized by the condition

$$x \cong y \quad \Longleftrightarrow \quad x \text{ and } y \text{ lie in the same element } A_i \text{ of the partition.}$$

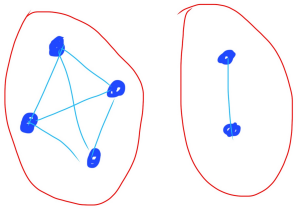

Figure 3.5: Equivalence relation

See Saracino, pp 80-82, Rotman, Appendix III for details.

Divisibility

Definition 3.3 Let $a, b \in \mathbb{Z}$. We say that a divides b if b is an integer multiple of a:

$$a \big| b \qquad \Longleftrightarrow \qquad \exists\, k \in \mathbb{Z} : b = ka \tag{3.5}$$

§4 Complex numbers

References

- L. Ahlfors, pp. 1-11, 76-88
-

In this section we cover

- Complex numbers
- Complex conjugate, norm
- Exponential function
- Polar coordinates

Complex numbers

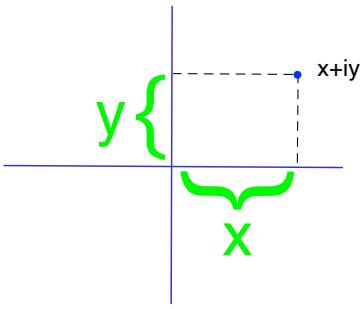

Figure 4.1: Complex numbers

The complex number system $\mathbb{C}$ is $\mathbb{R}^2$ equipped with addition and a nonobvious multiplication. Let

$$(x, y) \in \mathbb{R}^2.$$

Define

$$(x, y) + (x', y') := (x + x', y + y'), \qquad (x, y) \cdot (x', y') := (xx' - yy', xy' + x'y).$$

If we define

$$1 := (1, 0), \qquad i := (0, 1), \qquad t(x, y) := (tx, ty) \quad \text{for } t \in \mathbb{R},$$

then we have effectively identified $\mathbb{R}$ with the x-axis via

$$t \mapsto t \cdot 1 = (t, 0).$$

Furthermore

$$z = (x, y)$$

may be rewritten as

$$z = x1 + yi = x + iy.$$

Then

$$(x+iy)+(x'+iy') = x+x'+i(y+y'), \qquad (x+iy)(x'+iy') = xx'-yy'+i(xy'+x'y).$$

These rules are consistent with (can be deduced from) the rule

$$i^2 = -1$$

together with the distributive law. Indeed, the complex numbers satisfy all the usual laws of algebra (field axioms), including the existence of inverses. For $z = x + iy \neq 0$, the multiplicative inverse is given by

$$z^{-1} = (x + iy)^{-1} = \frac{x - iy}{x^2 + y^2},$$

as may easily be verified.

Complex conjugation

If

$$z = x + iy,$$

we call x the *real part* of z and y the *imaginary part*, and define

$$\operatorname{Re} z := x, \qquad \operatorname{Im} z := y.$$

We define the complex conjugate of z by

$$\bar{z} := x - iy.$$

The complex conjugate operation preserves all algebraic operations; it is an algebraic isomorphism (field isomorphism) $\mathbb{C} \to \mathbb{C}$. In particular,

$$\overline{z + w} = \bar{z} + \bar{w}, \qquad \overline{zw} = \bar{z}\bar{w}, \qquad \overline{z^{-1}} = \bar{z}^{-1},$$

as may easily be verified.

Norm

The *absolute value*, *norm*, or *length* of z is defined by

$$|z| := \sqrt{x^2 + y^2}, \qquad z = x + iy \in \mathbb{C}.$$

The reader may verify that

$$|z|^2 = z\bar{z}, \qquad z \in \mathbb{C},$$

so

$$|z| = \sqrt{z\bar{z}}, \qquad z \in \mathbb{C}.$$

There is also the multiplicative property

$$|zw| = |z||w|, \qquad z, w \in \mathbb{C}.$$

and the inverse formula

$$\frac{1}{z} = \frac{\bar{z}}{|z|^2}$$

The *unit circle* is defined by

$$S^1 := \{z \in \mathbb{C} : |z| = 1\}.$$

Exponential function

The exponential function is best defined as the limit of the convergent sequence

$$e^z = \exp(z) := \sum_{n=0}^{\infty} \frac{z^n}{n!}, \qquad z \in \mathbb{C},$$

where

$$n! := n(n-1)(n-2) \cdot \cdots \cdot 2 \cdot 1, \qquad 0! := 0,$$

is the factorial function. But even without knowing how this works, we can characterize e^z by its properties in such a way that we can work with it. We have

$$e^{z+w} = e^z e^w, \qquad z, w \in C. \tag{4.1}$$

$$e^{i\theta} = \cos\theta + i\sin\theta, \quad \theta \in \mathbb{R}, \qquad e^{2\pi i} = 1. \tag{4.2}$$

So

$$\theta \mapsto e^{i\theta}, \qquad \theta \in \mathbb{R},$$

is surjective and traces S^1 infinitely many times in both directions. In other words, for every

$$w \in S^1$$

there is t such that $e^{it} = w$, and all the numbers

$$\ldots t - 2\pi,\ t,\ t + 2\pi, \ldots$$

map to w under the exponential map.

This picture shows a rectangular grid in $\mathbb{C}$ and its image under the exponential map.

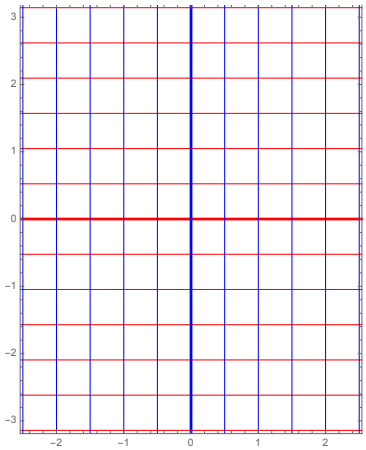

Figure 4.2: Grid (Mathematica)

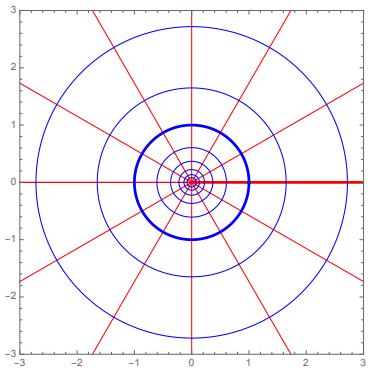

Figure 4.3: Exponential image of grid (Mathematica)

Exercise 4.1 *Fix $z \in \mathbb{C}$. Verify that $dx^z/dx = zx^{z-1}$, where x is a real variable.*

Polar coordinates

Fix $z \in \mathbb{C}$. Set

$$r := |z|.$$

Then

$$\frac{z}{|z|} = e^{i\theta}$$

for some $\theta \in \mathbb{R}$. θ is only well-defined up to adding a multiple of 2π. Typically we require $0 \leq \theta < 2\pi$, which makes θ unique, but not continuous as a function of z. So we have

$$z = re^{i\theta}.$$

This is called the *polar representation* of z. r is called the *magnitude* and θ is called the *argument* of z.

Multiplication is rather easy in the polar representation: if

$$z = re^{i\phi}, \qquad w = se^{i\psi},$$

then

$$zw = rse^{i(\phi+\psi)}.$$

2

Introduction

§5 First off

Here's what the course is about:

I Metric spaces
II The sphere
III The hyperbolic plane

It's different from last year.

We'll start with hyperbolic. Actually, we'll start with three spaces.

S^2	sphere	compact(finite)	positive curvature
$\mathbb{R}^2$	Euclidean space	infinite	zero curvature (flat)
$\mathbb{H}^2$	hyperbolic space	infinite	negative curvature

Here is the 2-sphere:

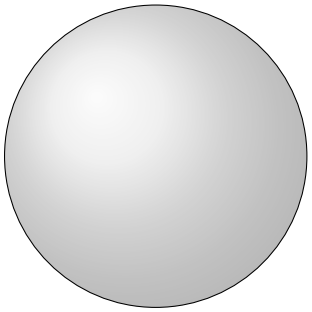

Figure 5.1: The 2-sphere

We consider the 2-sphere as a world in itself. That is, we take the point of view of an ant that lives on the surface of the sphere, and wanders around. He can't see off the sphere. Even his light rays travel along the surface of the sphere.

He experiences the geometry of the surface by walking. So, for him, the little ant scientist, distance is the distance he walks. The shortest distance between two points is a geodesic arc.

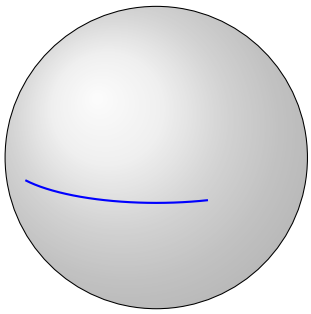

Figure 5.2: A geodesic arc

Wherever he goes on the sphere, it looks the same. We call this *homogeneous*. Also, whichever direction he looks, it looks the same. We call this *isotopic*. So the 2-sphere is homogeneous and isotropic.

Here is the Euclidean plane:

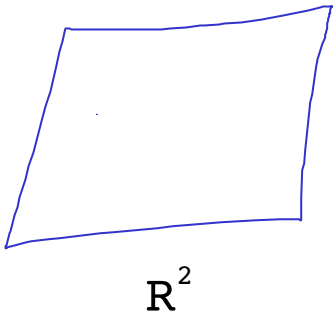

Figure 5.3: The Euclidean plane

The plane is also homogeneous and isotropic.

A space is called *simply-connected* if every loop can be contracted to a point within the space. The above three spaces are simply connected, whereas the surface of a torus is not.

It turns out that (up to scale) S^2, $\mathbb{R}^2$ and $\mathbb{H}^2$ are the only simply-connected, homogeneous, isotropic spaces in dimension 2.[1] to a sphere They are called *2-dimensional space forms*. They are the

But what is hyperbolic space? That is harder to define, and will be a major topic of the class. Here is a picture to give you an idea.

[1] We say "up to scale" because you can always multiply distances by a constant. This leads to a different space, but it's just a rescaling of the old space.

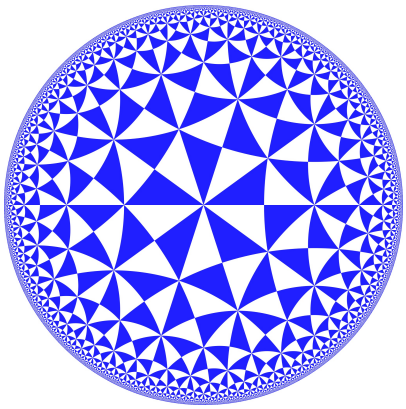

Figure 5.4: Order-4 bisected pentagonal tiling of the hyperbolic plane (Rocchini, Wikipedia)

The blue triangles form a tessellation, or tiling, of the hyperbolic plane. They are there to give you an idea of the geometry.[2]

The true distances on the hyperbolic plane are not as they appear. In fact, by declaration, all the triangles are the same size. Also, the sides of the triangles are "straight lines" for the inhabitants. That is, in the local geometry, they are the shortest distance between two points.

Notice that as one goes to the edge of the disk, there are more and more triangles. This shows that the distance to the edge is really infinite. For the inhabitants, there is no edge; their world goes on forever.

It also suggests another property of the hyperbolic plane: there is a huge amount of area out towards infinity. It turns out that

1) The area of a disk grows roughly exponentially as a function of radius.

To be precise,

$$A(r) \sim Ce^{cr} \qquad \text{for large } r.$$

So area grows much faster than it does in the Euclidean plane, where $A(r) = \pi r^2$.

We will discuss this later in detail. Now I'm just giving an idea.

Hyperbolic space has many other strange features. For example,

2) To an inhabitant, objects of a given size at a given distance appear far smaller in hyperbolic space than they do in Euclidean space.
3) Bodies moving in a straight line experience internal tidal effects, in contract to Euclidean space.

Here is something very odd. Despite the huge size of hyperbolic space:

4) There is a universal upper bound to the area of a triangle.

[2]See *Uniform tilings in hyperbolic plane*, Wikipedia.

To be precise, Very strange.

There is a hyperbolic space $\mathbb{H}^n$ in every dimension. Here is a screenshot from J. Weeks' Curved Spaces app:

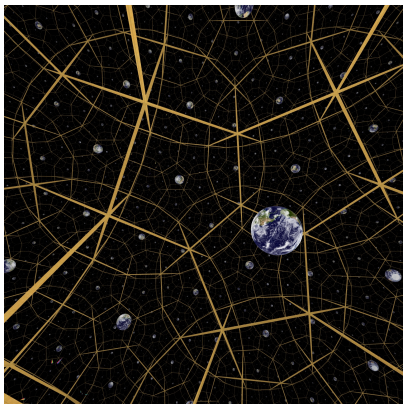

Figure 5.5: A tessellated hyperbolic space (J. Weeks' Curved Spaces app)

Let's fly around in hyperbolic space.

The following Curved Spaces app is by J. Weeks. There are various hyperbolic tessellations you can view.

- `http://www.geometrygames.org/CurvedSpaces/index.html`

On the net, there are thousands of graphics, videos and blogs on hyperbolic space. I found dozens on youtube alone. It's everybody's favorite subject. B. Loustau wrote:[3]

> *Hyperbolic geometry... is the star of geometries, and geometry is the star of mathematics!*

[3]Loustau, p. 4.

§6 Euclidean space

In this section we cover

- $\mathbb{R}^n$ with the Euclidean metric
- Norms, inner products

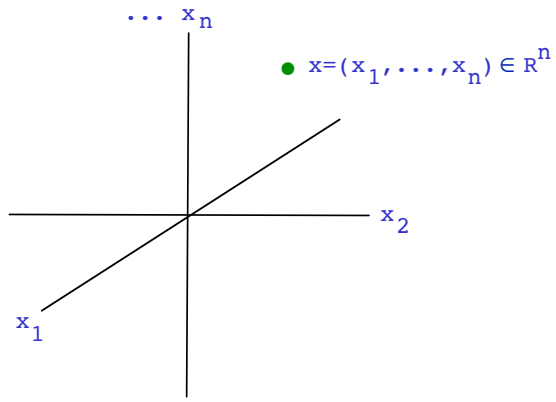

Figure 6.1: Point x in $\mathbb{R}^n$

By $\mathbb{R}^n$ we mean the set of ordered n-tuples

$$x = (x_1, \ldots, x_n),$$

of real numbers $x_1, \ldots, x_n$. The numbers x_i vary freely in $\mathbb{R}$. The formula

$$x \in \mathbb{R}^n$$

reads

$$x \text{ is an element of } \mathbb{R}^n.$$

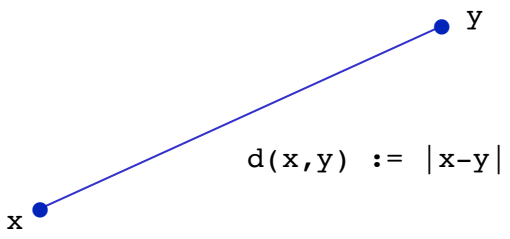

Figure 6.2: Distance between x and y

Definition 6.1 Let $x = (x_1, x_2, \ldots, x_n)$ and $y = (y_1, y_2, \ldots, y_n)$ be two points in $\mathbb{R}^n$. The *distance* $d(x, y)$ between x and y is given by

$$d(x, y) \coloneqq \sqrt{(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2}.$$

The function d is called the *Euclidean metric* on $\mathbb{R}^n$.

Inner product and norm

We can relate the distance in $\mathbb{R}^n$ to the inner product and norm as follows:

Definition 6.2 Let $x = (x_1, x_2, \ldots, x_n)$ and $y = (y_1, y_2, \ldots, y_n)$.

a) The *inner product* of x and y is defined by

$$x \cdot y := \sum_i x_i y_i.$$

b) The *norm* of the vector x is defined by

$$|x| := \sqrt{\sum_i x_i^2} = \sqrt{x \cdot x}.$$

Then

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

3

Metric spaces

§7 Definition of metric spaces

References

- Ahlfors, pp. 51-54.

In this section we cover

- Metric spaces
- A three-point example
- $\mathbb{R}^n$ example
- Sierpinski gasket

The definition

The idea of a metric space is to enthrone a notion of *distance* in a fully abstract setting. Distance is the essence of geometry.

Let X be a set.

Definition 7.1 A function

$$d\colon X \times X \to \mathbb{R}$$

is a *metric* on X if for all $x, y, z \in X$ we have

i)	$d(x,y) \geq 0$	(positivity)
ii)	$d(x,y) = 0 \iff x = y$	(definiteness)
iii)	$d(x,y) = d(y,x)$	(symmetry)
iv)	$d(x,z) \leq d(x,y) + d(y,z)$	(triangle inequality)

The pair (X, d) is called a *metric space*. We often use X as an abbreviation for (X, d).

The triangle inequality means that it shouldn't take longer to go from x to z than it takes to go from x to y, then y to z.

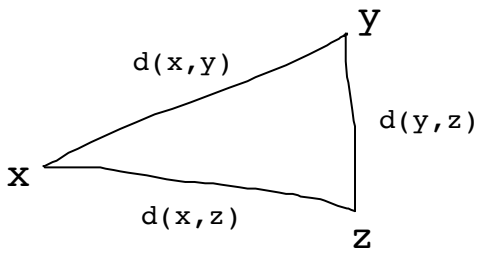

Figure 7.1: Triangle inequality

Notice that metric spaces don't have angles, lengths, or areas – at least not at first. We would have to work hard to define usable versions of these concepts in a general metric space, if it works at all. But we won't do that in this course.

A three-point metric space

This metric space has 3 points.

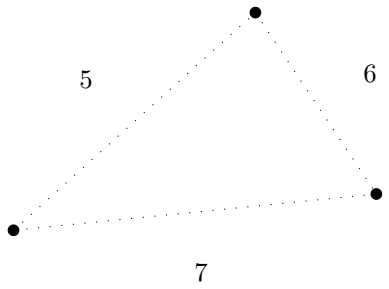

Figure 7.2: A three-point metric space

$\mathbf{R^n}$

A fundamental example of a metric space is $\mathbb{R}^n$ with the Euclidean metric d.

Positive definiteness and symmetry of d are obvious. These are i)-iii).

To prove iv) the triangle inequality for d, the reader must establish the following basic propostion of analysis:

Proposition 7.2 *Let $x, y, z \in \mathbb{R}^n$. Then*

a) (Cauchy-Schwarz inequality) $|x \cdot y| \leq |x||y|$
b) (Triangle inequality for norms) $|x + y| \leq |x| + |y|$
c) (Triangle inequality for distances) $|x - z| \leq |x - y| + |y - z|$.

Proof 1. Step a) is tricky. For any $t \in \mathbb{R}$, we have

$$|x + ty|^2 = |x|^2 + 2tx \cdot y + t^2|y|^2.$$

Since this is nonnegative for all real values of t, there can be at most one real root t. So the discriminant $b^2 - 4ac$ must be nonpositive. Here $a = |y|^2$, $b = 2x \cdot y$, $c = |x|^2$. So

$$\begin{aligned} 0 &\geq b^2 - 4ac \\ &= (2x \cdot y)^2 - 4|y|^2|x|^2, \end{aligned}$$

which becomes

$$|x \cdot y| \leq |y||x|,$$

which is a).

2. To prove b), compute

$$\begin{aligned} |x + y|^2 &= (x + y) \cdot (x + y) \\ &= |x|^2 + 2x \cdot y + |y|^2 \\ &\leq |x|^2 + 2|x||y| + |y|^2 \qquad \text{by a)} \\ &= (|x| + |y|)^2, \end{aligned}$$

from which b) follows by taking the square root.

3. We obtain c) from b) by substituting $x \to x - y$, $y \to y - z$ in b). Because b) and c) are so closely related, they are both called the triangle inequality.

□

Sierpinski gasket

The *Sierpinski gasket* G is the limit of the following construction as the number of "levels" goes to infinity.

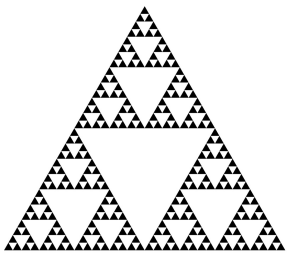

Figure 7.3: The Sierpinski gasket

Use the metric of $\mathbb{R}^2$ to give it a metric. It is a fractal. That means, it has a non-integer dimension. Its dimension[1] is

$$\frac{\log(3)}{\log(2)} \approx 1.5849625.$$

This says that it has a higher dimension than a curve (dimension one), but is "thinner" than $\mathbb{R}^2$.[2]

[1] Hausdorff dimension, for the experts

[2] We don't claim to have defined dimension. It is defined in the subject of *Geometric Measure Theory.*

§8 Isometries

In this section we cover

- Isometries of metric spaces
- Example
- Isometries of $\mathbb{R}^n$

Isometries

Let X, Y be metric spaces.

Definition 8.1 An *isometry* from (X, d_X) to (Y, d_Y) is a bijection

$$f : X \to Y, \qquad x \mapsto f(x),$$

that preserves distances between points:

$$d_X(x, y) = d_Y(f(x), f(y)), \qquad x, y \in X. \tag{8.1}$$

Here, x is a point of X, and $f(x)$ is its image under the mapping f.

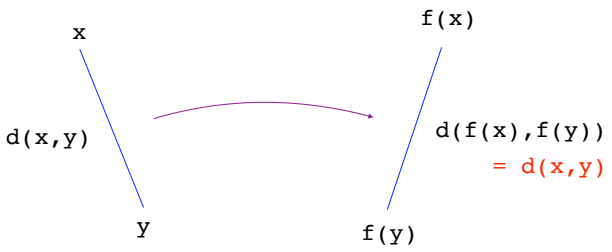

Figure 8.1: Distance is preserved

It means the metric spaces look the same, for all practical purposes.

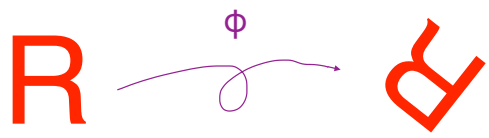

Figure 8.2: An isometry of an **R** to another **R**

Note: The injective part of "bijective" follows automatically from the distance-preserving property. So only the surjectivity must be checked.[3]

Define

$$\mathrm{Isom}(X, Y) := \{f : X \to Y | f \text{ is an isometry})$$

A *self-isometry* (usually just called an isometry) of X is an isometry from X to X. Write

$$\mathrm{Isom}(X) := \mathrm{Isom}(X, X).$$

Exercise 8.1 *Show that the set of self-isometries has the following 3 properties:*

a) $\mathrm{id}_X \in Isom(X)$,
b) $f, g \in Isom(X)$ *implies* $f \circ g \in Isom(X)$,
b) $f \in Isom(X)$ *implies* $f^{-1} \in Isom(X)$.

Isometries of the Sierpinski gasket to itself

Consider the Sierpinski gasket again:

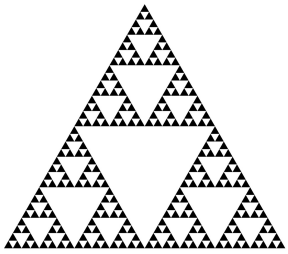

Figure 8.3: The Sierpinski gasket

Exercise 8.2 *How many self-isometries does it have?*

Isometries of $\mathbb{R}^n$ to itself

A self-isometry of $\mathbb{R}^n$ is often called a *rigid motion*.

Distance-preserving maps from $\mathbb{R}^n$ to $\mathbb{R}^n$ are automatically surjective, and therefore bijective. This is not obvious, but requires a linear algebra proof.

[3] *Surjective* means that every point of $\mathbb{R}^n$ is hit by at least one point x under ϕ. *Injective* means that every point of $\mathbb{R}^n$ is hit by at most one point x under ϕ. *Bijective* means injective and surjective. A bijective function is a one-to-one correspondence.

Isometries of $\mathbb{R}^n$ also preserve angles, areas, and volumes. This requires a proof.[4]

Here are some isometries of $\mathbb{R}^n$:

- The identity map $\mathrm{id}_{\mathbb{R}^n} : \mathbb{R}^n \to \mathbb{R}^n, \quad x \mapsto x$
- Translations, rotations
- Reflections in planes, lines, or points.

Then there are some exotic ones:

- Roto-reflections, glide reflections, screw motions.

These rigid motions are amply explained in Geometrie 2020, `https://metaphor.ethz.ch/x/2020/hs/401-1511-00L/literatur/script.pdf`, in many places: §§2,12,14,15,16,17,32,48 and others.

The following kinds of maps are generally not isometries:

> expansions, contractions, shears, projections, distortions, rips, tears, constant maps.

$\mathrm{Isom}(\mathbb{R}^n)$ is called the *Euclidean group*.

[4]For general metric spaces, such notions are more subtle or don't exist. But when they exist, they are preserved.

§9 Examples of metric spaces

References

- Wikipedia, *Taxicab geometry.*
- Falconer, *The geometry of fractal sets.*

In this section we cover

- L^1 metric on $\mathbb{R}^n$
- sup metric on $\mathbb{R}^n$
- Infinite-dimensional vector spaces, function spaces

L^1 metric on $\mathbb{R}^n$

Define on $\mathbb{R}^n$ the distance function

$$d_1(x, y) := \sum_i |x_i - y_i|, \qquad x, y \in \mathbb{R}^n.$$

This is called the L^1 *metric.*

Exercise 9.1 *Prove that this is a metric.*

The L^1 metric is often called the *taxicab metric*, because if $\mathbb{R}^2$ gets a New-York style grid of streets parallel to the coordinate axes, it would be the distance a taxi has to drive to get from x to y.

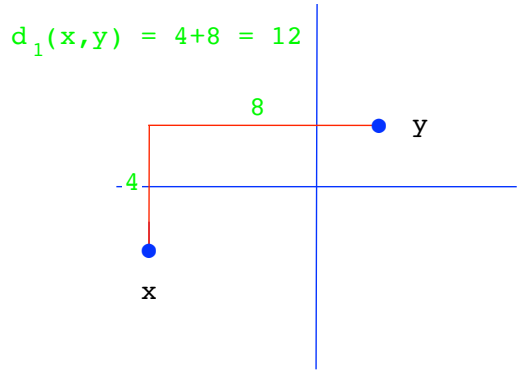

Figure 9.1: Taxi metric

Indeed, notice that there are often many alternate taxi routes from x to y of minimum length:

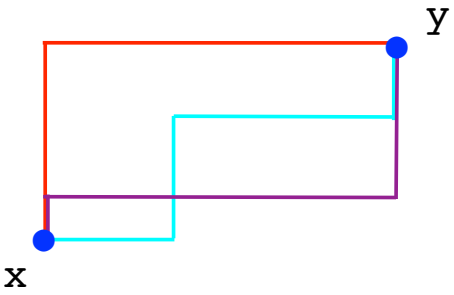

Figure 9.2: Many pathways of the same length

This contrasts starkly with the situation in Euclidean space, where there is only one shortest path between two points.

Now, in $(\mathbb{R}^2, d_1)$, if x, y differ by only one coordinate, then they *do* have a shortest path between them, whereas if they differ in both coordinates, they have infinitely many shortest paths between them.

Indeed, define

$$R(x,y) := \{z \in \mathbb{R}^2 : d(x,z) + d(z,y) = d(x,y)\},$$

that is, the triangle inequality is *saturated*. If x, y differ by only one coordinate, then $R(x,y)$ is a line, whereas if they differ in both coordinates, then $R(x,y)$ is a rectangle.

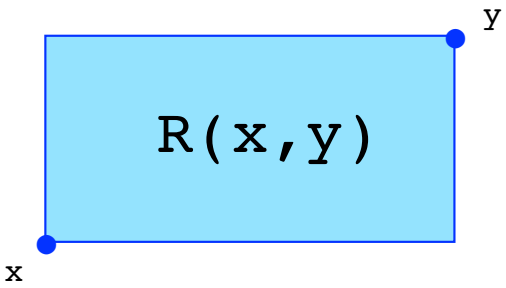

Figure 9.3: Set of points that saturate the triangle inequality

So $(\mathbb{R}^n, d_1)$ does not look the same in all directions. The directions given by the coordinate axes are special.

Exercise 9.2 *What is the set* $\{x : d_1(x,0) \leq 1\}$*?*
(This is called the *unit ball* of $(\mathbb{R}^n, d_1)$.)

Exercise 9.3 *What are the isometries of $(\mathbb{R}^n, d_1)$? Are there less of them than with the Euclidean metric?*

Sup metric on $\mathbb{R}^n$

Define on $\mathbb{R}^n$ the distance function

$$d_\infty(x, y) := \max_i |x_i - y_i|, \qquad x, y \in \mathbb{R}^n.$$

This is called the *sup metric* or L^∞ *metric*.

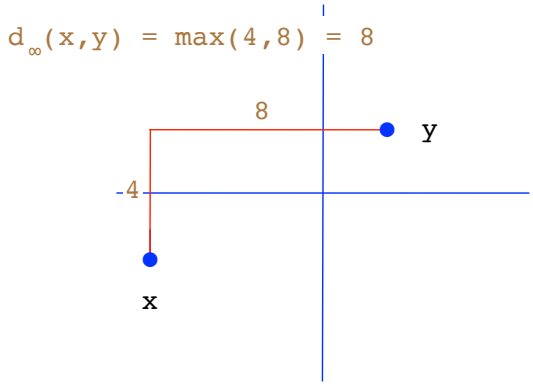

Figure 9.4: Sup metric

Exercise 9.4 *We can ask the same questions for d_∞ as we did for d_1.*

Exercise 9.5 *Prove that $(\mathbb{R}^2, d_1)$ is isometric to $(\mathbb{R}^2, 2d_\infty)$.*

Function spaces

We can make infinite infinite-dimensional vector spaces into metric spaces, called *Banach spaces*. Very often these are spaces of functions, called *function spaces*.

This can be done in many ways, and it is very subtle. But it is essential in Analysis and Partial Differential Equations. The subject is called *Functional Analysis*.

§10 More examples of metric spaces

In this section we cover

- Metric subspaces
- Path-length metrics
- S^2 with the path-length metric
- Sierpinski gasket with the path-length metric
- Koch snowflake
- $\sqrt{\ }$ metric
- The infinite-dimensional simplex

Metric subspaces

Any subset of a metric space becomes a metric space in a natural way.

Indeed, let (X, d_X) be a metric space. Let $Y \subseteq X$ be any subset. Then Y inherits a metric space structure, called the *subspace metric*, defined by

$$d_Y := d_X|(Y \times Y).$$

That is, we just use the same distances in Y that we were already using in X. It is trivial to verify that

$$(Y, d_Y)$$

is a metric space. (Y, d_Y) is called a *metric subspace* of (X, d_X).

Example: The Sierpinski gasket is a metric subspace of $\mathbb{R}^2$. But then so is any other subset.

Path-length metrics

For a subset Y of $\mathbb{R}^n$, we can define another metric d'_Y on Y called *geodesic distance* or the *path-length metric*.

Let γ be any path in Y. If γ is not too irregular, we can define the *length* $L(\gamma)$ of γ.

Let $x, y \in Y$. Define $d'_Y(x, y)$ to be the length of the shortest path between x and y that stays within Y. If there is no shortest path, then use the infimum instead:

$$d'_Y(x, y) := \inf\{L(\gamma) : \gamma \text{ is a path in } Y \text{ connecting } x \text{ to } y.\}$$

Observe:

1) d'_Y is not always a metric, because it may be impossible to connect x and y by a finite-length path.

2) *Exercise:* Prove that if every x and y in Y can be connected by a finite-length path, then d'_Y is a metric on Y.

3) $d'_Y \geq d_Y$ always.

4) Actually, $L(\gamma)$ can be defined for any continuous path in any metric space, though it might be infinite. So d'_Y can be defined for any subset Y of any metric space X, and is a metric if every two points of Y can be connected by a finite-length path lying in Y. But we won't do that here.

Sometimes d_Y is called the *extrinsic distance* and d'_Y is called the *intrinsic distance*. So Y gets two different induced metrics.

Path-length metric on the Sierpinski gasket

Recall the Sierpinski gasket. Call it G. Use the path-length metric d'.

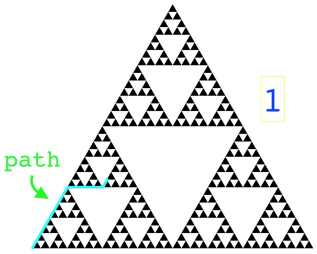

Figure 10.1: The Sierpinski gasket

Exercise 10.1 *Suppose a side of the containing triangle has length 1.*

a) What is the average distance from a randomly chosen point of G to a corner point?

b) What is the average distance between two randomly chosen points of G?*

Here is a hint on how to understand the probability. Let x be a random point. Require

1) For any closed subsets A, B of G that are isometric, we have

$$\text{Prob}(x \in A) = \text{Prob}(x \in B).$$

2) The probability of hitting a particular point is zero.

Koch snowflake

Here is another fractal. The limit of the following process is called the Koch snowflake S.

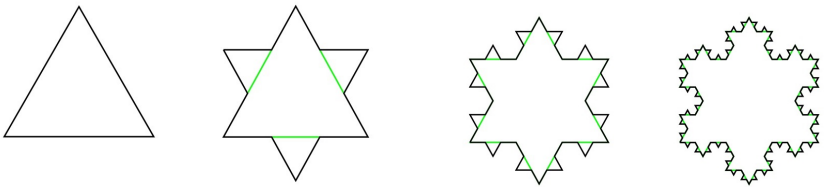

Figure 10.2: Koch snowflake (Wxs, Wikipedia)

The dimension[5] of the Koch snowflake is

$$\frac{\log(4)}{\log(3)} \approx 1.261860$$

It is a bit "thinner" than the Sierpinski gasket.

Exercise 10.2 *What is the path-length metric on S?*

For more about fractals, see Falconer, *The geometry of fractal sets.*

$\mathbb{R}$ with $\sqrt{\ }$ metric

Another way to define a fractal metric space is by modifying the Euclidean metric in place. Define a new metric $d_{\sqrt{\ }}$ on $\mathbb{R}$ by

$$d_{\sqrt{\ }}(x, y) = \sqrt{|x - y|}, \qquad x, y \in \mathbb{R}.$$

This is easily seen to be a metric. In particular, the triangle inequality follows from the well-known inequality

$$\sqrt{a + b} \leq \sqrt{a} + \sqrt{b}, \qquad a, b \geq 0. \tag{10.1}$$

Exercise 10.3 *Prove (10.1).*

Let's call this metric space $(\mathbb{R}, d)$ the $\sqrt{\ }$-space.

An *isometric embedding* is a distance-preserving function. It must be injective; it need not be surjective. It is the same as an isometry onto a metric subspace of the target space.

[5] Hausdorff dimension, for the experts

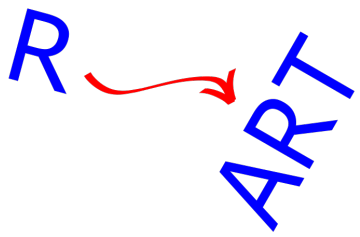

Figure 10.3: Isometric embedding

Exercise 10.4 *Prove or disprove: The $\sqrt{}$ space cannot be isometrically embedded in $\mathbb{R}^n$ for any n.*

Exercise 10.5 *What are the isometries of $(R, d_{\sqrt{}})$?*

The infinite-dimensional simplex

Let Z be the set of all unit vectors of the form

$$e_i = (0, \ldots, 0, 1, 0, \ldots),$$

where the 1 occurs in the i'th place. Note that

$$d(x, y) = \sqrt{2}, \qquad x \neq y \in Z.$$

Z is sometimes called the infinite-dimensional simplex. It is a so-called *discrete space*, since no point has other points arbitrarily close to it.

Exercise 10.6 *Show that Z cannot be isometrically embedded in $\mathbb{R}^n$ for any n.*

Exercise 10.7 *Find a metric space with four points that cannot be isometrically embedded in $\mathbb{R}^n$ for any n.*

Exercise 10.8 *Find a metric on $\mathbb{R}^2$ so that every bijective map $f\colon \mathbb{R}^2 \to \mathbb{R}^2$ is an isometry.*

Part II

Spheres

4

Area and circumference of an intrinsic disk in S^2

§11 Length and distance in S^2

Length of paths in S^2

Let

$$S^2 := \{x \in \mathbb{R}^3 : |x| = 1\}$$

be the 2-sphere $\mathbb{R}^{n+1}$.

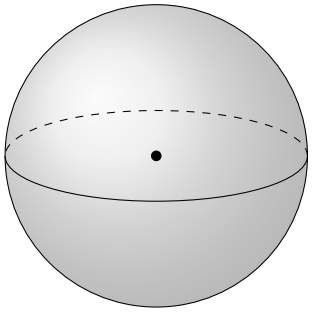

Figure 11.1: Two-sphere

Generally, for a continuously differentiable curve

$$\gamma : [a, b] \to \mathbb{R}^n,$$

we define its length by the integral

$$L(\gamma) := \int_a^b \left| \frac{d\gamma}{dt}(t) \right| dt$$

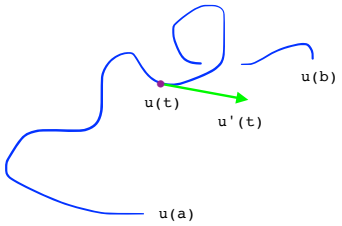

Figure 11.2: Path with velocity vector

The idea is that distance traveled is the integral of speed over time. The units check out:

$$x \sim \frac{x}{t} t.$$

The length of a path in S^2 is just its length in $\mathbb{R}^3$. So we write $L(\gamma)$ without a subscript.

For simple curves such as geodesic arcs, we can compute the length directly: it is just equal to the central angle subtended by the arc.

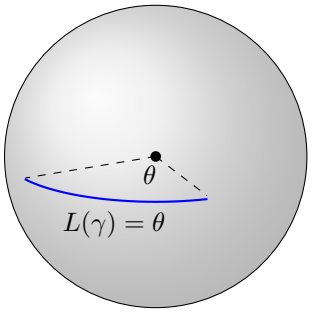

Figure 11.3: Central angle is length

S^2 with two metrics

Write $d' = d'_{S^2}$ for the path-length metric on S^2, and $d = d_{S^2}$ for the subspace metric.

Then d' is the length of the shorter geodesic arc γ connecting x and y, namely

$$d'(x, y) := L(\gamma).$$

On the other hand, d is the length of the chord connecting x and y, namely

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

See the picture.

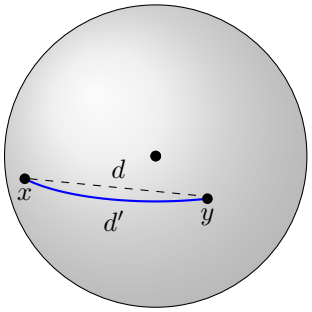

Figure 11.4: Arc versus chord

For example, if N is the north pole and S is the south pole, then

$$d(N,S) = 2, \qquad d'(N,S) = \pi.$$

We won't prove them at this point, but you can try your hand at the following two exercises. They are a little tricky.

Exercise 11.1 *In $\mathbb{R}^2$, a line is the shortest curve between two points.*

Exercise 11.2 *In (S^2, d'), a geodesic arc (of length $\leq \pi$) is the shortest curve between two points.*

In particular, the infimum is really a minimum in both cases.

Let us focus on d'. We exclude geodesic arcs of length $> \pi$ because they go the "long way around" and aren't the most efficient way to get from x to y.

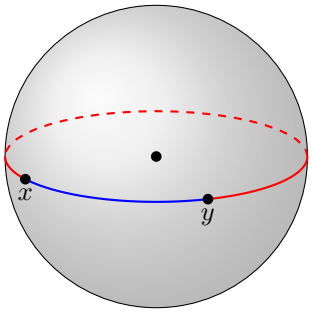

Figure 11.5: The red path is the long way around

Note that there can be many paths that realize the minimum. Indeed, if x and y are antipodal points, then $d'(x,y) = \pi$, and every semicircle from x to y has length π. The reader can compare this to the nonunique minimizing path in the taxicab metric, but the mechanisms seem to be quite different.

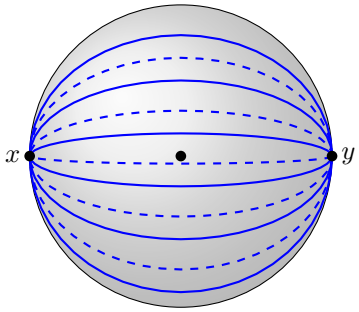

Figure 11.6: Many paths

Angle

A curve $\alpha(t)$ is called *regular* if it is continuously differentiable and its velocity vector $d\alpha/dt$ is nowhere vanishing.

The angle between two regular curves $\alpha(t)$, $\beta(t)$ in S^2 that meet at a point

$$\alpha(t_0) = \beta(t_0)$$

is the angle between their velocity vectors at the point of intersection.

$$\angle\left(\frac{d\alpha}{dt}(t_0), \frac{d\beta}{dt}(t_0)\right) \in [0, \pi].$$

§12 Area and circumference of an intrinsic disk

References

- J. R. Weeks, *The Shape of Space,* pp. 125-134.

We compute the area and circumference of an intrinsic disk in S^2, and reflect upon our result. We briefly mention the hyperbolic case.

Let N be the north pole of S^2. Let $D_r = D_r(N)$ be the intrinsic disk of radius r defined by

$$D_r := \{x \in S^2 : d'(x, N) < r\},$$

where d' is the geodesic distance on S^2. Let

$$K_r := \{x \in S^2 : d'(x, N) = r\}$$

be its boundary, the intrinsic circle of radius r with center N.

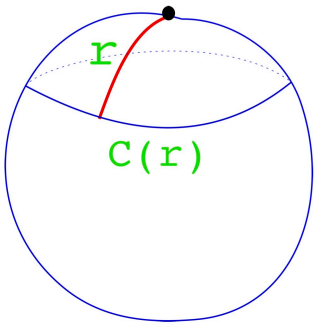

Figure 12.1: Circumference of a disk

Circumference

We can easily compute the circumference of K_r.

Let θ be the central angle subtended by N and any point x on K_r. Then

$$\theta = r$$

So the radius of C_r *as a circle in 3-space* is given by

$$\tilde{r} := \sin\theta = \sin r.$$

(see figure).

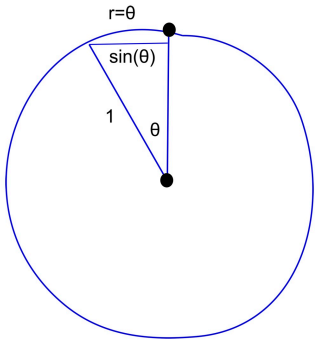

Figure 12.2: Euclidean radius of circle

So the circumference of K_r is $2\pi\tilde{r}$, which is

$$C_S(r) = 2\pi \sin(r), \qquad 0 \le r \le \pi.$$

Now let us reflect on this: $\sin(r)$ has the expansion

$$\sin(r) = r - \frac{r^3}{6} + O(r^5) \qquad \text{as } r \to 0,$$

where the big-O notation $O(g)$ refers to any function $\varepsilon(r)$ that satisfies

$$|\varepsilon(r)| \le Cg(r) \qquad \text{for small } r,$$

where C is some constant independent of r. So we have

$$C_S(r) = 2\pi r - \frac{\pi r^3}{3} + O(r^5) \qquad \text{as } r \to 0.$$

Compare this to the Euclidean formula

$$C_E(r) = 2\pi r, \qquad r \ge 0.$$

So the spherical formula is asymptotically equal to the Euclidean formula as $r \to 0$, but it is a little smaller. If r is much greater than 0, it is a lot smaller.

The hyperbolic plane. A glimpse into the future. For the hyperbolic plane, the circumference will be $2\pi \sinh(r)$, which grows exponentially. See §22, §28.

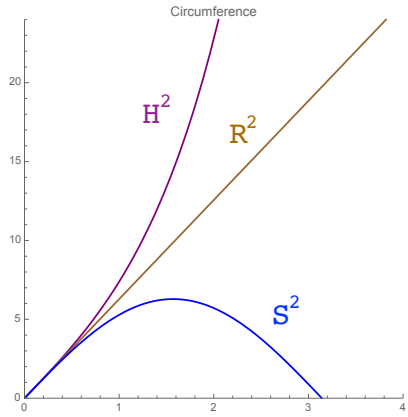

Figure 12.3: Circumference of a circle: $\mathbb{H}^2$, $\mathbb{R}^2$ and S^2 (Mathematica)

Area (by calculus)

Next let us find the area of D_r. We will do it two ways, by calculus and by quoting Archimedes' theorem on the area of spherical sectors.

For the calculus proof, we fill the region between N and K_r by "parallel" circles

$$K_s, \qquad 0 < s \leq r.$$

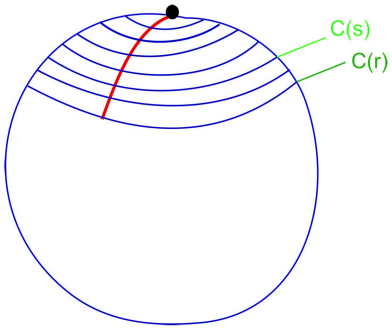

Figure 12.4: Area of the disk by integration of shells

The circles are everywhere equidistant, and the distance between K_s and K_t is

$|t-s|$. So we can compute the area of D_r by

$$\begin{aligned} A_S(r) &= \int_0^r L(K_s)\,ds \\ &= \int_0^r C_S(s)\,ds \\ &= \int_0^r 2\pi\sin(s)\,ds \qquad \text{from above} \\ &= (-2\pi\cos(s)]_{s=0}^{s=r} \\ &= -2\pi\cos(r) + 2\pi\cos(0) \\ &= 2\pi(1-\cos(r)). \end{aligned}$$

So the spherical area of the disk is

$$\boxed{A_S(r) = 2\pi(1-\cos(r)), \qquad 0 \le r \le \pi.}$$

Let us compare this to the Euclidean result. $\cos(r)$ has the expansion

$$\cos(r) = 1 - \frac{r^2}{2} + \frac{r^4}{24} + O(r^6) \qquad \text{as } r \to 0,$$

So we have

$$A_S(r) = \pi r^2 - \frac{\pi r^4}{12} + O(r^6) \qquad \text{as } r \to 0.$$

Compare this to the Euclidean formula

$$A_E(r) = \pi r^2, \qquad r \ge 0.$$

That means, the area is asymptotically equal to the Euclidean value as $r \to 0$, but it is smaller. It resembles the length in this way.

A visual proof that the area is smaller for the sphere case is given by this picture, adapted from J. Weeks, p. 133.

We slice up a spherical disk into angular sectors in order to be able to press it flat onto the plane. Space opens up between the sectors. This establishes that the area of a spherical disk is less than the area of the Euclidean disk of the same radius.

The hyperbolic plane. Message from the future: For the hyperbolic plane we get that the area is $2\pi(\cosh(r)-1)$, which grows exponentially. Here is a comparison of all three.

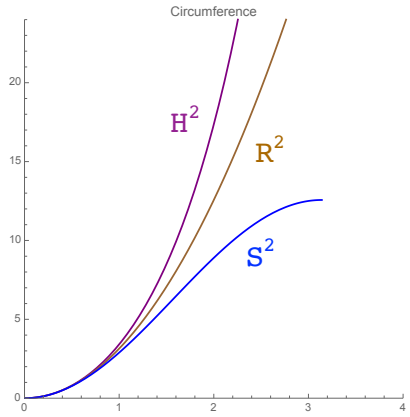

Figure 12.5: Area of a circle: $\mathbb{H}^2$, $\mathbb{R}^2$ and S^2 (Mathematica)

Area (by Archimedes' theorem)

Archimedes' theorem says the following:[1]

> *The area of a sphere cut by two parallel planes normal to an axis equals the corresponding area of a circumscribing cylinder with the same axis.*

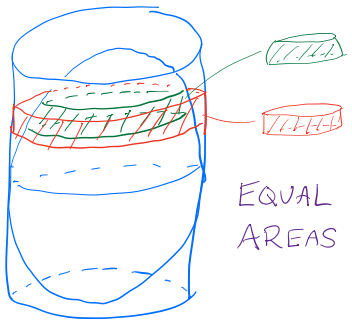

Figure 12.6: The two sectors have equal areas

Such an annular region on the sphere is called a *spherical sector*. The annular region on the cylinder is called a *cylindrical sector*.

Archimedes' theorem has a purely geometric proof, without using symbolic integration.[2] However, it does use "geometric" integration and differentiation, as

[1] Archimedes, *On the Sphere and Cylinder* (Περὶ σφαίρας καὶ κυλίνδρου), around 225 B.C.

[2] T. M. Apostol & M. A. Mnatsakanian, *A fresh look at the method of Archimedes*, Math. Assoc. of America Monthly 111, 2004.

invented by Archimedes to solve problems like this.

Archimedes' theorem can be used to *calculate* the area of a unit sphere – namely, it is the same as the lateral area of the circumscribed cylinder of the same height. This leads to the standard formula 4π.

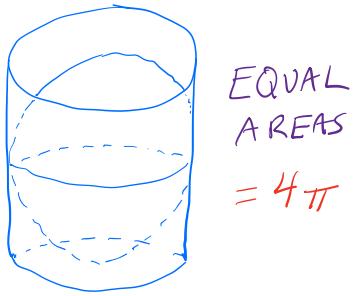

Figure 12.7: A sphere inscribed in a cylinder

If we apply Archimedes' theorem to the spherical cap D_r, we find that it has the same area as a cylinder of radius 1 and height h.

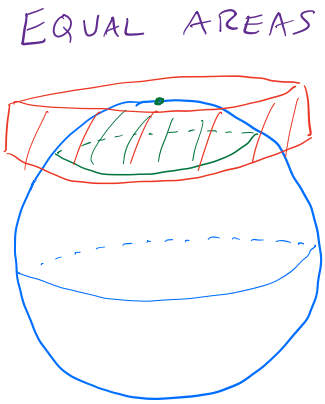

Figure 12.8: A spherical cap compared to a short cylinder

From Figure 12.2,

$$h = 1 - \cos(r).$$

Therefore

$$\begin{aligned} A_S(r) &= 2\pi r_{\text{cylinder}} h \\ &= 2\pi h \\ &= 2\pi(1 - \cos(r)), \end{aligned}$$

as before.

Summary

The circumference and area of an intrinsic disk of radius r in the sphere are

$$\begin{aligned} C_S(r) &= 2\pi \sin(r), \qquad 0 \leq r \leq \pi, \\ A_S(r) &= 2\pi(1 - \cos(r)), \qquad 0 \leq r \leq \pi. \end{aligned}$$

Exercise 12.1 (Dido on the sphere) *Suppose you are given a rope of length a. You are told: you can claim as much land as you can enclose with the rope. Assuming that you want as much land as possible, are you better off in $\mathbb{R}^2$ or in S^2? Does it depend on the length of the rope?*

5

Angle excess

§13 Angle excess

Recall the famous angle sum formula in $\mathbb{R}^2$ (known to Euclid). It says that for a triangle in $\mathbb{R}^2$,

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

Here is the triangle.

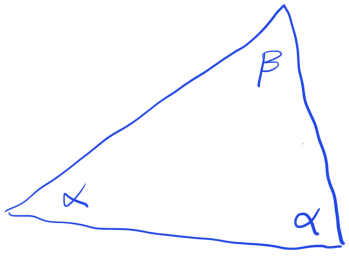

Figure 13.1: Triangle in $\mathbb{R}^2$

What is the corresponding fact in S^2?

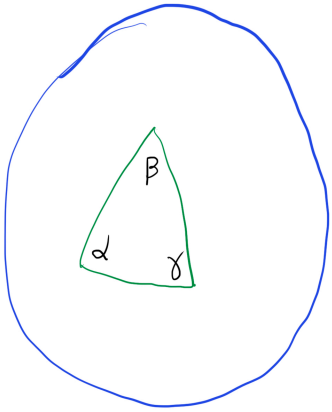

Figure 13.2: Triangle in S^2

Obviously the Euclidean formula does not hold in the sphere. Just consider a triangle with three right angles: one at the north pole and two on the equator. Then the angle sum is $3\pi/2$, significantly larger than π.

On the other hand, a very small triangle is nearly Euclidean, so its angle sum is nearly π.

The larger the triangle, the more it partakes of the sphere's curvature. Here is a half orange slice:[1]

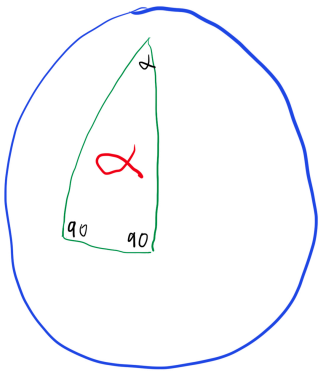

Figure 13.3: Half an orange slice

Its angle sum is

$$\frac{\pi}{2} + \frac{\pi}{2} + \alpha = \pi + \alpha.$$

So it's more than π. The *angle excess* is

$$\begin{aligned} X(T) &:= (\text{angle sum}) - \pi \\ &= (\pi + \alpha) - \pi \\ &= \alpha. \end{aligned}$$

It's proportional to the angle α.

On the other hand, the area of T is also proportional to α. Let us calculate the area of T.

If we take two copies of T, we get a full orange-slice L. This is also called a *lune*.[2]

A lune is characterized by an angle α. If α reaches 2π, then L becomes S^2. So

$$\begin{aligned} A(L) &= \frac{\alpha}{2\pi} A(S^2) \\ &= \frac{\alpha}{2\pi}(4\pi) \\ &= 2\alpha, \end{aligned}$$

[1] Halborangenscheibe.
[2] Kugelzweieck.

so

$$\begin{aligned} A(T) &= \frac{1}{2}A(L) \\ &= \alpha. \end{aligned}$$

So for the half-lune,

$$\text{angle excess} = \text{area}.$$

Amazing.

In fact, this is always true:[3]

Theorem 13.1 (Angle excess formula) *Let T be a triangle in S^2 of interior angles α, β, γ. Then*

$$(\alpha + \beta + \gamma) - \pi = A(T).$$

We will prove this in the next section.

Here is another example. The lune itself is a triangle. It may not look like a triangle, but is a degenerate triangle as shown in the picture:

It has angles α, α, π and angle excess

$$X(T) = (\alpha + \alpha + \pi) - \pi = 2\alpha.$$

And as computed above,

$$A(T) = 2\alpha.$$

So again, for the lune,

$$\text{angle excess} = \text{area}.$$

Hyperbolic case

Here is a glimpse into the future:

The hyperbolic plane is somehow the opposite of the 2-sphere. It has an *angle deficit*. Namely, $\alpha + \beta + \gamma < \pi$, and

$$\pi - (\alpha + \beta + \gamma) = A(T),$$

just the opposite of S^2. Triangles in $\mathbb{H}^2$ are skinnier than Euclidean ones, whereas spherical ones are fatter.

[3] J. Weeks tells us that this result appeared in "De la mesure de la superfice des triangles et polygones sphéricques, nouvellement inventée" in the book *Invention nouvelle en l'algebre* by Albert Girard, 1629.

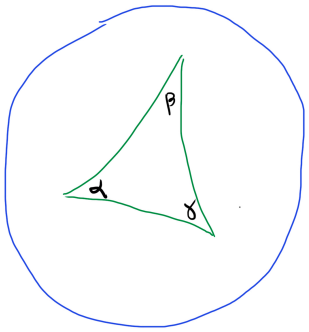

Figure 13.4: Triangle in $\mathbb{H}^2$

Tantalizing question How large can the area of a triangle in $\mathbb{H}^2$ be?

Surprisingly, it turns out that the area is limited by π (proof needed).

But at the same time, there is exponentially much area in $\mathbb{H}^2$.

So it *seems* that very large triangles (triangles with long sides) ought to have extremely much area. Yet they don't.

How can these apparently contradictory assertions be reconciled? At this point, we have no idea. We're swimming. We'll have to wait.

Here is a comparison of triangles in the three spaces, borrowed from Jeff Weeks.

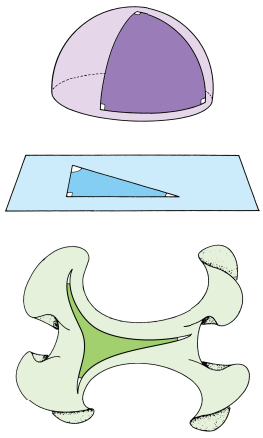

Figure 13.5: Three types of triangles (J. Weeks)

§14 Proof of the angle excess formula

References

- J. R. Weeks, *The Shape of Space,* pp. 125-134.

Proof of Theorem 13.1 Note that in going from the half orange-slice to the full orange-slice, there seemed to be a kind of additivity. Let's try to exploit this to prove the Angle Excess Formula.

Let T be a triangle with angles α, β, γ. Extend all three sides until they are three great circles. Then the three circles divide S^2 into 8 triangles.

Now the proof can be seen at a glance from this picture and the one below, but the verbal description takes a little longer.

Consider the *antipodal map*[4]

$$Z : S^2 \to S^2, \qquad x \to -x.$$

It takes T to its opposite triangle

$$Z(T) = -T$$

and back again.

Now T and $-T$ are isometric. To see this, observe that Z can be expressed as the composition

$$Z = R_x \circ \mathbb{R}_y \circ R_z$$

of three mirror reflections, namely

$$\begin{aligned} R_x &:= \text{reflection across the } y\text{-}z \text{ plane,} \\ R_y &:= \text{reflection across the } x\text{-}z \text{ plane,} \\ R_z &:= \text{reflection across the } x\text{-}y \text{ plane.} \end{aligned}$$

Since each of these is an isometry of S^2, Z is an isometry of S^2 (maybe this was obvious).

So T is isometric to $-T$. So T and $-T$ have the same angles and the same area.

Now the 8 triangles arrange thenselves into lunes in many ways. Any two adjacent triangles form a lune.

The remaining 6 triangles besides T and $-T$ form a cyclic chain that lies between

[4] Antipodenabbildung.

T and $-T$. The successive triangles in this chain have angles

$$\begin{array}{lll} \alpha, & \pi-\beta, & \pi-\gamma \\ \pi-\alpha, & \beta, & \pi-\gamma \\ \pi-\alpha, & \pi-\beta, & \gamma \\ \alpha, & \pi-\beta, & \pi-\gamma \\ \pi-\alpha, & \beta, & \pi-\gamma \\ \pi-\alpha, & \pi-\beta, & \gamma. \end{array}$$

We can fuse the 1st and 2nd triangle, the 3rd and 4th triangle, and the 5th and 6th triangle to form 3 disjoint lunes, defined by angles

$$\pi-\gamma, \qquad \pi-\beta, \qquad \pi-\alpha.$$

Together with the two original triangles, these three lunes exactly fill the sphere. The following picture is adapted from Weeks, p. 130.

Recalling our observation in the last section, the lunes have areas

$$2(\pi-\gamma), \qquad 2(\pi-\beta), \qquad 2(\pi-\alpha).$$

It follows that the two triangles T, $-T$ have total area

$$\begin{aligned} A(T)+A(-T) &= A(S^2) - (\text{area of 3 lunes}) \\ &= 4\pi - 2(\pi-\gamma) - 2(\pi-\beta) - 2(\pi-\alpha) \\ &= 2(\gamma+\beta+\alpha) - 2\pi. \end{aligned}$$

Since T and $-T$ are isometric, they must have the same area. So

$$A(T) = (\alpha+\beta+\gamma) - \pi,$$

the angle excess of T, as was desired to prove.

□

6

Stereographic projection

§15 The map problem

References

- *Mercator projection*, Wikipedia
- *Conformal map projection*, Wikipedia
- *Equal-area map*, Wikipedia
- *Lambert cylindrical equal-area projection*, Wikipedia

Let's discuss maps of the earth's surface.

Maps

A *geographical map* is a bijection

$$f : U \subseteq S^2 \to V \subseteq \mathbb{R}^2$$

between an open subset of the sphere and an open subset of the plane. We use f to transport all features (cities, coastlines, graticule,[1] etc.) from U to V.

Here is the familiar Mercator map:

Figure 15.1: Mercator map (Strebe, Wikipedia)

It extends infinitely far upwards and downwards.

Here is the bijection (called the *Mercator projection*) that produces it.

It takes circles of latitude to horizontal lines and lines of longitude to vertical lines. It is a *cylindrical projection* followed by unrolling[2] of the cylinder.

[1] Gradnetz oder Kartennetz.
[2] Entrollen.

The exact function that assigns latitude circles to cylinder circles in the Mercator projection is carefully computed so that the Mercator projection preserves angles. It is not obvious (it's not straight-line projection).

Preservation of properties

In constructing a map, one wishes to preserve all geometric properties: length, distance, angle, area.

It is impossible to preserve all of these at once. So some kind of compromise has to be made.

Why can't we preserve all of these at once?

First of all, one can show in all generality:[3]

Lengths determine both angles and areas.

So it would be enough to preserve lengths. Then everything is preserved.

On the other hand, one can also show:[4]

If both angle and area are preserved, then length is preserved.

We haven't proven these things, but you can imagine how they might be proven.[5]

Now here's the thing:

The function f cannot preserve lengths.

Here's why: In §§12-14, we've shown that the intrinsic geometric properties of S^2 differ in essential ways from those of $\mathbb{R}^2$. In particular:

- The circumference and area of a disk are different.
- The angle sum of a triangle is different.

So there cannot be a isometry between any open subset of S^2 and an open set of $\mathbb{R}^2$.

It then follows:

f cannot preserve both angle and area.

For if both were preserved, then length would be preserved, which is impossible.

Types of maps

So there are, broadly speaking, two categories of geographical map in use:

- Angle-preserving,
- Area preserving,

[3] In any dimension.
[4] In any dimension.
[5] Linear algebra.

as well as some hybrid forms. You have to make tradeoffs and there is no perfect solution.

An angle-preserving function is called *conformal*. The Mercator projection above is conformal.

A well-known area-preserving map is the Lambert cylindrical equal area projection:

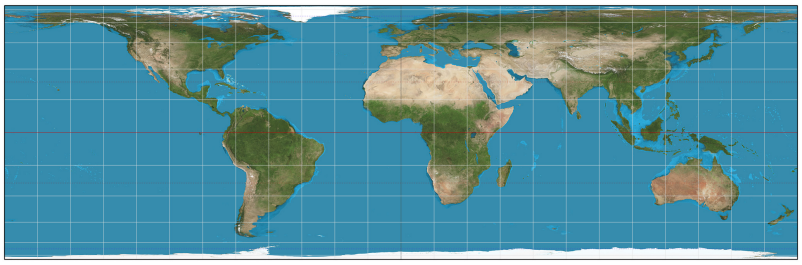

Figure 15.2: Lambert cylindrical equal area projection (Strebe, Wikipedia)

It is obtained by projecting horizontally outward from the vertical axis onto the cylinder of height 2 that is tangent to S^2 along the equator, and then unrolling the cylinder. Like the Mercator projection, it is a cylindrical projection.

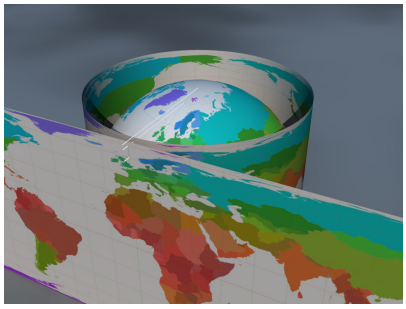

Figure 15.3: How Lambert is done (KoenB, Schuyler Erle, Wikipedia)

The Lambert projection preserves area because of Archimedes' Theorem.

Exercise 15.1 *The Lambert cylindrical equal area projection has a lot of angular distortion, especially toward the poles. What is a very simple way to reduce the distortion? (Hint: Wikipedia)*

Wikipedia lists at least 13 conformal projections and 21 equal area projections, and some mixed ones. These are ones that were significant enough to get names. There are infinitely many.

§16 Stereographic projection

References

- *Stereographic projection*, Wikipedia
- *Stereographic map projection*, Wikipedia

Note: The notation in this section has been changed.

Mathematically, perhaps the most important mapping from S^2 to $\mathbb{R}^2$ is *stereographic projection*. It will turn out to be conformal.

It is defined as follows. Let $N = (0, 0, 1)$ be the north pole of S^2, and let $\mathbb{R}^2$ be the x-y plane. Then $\mathbb{R}^2$ is the horizontal plane containing the equator of S^2.

Fix $P \in S^2$, $P \neq N$. Draw a line L through N and P. Let $\sigma(P) = Q$ be the point where L meets $\mathbb{R}^2$. Then

$$\sigma : S^2 \setminus \{N\} \to \mathbb{R}^2$$

is stereographic projection.

Geometrically, it is clear that σ is a bijection.

Calculate stereographic projection

What is the formula for $\sigma(P)$?

Write

$$P = (X, Y, Z), \qquad \sigma(P) = Q = (x, y).$$

The line L has the parametrization

$$t \mapsto N + t(P - N), \qquad t \in \mathbb{R}.$$

To obtain the intersection point Q of L with $\mathbb{R}^2$, we must solve

$$N + t(P - N) = (x, y, 0)$$

for t, x, y in terms of X, Y, Z. It becomes

$$(0, 0, 1) + t((X, Y, Z) - (0, 0, 1)) = (x, y, 0)$$

i.e.

$$tX = x, \qquad tY = x, \qquad 1 + t(Z - 1) = 0.$$

so

$$t = 1/(1 - Z) > 0$$

since $Z < 1$. So

$$x = \frac{X}{1 - Z}, \qquad y = \frac{Y}{1 - Z}.$$

So

$$\boxed{\sigma(P) = \sigma(X, Y, Z) = \left(\frac{X}{1-Z}, \frac{Y}{1-Z}\right), \qquad \sigma : S^2 \setminus \{N\} \to \mathbb{R}^2.}$$

Calculate the inverse map

The inverse map

$$\tau = \sigma^{-1} : \mathbb{R}^2 \to S^2, \qquad Q \mapsto \tau(Q)$$

can be found by solving

$$\sigma(X, Y, Z) = (x, y), \qquad X^2 + Y^2 + Z^2 = 1,$$

for (X, Y, Z) in terms of x, y.

This takes a few steps. Here are the details.

We get

$$\frac{X}{1-Z} = x$$
$$\frac{Y}{1-Z} = y$$
$$X^2 + Y^2 + Z^2 = 1.$$

So

$$|(X, Y)| = (1 - Z)|(x, y)| = (1 - Z)|Q|.$$

So

$$1 = X^2 + Y^2 + Z^2 = (1 - Z)^2|Q|^2 + Z^2.$$

So

$$(1 - Z)(1 + Z) = (1 - Z)^2|Q|^2.$$

Because we can't have $Z = 1$, we can divide by $1 - Z$ and get

$$1 + Z = (1 - Z)|Q|^2.$$

So

$$Z = \frac{|Q|^2 - 1}{|Q|^2 + 1}. \tag{16.1}$$

So

$$X = (1 - Z)x = \frac{2x}{|Q|^2 + 1}, \qquad Y = (1 - Z)y = \frac{2y}{|Q|^2 + 1}.$$

So the inverse of stereographic projection is

$$\tau(Q) = \tau(x, y) = (X, Y, Z) = \frac{(2x, 2y, |Q|^2 - 1)}{|Q|^2 + 1}.$$

For future use, we write $z = x + iy$ instead of $Q = (X, Y)$. We get

$$\tau(z) = \tau(x, y) = \frac{(2x, 2y, |z|^2 - 1)}{|z|^2 + 1}, \qquad \tau = \sigma^{-1} : \mathbb{C} \to S^2 \setminus \{N\}.$$

More images

Here is an image from Wikipedia:

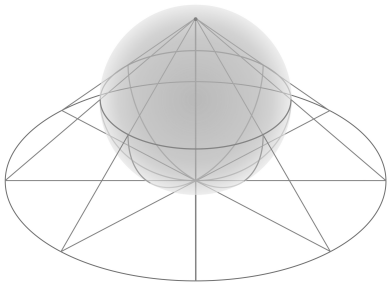

Figure 16.1: Stereographic projection (Che Che, Mark.Howison, Wikipedia)

In this figure, instead of using the plane $\mathbb{R}^2 = \{Z = 0\}$ as the target, we use $\{Z = -1\}$. This does not make much difference. We could use any horizontal plane $\{Z = a\}$, $a < 1$ as the target. The maps we obtain are all related by a scale factor.[6]

Exercise 16.1 *Verify this.*

Here is what it looks like as a geographic map. The map is infinite in extent, with an arbitrarily large amount of expansion of the Antarctic region. I have to admit that the infiniteness is not very well shown on the map. What is the projection point in this case?

[6] See *Stereographic projection*, Wikipedia.

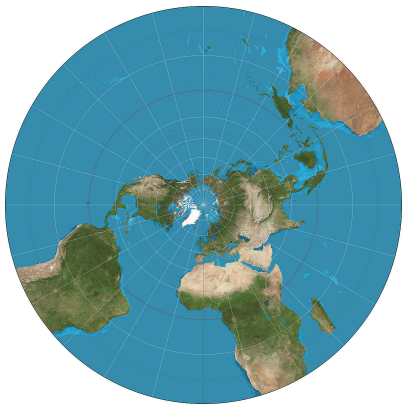

Figure 16.2: Stereographic projection (Strebe, Wikipedia)

§17 The spherical metric on $\mathbb{R}^2$

How can we use stereographic projection to express the metric of S^2 by calculations on $\mathbb{R}^2$?

An ant lives on $\mathbb{C}$, but believes she lives in S^2. That is, she walks around in $\mathbb{C}$, but she experiences the geometry of the corresponding points of S^2 under the inverse stereographic projection τ. This is considered a delusion by the other ants.

Let us express her notion of distance in terms of $\mathbb{C}$ distances.

Answer:

It is just a calculation. Let z, w be points of $\mathbb{C}$. Then the ant's private notion of distance is obtained by mapping the points z, w to S^2 via

$$\tau : \mathbb{C} \to S^2 \setminus \{N\},$$

finding the S^2 distance between them (geodesic distance), and that's the result.

That is, her private notion of distance is

$$d_{private}(z, w) = d_S(\tau(z), \tau(w)).$$

From the previous section, with $z = x + iy$, $w = u + iv$

$$P := \tau(z) = \frac{(2x, 2y, |z|^2 - 1)}{|z|^2 + 1},$$

and similarly

$$P' := \tau(w) = \frac{(2u, 2v, |w|^2 - 1)}{|w|^2 + 1}.$$

Then

$$\begin{aligned} d_{private}(z, w) &= d_S(P, P') \\ &= \theta \qquad\qquad \text{from §11,} \end{aligned}$$

where P and P' are the image points in S^2 and θ is the central angle between them.

Then by the rule relating angle to inner product (linear algebra)

$$\begin{aligned}
\cos(\theta) &= \frac{P \cdot P'}{|P||P'|}, \\
&= P \cdot P' \qquad\qquad \text{since } |P| = |P'| = 1, \\
&= \left(\frac{(2x, 2y, |z|^2 - 1)}{|z|^2 + 1}\right) \cdot \left(\frac{(2u, 2v, |w|^2 - 1)}{|w|^2 + 1}\right) \\
&= \frac{4xu + 4yv + (|z|^2 - 1)(|w|^2 - 1)}{(|z|^2 + 1)(|w|^2 + 1)} \\
&= 1 + \frac{4xu + 4yv + (|z|^2 - 1)(|w|^2 - 1) - (|z|^2 + 1)(|w|^2 + 1)}{(|z|^2 + 1)(|w|^2 + 1)} \\
&= 1 + \frac{4xu + 4yv - 2|z| - 2|w|^2}{(|z|^2 + 1)(|w|^2 + 1)} \\
&= 1 - 2\frac{-2xu - 2yv + |z|^2 + |w|^2}{(|z|^2 + 1)(|w|^2 + 1)} \\
&= 1 - 2\frac{-2xu - 2yv + x^2 + y^2 + u^2 + v^2}{(|z|^2 + 1)(|w|^2 + 1)} \\
&= 1 - 2\frac{(x - u)^2 + (y - v)^2}{(|z|^2 + 1)(|w|^2 + 1)} \\
&= 1 - 2\frac{|z - w|^2}{(|z|^2 + 1)(|w|^2 + 1)}.
\end{aligned}$$

We conclude

$$d_{private}(z, w) = \arccos\left(1 - 2\frac{|z - w|^2}{(|z|^2 + 1)(|w|^2 + 1)}\right).$$

This is how to calculate her private fantasy of living on the sphere in terms of the real-life geometry of the Euclidean plane.

Later every ant comes to share her perceptions, so they no longer have any way to refute it, so it becomes reality.[7] *There is only one ant who's left out, but he's a lightning calculator, so he can use the above formula to* pretend *he fits in with the other ants.*

In summary (changing the notation slightly):

Proposition 17.1 *The metric of S^2, expressed on $\mathbb{R}^2$ via stereographic projection, is*

$$\boxed{d_S(z, w) = \arccos\left(1 - \frac{2|z - w|^2}{(|z|^2 + 1)(|w|^2 + 1)}\right), \qquad z, w \in \mathbb{C}.}$$

[7] Shades of Philip K. Dick.

7

Spherical arclength on S^2

§18 Similarities of $\mathbb{R}^2$

References

- Ilmanen, *Geometrie 2020*, §10, "The bare minimum for orientation".
- Jänich, *Lineare Algebra*, Springer, 11th edition, 2010, pp. 70-73, 157 (orientation).
- Fischer, *Lineare Algebra: Eine Einführung für Studienanfänger*, 18th edition, Springer, 2014, pp. 212-221 (orientation).
- Weeks, pp. 41-58 (orientation of manifolds - lots of information).

In this section we:

- Define similarities
- Show that all similarities of $\mathbb{R}^n$ are conformal
- Find all the similarities of $\mathbb{R}^2$, and express them in complex notation.
- Classify them as orientation-preserving and orientation-reversing.

These ideas will help us to understand the result in the next section, and we'll use them later for Möbius transformations as well.

Similarities

A bijection

$$f : \mathbb{R}^n \to \mathbb{R}^n$$

is a *similarity*[1] if it stretches all distances by a constant factor:

$$|f(x) - f(y)| = \lambda |x - y|, \qquad x, y \in \mathbb{R}^n,$$

where λ is a constant. If $\lambda = 1$ then it is an isometry. So a similarity is an isometry, plus a scale factor. We can define the same concept for maps between metric spaces.

It is obvious that compositions and inverses of similarities are similarities.

Similarities of $\mathbb{R}^n$ are conformal

Recall that a map is called *conformal* if it is angle-preserving. We have:

Proposition 18.1 *A similarity of $\mathbb{R}^n$ is angle-preserving.*

Proof Recall the law of cosines:

$$2ab\cos(\gamma) = a^2 + b^2 - c^2.$$

[1] Ähnlichkeitstransformation.

But if we multiply all distances by the same constant λ, we don't change $\cos(\gamma)$. So we don't change γ. This proves the result.

□

Note: The law of cosines is why we say that lengths determine angles.

Similarities of $\mathbb{R}^2$

What are the similarities of $\mathbb{R}^2$?

Here are some:

- Rotations about 0 by some angle θ (here $\lambda = 1$)
- Dilations (expansions, homotheties)[2] by a factor $\lambda > 0$
- Reflections across a line through 0 (again $\lambda = 1$)
- Translations by some vector b (again $\lambda = 1$)

These can be represented in complex notation by

- Rotations: $z \mapsto e^{i\theta} z$, where $\theta \in \mathbb{R}$
- Dilations: $z \mapsto \lambda z$, where $\lambda \in \mathbb{R}$, $\lambda > 0$
- Conjugation: $z \mapsto \bar{z}$
- Translations: $z \mapsto z + b$, where $b \in \mathbb{C}$.

In fact, as we shall see, these generate all the similarities of $\mathbb{C}$.

Rotation-expansions

Rotations and expansions can be combined into one operation by defining

$$a := \lambda e^{i\theta}$$

and become

- Rotation-expansions: $z \mapsto az$, where $a \in \mathbb{C}$, $a \neq 0$.

So a rotation-expansion is nothing more than a complex-linear map of $\mathbb{C}$.

Reflections across lines

The third operation, conjugation, is only one example of reflection across a line.

Exercise 18.1

[2] Streckungen.

a) Show that $z \mapsto e^{i\phi}\bar{z}$ is the most general reflection across a line through 0.
b) What line does it reflect across?

Composition rules for orientation

A map is *orientation-preserving* if it takes right hands to right hands and left hands to left hands. (See the references at the beginning of the section.)

A map is *orientation-reversing* if it takes right hands to left hands and left hands to right hands.

Let

$$E = \text{orientation-preserving}, \qquad U = \text{orientation reversing}.$$

We have the following rules:

$$E \circ E = E$$
$$E \circ U = U$$
$$U \circ E = U$$
$$U \circ U = E$$

It's a matter of parity, or lex talionis – two wrongs DO make a right.

Orientation-preserving similarities

Rotations, expansions, and translations are all orientation-preserving.

So by the composition rules for orientation, any composition of them is an orientation-preserving similarity of $\mathbb{C}$.

It turns out that the most general orientation-preserving similarity of $\mathbb{C}$ is a rotation-expansion followed by a translation, namely

$$z \mapsto az + b,$$

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

Exercise 18.2 *Prove this.*

So: an *orientation-preserving similarity* of $\mathbb{C}$ is the same as a *complex affine map.*

It follows that the orientation-preserving similarities *that fix zero* are precisely the complex-linear maps

$$z \mapsto az,$$

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

Orientation-reversing similarities

Conjugation $z \mapsto \bar{z}$ is *orientation-reversing.*

So by the composition rules for orientation, any composition of rotations, expansions, and translations, together with an odd number of conjugations, is an orientation-reversing similarity of $\mathbb{C}$.

It turns out that the most general orientation-reversing similarity of $\mathbb{C}$ is conjugation, followed by a rotation-expansion, followed by a translation, namely

$$z \mapsto a\bar{z} + b,$$

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

Exercise 18.3 *Prove this.*

Non-similarities

What linear maps of R^3 are not similarities?

Two examples are

- shears, such as $(x, y) \mapsto (x, x + y)$.
- maps with two different stretch factors, such as $(x, y) \mapsto (2x, 3y)$.

Exercises

Exercise 18.4 *Let $p \in \mathbb{C}$. Let R_p be reflection through p, that is, $R_p(z)$ is the point lying on the opposite size of p from z, but at the same distance from p as z.*

a) Is R_p a similarity transformation?
b) Show $R_p^2 = \mathrm{id}$.
c) Is R_p orientation-preserving or orientation-reversing?
d) Express $R_p(z)$ in terms of z using complex addition, multiplication, and conjugation.
e) What is $R_q \circ R_q$?

Exercise 18.5 *Let T_a be translation by a and H_λ multiplication by $\lambda > 0$.*

a) Describe the effect of $T_a \circ H_\lambda \circ T_{-a}$ geometrically. Express it in terms of complex addition and multiplication.
b) Describe the effect of $H_\lambda \circ T_a \circ H_{1/\lambda}$ geometrically. Express it in terms of complex addition and multiplication.

§19 Spherical arclength

References

- Anderson, section 3.1.

In order to better understand the experience of the ant, we want to find the infinitesimal stretching factor of the function

$$\tau : (\text{map}) \to (\text{territory}),$$

that is, of inverse stereographic projection from the flat map to the round territory.[3]

Infinitesimal similarity

We have the following theorem.

Theorem 19.1 *Let $\tau : \mathbb{C} \to S^2 \setminus \{N\}$ be the inverse of stereographic projection. Then*

a) For a very small neighborhood U around any point z, τ is nearly a similarity between U and $\tau(U)$.

b) The stretch factor of τ at z is

$$g(z) := \frac{2}{|z|^2 + 1}, \qquad z = x + iy \in \mathbb{C}.$$

One could say that at each point, τ is a similarity at the infinitesimal level, with the given stretch factor.

(It will follow from this in the next section that τ is conformal.)

Note that the stretch factor goes to zero as $|z| \to \infty$, reflecting the fact that distances on S^2 are a very tiny multiple of distances on $\mathbb{C}$ when $|z|$ is large.

The proof is a calculation.

Proof outline

1. We prove b).

Fix z. If w, w' are very close to z, say $w, w' \in U = B_\delta(z)$, let us see how the tiny segment

$$[w, w']$$

in $\mathbb{C}$ gets stretched to form a tiny segment

$$[\tau(w), \tau(w')]$$

[3] A. Korzybski, "The map is not the territory".

in R^3.

Effectively, we have to do a differentiation at z, but we do it with O-notation.

Recall from the previous section that the metric d_S of S^2 can be expressed in terms of the metric $d_E = |\cdot - \cdot|$ of $\mathbb{R}^2$. Write

$$D := d_S(\tau(w), \tau(w')).$$

Then D is small when δ is small because τ is continuous. Then from the previous section,

$$\begin{aligned}
\cos(D) &= 1 - \frac{2|w-w'|^2}{(|w|^2+1)(|w'|^2+1)} \\
1 - \frac{D^2}{2} + O(D^4) &= 1 - \frac{2|w-w'|^2}{(|w|^2+1)(|w'|^2+1)} \\
\frac{D^2}{2}(1+O(D^2)) &= \frac{2|w-w'|^2}{(|z|^2+1)^2}(1+O(\delta^2)) \qquad \text{(verify)}.
\end{aligned}$$

Taking the square root,

$$D(1+O(D^2)) = \frac{2|w-w'|}{|z|^2+1}(1+O(\delta^2)) \qquad \text{(verify)}.$$

Since D is small, it follows that

$$D = O(|w-w'|)(1+O(\delta^2)) = O(\delta).$$

(Interestingly, the second equal sign is not symmetric.) We feed this back in to the previous equation to get

$$D = \frac{2|w-w'|}{|z|^2+1}(1+O(\delta^2)).$$

or spelled out,

$$d_S(\tau(w), \tau(w')) = \frac{2|w-w'|}{|z|^2+1}.$$

So $\tau|U$ is *nearly* a similarity transformation from U to $\tau(U)$, with stretch factor equal to

$$\frac{2}{|z|^2+1}(1+O(\delta^2))$$

on $U = B_\delta(z)$. This proves a).

2. As $\delta \to 0$, the stretch factor on $B_\delta(z)$ converges to $2/(|z|^2+1)$. That is what we mean by saying that τ has a stretch factor of

$$\frac{2}{|z|^2+1}$$

at the point z. This proves b).

□

Lengths of curves

Now we find out how to calculate path lengths of curves in S^2, by transferring the arclength of S^2 to $\mathbb{C}$ via τ^{-1}.

Let

$$ds_E = \sqrt{dx^2 + dy^2}$$

denote Euclidean arclength in $\mathbb{R}^2$. This is shorthand for

$$\begin{aligned} L_E(\gamma) &= \int_\gamma ds_E \\ &= \int_a^b \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2}\, dt. \end{aligned}$$

As we saw above, the stereographic model

$$\tau : \mathbb{C} \to S^2 \setminus \{0\},$$

has length multiplier

$$g(z) := \frac{2}{|z|^2 + 1}, \qquad z = x + iy \in \mathbb{C}.$$

So we define a "spherical" arclength on $\mathbb{R}^2$ by

$$\boxed{ds_S := \frac{2}{|z|^2 + 1} ds_E.}$$

This ds_S lives on $\mathbb{C}$, whereas the original ds_S lives on S^2.

This "spherical" arclength can now be used by the prophet ant of the last section (or by her skeptical brother) to compute S^2 lengths while living in $\mathbb{C}$, as follows.

We get for a continuously differentiable curve $\gamma : [a, b] \to \mathbb{C}$ in $\mathbb{C}$:

$$\begin{aligned} L_S(\gamma) &= \int_\gamma ds_S \\ &= \int_\gamma \frac{2}{|z|^2 + 1} ds_E \\ &= \int_a^b \frac{2}{|z|^2 + 1} \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2}\, dt, \end{aligned}$$

This is the "spherical" length of γ in c, and it equals the actual length of $\tau \circ \gamma$ in S^2.

§20 Stereographic projection is conformal

As a corollary, we have that stereographic projection is conformal.

Corollary 20.1 *Let $\sigma : \mathbb{C} \to S^2 \setminus \{N\}$ be stereographic projection. Let $\tau : \mathbb{C} \to S^2 \setminus \{N\}$ be its inverse. Then σ and τ τ preserve angles, that is, they are is conformal.*

This is obvious from Theorem 19.1, but we will give a long-winded proof just to be sure.

It suffices to prove it for τ.

The brief sketch is this: By Theorem 19.1, τ is a similiarity transformation in an infinitesimal neighborhood of each $a \in \mathbb{R}^2$, so it preserves angles at each a, so it is conformal. If you are satisfied with this, you need read no further.

A fuller sketch: Using the law of cosines, we proved (Proposition 18.1) that if the stretch factors are exactly constant then the function exactly preserves angles.

For the case at hand, fix a. By Theorem 19.1, the stretch factors in a small neighborhood U about a are nearly constant, so τ nearly preserves angles near a.

We pass the size of U to zero to show that τ exactly preserves angles at a. This is true for all a in $\mathbb{R}^2$. So τ is conformal.

An alternative way to prove this is to combine Theorem 31.2 and §34, Remark 3.

Yet another alternative is to use calculus (derivatives).

Here is the full proof.

Proof 1. We will use the law of cosines to make the proof rigorous. It reduces angles to distances, as in the last section.

Note that it suffices to prove that τ is c onformal.

Let L, M be lines in $\mathbb{R}^2$ that meet at a at an angle α.

Then $\tau(L)$ and $\tau(M)$ are curves in S^2 that meet at $a' = \tau(a)$ at some angle $\tilde{\alpha}$.

We may assume $0 < \alpha, \tilde{\alpha} < \pi$. Our aim is to prove

$$\tilde{\alpha} = \alpha.$$

2. This will be successful due to the fact that τ is nearly a similarity transformation for points close to a.

Select points b, c on L, M close to a and consider the tiny triangle bac. Then the angle of this triangle at a is α.

Let

$$a' = \tau(a), \qquad b' = \tau(b), \qquad c' = \tau(c).$$

Let α, β, γ be the angles at a, b, c. Let α', β', γ' be the angles at a', b', c'. Let B, C, A', B', C' be the lengths of the edges opposite a, b, c, a', b', c'.

For technical reasons we will require

$$B = C$$

so that the triangle is not too distorted. Then $\alpha' \approx \tilde{\alpha}$, and indeed

$$\alpha' \to \tilde{\alpha}$$

as $b, c \to a$.

3. Now here is the punchline – because τ is nearly a similarity transformation near a, we can use the law of cosines to show that

$$\alpha' \approx \alpha.$$

Indeed, by the law of cosines, in $\mathbb{R}^2$ we have

$$2BC\cos\alpha = B^2 + C^2 - A^2 \tag{20.1}$$

and in $\mathbb{R}^3$ we have

$$2B'C'\cos\alpha' = (B')^2 + (C')^2 - (A')^2, \tag{20.2}$$

So the distances determine the angle.

But τ is nearly a similarity transformation near a. This yields (examining the proof of Theorem 19.1),

$$A' = (1 + O(\varepsilon^2))\,\lambda A, \qquad B' = (1 + O(\varepsilon^2))\,\lambda B, \qquad C' = (1 + O(\varepsilon^2))\,\lambda C,$$

where $\varepsilon := B = C \geq A/2$ and $\lambda := g(a)$ is the stretch factor at a. So (20.2) becomes

$$\begin{aligned}(1 + O(\varepsilon^2))^2\, 2(\lambda B)(\lambda C)\cos\alpha' \\ = (1 + O(\varepsilon^2))^2\,(\lambda A)^2 + (1 + O(\varepsilon^2))^2\,(\lambda B)^2 - (1 + O(\varepsilon^2))^2\,(\lambda C)^2.\end{aligned}$$

which becomes

$$2BC\cos\alpha' = B^2 + C^2 - A^2 + O(\varepsilon^4).$$

Comparing with (20.1), we get

$$2BC\cos\alpha = 2BC\cos\alpha' + O(\varepsilon^4).$$

Dividing by $2BC$, we get

$$\cos\alpha = \cos\alpha' + O(\varepsilon^2).$$

Passing $b, c \to a$, and recalling that $\alpha' \to \tilde{\alpha}$, and we get

$$\cos\alpha = \cos\alpha' + O(\varepsilon^2) \to \cos\tilde{\alpha}.$$

So

$$\cos\alpha = \cos\tilde{\alpha}.$$

But $\cos : [0, \pi] \to [-1, 1]$ is bijective, and $0 < \alpha, \tilde{\alpha} < \pi$, so

$$\tilde{\alpha} = \alpha.$$

So τ preserves angles. So τ is conformal.

□

Part III

Hyperbolic space

8

The hyperbolic metric

§21 The hyperbolic metric

References

- Weeks, chapter 10.
- Hitchman, section 5.1 (but he has already covered the Möbius group).
- Anderson, section 3.2.

We'll present

- The hyperbolic plane
- Path length
- Distance
- Angles

Recall the Euclidean length element on the plane, called

$$ds_E.$$

In §19, we derived the spherical length element, inherited from S^2 via stereographic projection. It is

$$ds_S = \frac{2}{1+|z|^2}\, ds_E, \qquad z \in \mathbb{R}^2.$$

We get the hyperbolic length element by flipping the sign. It is

$$ds_H := \frac{2}{1-|z|^2}\, ds_E, \qquad |z| < 1.$$

The length stretching factor is

$$h(z) := \frac{2}{1-|z|^2}, \qquad |z| < 1.$$

The unit disk B_1, equipped with this notion of arclength, is called the *Poincaré disk model* of the hyperbolic plane. We write

$$(B_1, ds_H)$$

to denote this structure. We write

$$\mathbb{H}^2$$

to denote any space isometric to it.

Let $\gamma : [a, b] \to \mathbb{C}$ be a continuously differentiable curve. Its length is given by

$$\begin{aligned} L_H(\gamma) &= \int_\gamma ds_H \\ &= \int_\gamma \frac{2}{1-|z|^2}\, ds_E \\ &= \int_a^b \frac{2}{1-|z|^2} \sqrt{\left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2}\, dt. \end{aligned}$$

Distance is given by

$$d_H(z, w) := \inf\{L_H(\gamma) : \gamma \text{ is a continuously differentiable curve connecting } z \text{ to } w\}.$$

What about angles?

We define hyperbolic angles (between two curves) to be the same as the euclidean angle.

We won't justify this completely at this time, but there is a very good reason for this: In a small neighborhood U of a point z_0, the hyperbolic metric is *nearly* a constant multiple of the Euclidean metric, namely by a factor

$$h(z) = h(z_0)(1 + o(1)), \qquad z \in U.$$

So the hyperbolic angles are *nearly* the same as the Euclidean angles. Since we can take U as small as we like without affecting the angle at z_0, this relationship is exact.

Exercise 21.1 *We said earlier "distance determines angle". Use the law of cosines*

$$2ab\cos(\gamma) = a^2 + b^2 - c^2,$$

valid in Euclidean space, to make the above heuristic argument more precise.

Because the two metrics have the same angles, we say that they are *conformally equivalent*.

§22 Sizes near the boundary

We return to the figure mentioned in §5.

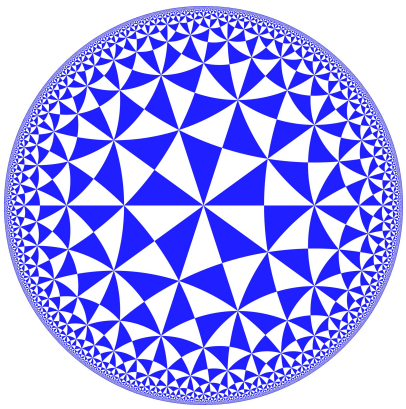

Figure 22.1: Order-4 bisected pentagonal tiling of the hyperbolic plane (Rocchini, Wikipedia)

The triangles appear to get smaller and smaller as $z \to \partial B_1$, but in the hyperbolic metric, they are the same size. Indeed, the stretch factor is

$$\frac{2}{1-|z|^2} = \frac{2}{(1+|z|)(1-|z|)} \sim \frac{1}{1-|z|}$$

as $|z| \to 1$. Asymptotically, the hyperbolic metric is proportional to the inverse of the Euclidean distance to the boundary.

So the equal-sized hyperbolic triangles must get smaller and smaller as we converge to the boundary. Indeed, they must have Euclidean size roughly proportional to the Euclidean distance to the boundary. And indeed, viewing the figure, this seems to be the case.

Exercise 22.1 *Suppose we use the "number of triangles" as a crude way of gauging distance. Start at the origin and proceed n triangles towards the boundary. Very roughly, how many triangles are at this "distance" from the origin?*

Answer: Say that the Euclidean size of a triangle is roughly λ times its Euclidean distance to the boundary. For convenience (because this is only a rough estimate) suppose $\lambda = 1$.

We make a chain

$$T_1, T_2, \dots$$

of triangles starting at the origin and proceeding toward the boundary.

Then, very roughly, T_1 has Euclidean size $1/2$, T_2 has size $1/4$, etc. In general

$$\text{size of } T_n = 2^{-n}, \qquad \text{dist}(T_n, \partial B_1) = 2^{-n}.$$

So T_n touches the circle C_r defined by $r = 1 - 2^{-n}$. The Euclidean length of C_r is roughly 6, so the hyperbolic length is roughly

$$\frac{6}{1-r} = 6 \cdot 2^n$$

Since the triangles are all the same hyperbolic size, say c, there are roughly

$$\frac{6 \cdot 2^n}{c} = C \cdot 2^n$$

such triangles along C_r. But we can't trust the number 2 – our estimate is not exact enough. So we expect roughly

$$Ce^{cn}$$

triangles along C_r, for some $c > 0$.

To summarize:

> *The number of triangles that are n triangles away from the origin is roughly Ce^{cn} for large n.*

That is, hyperbolic space grows exponentially fast.

For example, in the actual figure, if I go out from the origin by 4 concatenated blue triangles, it appears, by inspection, that there are around 100 blue triangles circling the origin at that distance.

And each time we add 1 triangle to the radius, we multiply the number of triangles around the circumference by a constant.

That is a lot.

9

Geodesics

§23 Minimizing curves and geodesics

We define

- Minimizing curves
- Locally minimizing curves
- Geodesics

Let X be any metric space whose metric is obtained by taking the infimum of path-lengths.

Definition 23.1 A curve $\gamma : [a, b] \to X$ is *length-minimizing* or *minimizing* if

$$L_X(\gamma) = \text{dist}_X(\gamma(a), \gamma(b)).$$

The word "minimizing" is used because, by definition,

$$\text{dist}_X(x, y) = \inf\{L_X(\gamma) : \gamma \text{ is a suitable curve connecting } x \text{ to } y\}.$$

So the definition says that γ minimizes the length, and realizes the infimum.

If γ is minimizing, it follows that any sub-curve

$$\beta = \gamma \,|\, [c, d], \qquad a \le c < d \le b,$$

is minimizing as well.

For if there were a shorter competitor for β, it could be used to construct a shorter competitor for γ.

The definition can be extended to open-ended curves (including infinite curves) $\beta : (a, b) \to X$ by saying that β is length-minimizing if every closed-ended sub-curve $\beta|[c, d]$, $a < c < d < b$, is length-minimizing in the original sense.

If this is true only for short sub-curves, we have the following definition.

Definition 23.2 A curve is *locally length-minimizing* if

a) every interior point of the curve is contained in the interior of a length-minimizing sub-curve.

b) every endpoint of the curve is an endpoint of a length-minimizing sub-curve.

Definition 23.3 A *geodesic* is a locally length-minimizing curve.

This figure from §11 shows minimizing and non-minimizing geodesics.

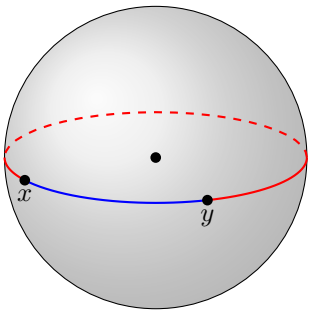

Figure 23.1: The blue path is minimizing. The red one is not.

On the other hand, the many geodesics that join two antipodal points of S^2 are all length-minimizing. See Figure 11.6.

A geodesic is allowed to cross itself, but as soon as it does so, it is not minimizing.

There is an equivalent definition of a geodesic that uses an ODE (ordinary differential equation). The ODE says: the curve turns neither to the right or to the left. The ODE is studied in differential geometry. The ODE only works if the space is smooth enough. It would not work in the Sierpinski gasket, for example, nor in the L^1 metric on $\mathbb{R}^n$. We don't need the ODE definition for now. See J. M. Lee, *Introduction to Riemannian Manifolds*, 2018, Theorem 4.27. (A great book.)

§24 The x-axis is a minimizing geodesic

Write

$$A := B_1 \cap (x\text{-axis}).$$

Theorem 24.1 *A is a minimizing geodesic in $\mathbb{H}^2$.*

Broadly speaking, this follows from the fact that the length stretch factor

$$g_H(z) = \frac{2}{1-|z|^2}$$

is a convex function of z and is symmetric under the reflection $z \mapsto \bar{z}$. But we will prove it concretely.

Proof 1. Let $-1 < a < b < 1$ be two points in A. Let

$$\gamma(t) = (x(t), 0) = (a + (b-a)t, 0), \qquad 0 \le t \le 1$$

be the interval from a to b along the x-axis. Let $\beta : [0,1] \to B^1$ given by

$$\beta(t) = (u(t), v(t)), \qquad 0 \le t \le 1,$$

be any other curve connecting a to b.

We must show that

$$L_H(\beta) \ge L_H(\gamma),$$

for any such β.

2. We will show that β gets shorter if it is projected to the x axis. That is because orthogonal projection decreases the Euclidean length, and also decreases the hyperbolic length factor.

The curve

$$\tilde{\beta}(t) := (u(t), 0), \qquad 0 \le t \le 1,$$

is the orthogonal projection of β to the x-axis.

Then

$$\begin{aligned} L_H(\beta) &= \int_0^1 g_H(\beta(t)) \sqrt{\left(\frac{du}{dt}\right)^2 + \left(\frac{dv}{dt}\right)^2}\, dt \\ &= \int_0^1 \frac{2}{1-u(t)^2-v(t)^2} \sqrt{\left(\frac{du}{dt}\right)^2 + \left(\frac{dv}{dt}\right)^2}\, dt \\ &\ge \int_0^1 \frac{2}{1-u(t)^2} \left|\frac{du}{dt}\right| dt \\ &= L_H(\tilde{\beta}). \end{aligned}$$

3. So

$$L_H(\beta) \geq L_H(\tilde{\beta}),$$

where $\tilde{\beta}$ stays on the x-axis. It suffices to prove that

$$L_H(\tilde{\beta}) \geq L_H(\gamma).$$

But this is intuitively obvious, because $\tilde{\beta}$ goes from a to b, but may have backtracking, whereas γ goes from a to b without backtracking.

It is more efficient to do it without backtracking. So γ is shorter that $\tilde{\beta}$.

4. Here are the details of the backtracking argument.

We have $\gamma(t) := (x(t), 0)$, $0 \leq t \leq 1$, and

$$L_H(\gamma) = \int_0^1 \frac{2}{1-x(t)^2} \left|\frac{dx}{dt}\right| dt,$$

whereas $\tilde{\beta}(t) := (u(t), 0)$, $0 \leq t \leq 1$, and

$$L_H(\tilde{\beta}) = \int_0^1 \frac{2}{1-u(t)^2} \left|\frac{du}{dt}\right| dt.$$

Both $x(t)$ and $u(t)$ accomplish the same journey, that is,

$$x(0) = u(0) = a, \qquad x(1) = u(1) = b. \tag{24.1}$$

The difference is that x increases the whole way ($dx/dt > 0$), whereas u can meander. That makes x more efficient than u.

We get

$$\begin{aligned}
L_H(\gamma) &= \int_0^1 \frac{2}{1-x(t)^2} \left|\frac{dx}{dt}\right| dt \\
&= \int_0^1 \frac{2}{1-x(t)^2} \frac{dx}{dt} dt \qquad \text{since } dx/dt > 0 \\
&= \int_a^b \frac{2}{1-x^2} dx \qquad \text{using (24.1)} \\
&= \int_a^b \frac{2}{1-u^2} du \\
&= \int_0^1 \frac{2}{1-u(t)^2} \frac{du}{dt} dt \qquad \text{using (24.1)} \\
&\leq \int_0^1 \frac{2}{1-u(t)^2} \left|\frac{du}{dt}\right| dt \\
&= L_H(\tilde{\beta})
\end{aligned}$$

The "waste" is accounted for by the inequality in the 6th line. This proves the result.

□

§25 Length along the x-axis

Ultimately, we want to calculate the hyperbolic distance between any two points z, w in $\mathbb{H}^2$. But this will take some time, and we can't do it today.

Today, we will calculate the hyperbolic distance between two points a, b in the x-axis. Write

$$A := \mathbb{H}^2 \cap (x\text{-axis}).$$

Recall that A is a minimizing geodesic.

We obtain the following formula:

Theorem 25.1 *Let a, b be points in A with $a < b$. Then*

$$d_H(a,b) = \log \frac{(1-a)(1+b)}{(1+a)(1-b)}.$$

Note that the right-hand side is positive since $a < b$. If we took the points in the other order, we would have to flip the expression to get a positive value.

The proof is easy. The fact that A is a minimizing geodesic (Theorem 52.1) is essential.

Proof Because A is a minimizing geodesic, we have

$$d_H(0,b) = L_H([0,b]).$$

Now along A, we have Euclidean arclength

$$ds_E = dx$$

and hyperbolic arclength

$$ds_H = f(x)\,dx = \frac{2}{1-x^2}dx$$

We integrate from $x = a$ to $x = b$ to get

$$\begin{aligned} d_H(a,b) &= L_H([a,b]) \\ &= \int_a^b ds_H \\ &= \int_a^b \frac{2}{1-x^2}dx \\ &= \int_a^b \frac{1}{1+x} + \frac{1}{1-x}dx \\ &= [\log(1+x) - \log(1-x)]_{x=a}^{x=b} \\ &= \log(1+b) - \log(1-b) - (\log(1+a) - \log(1-a)) \\ &= \log \frac{(1-a)(1+b)}{(1+a)(1-b)}. \end{aligned}$$

□

Distance to infinity

Take $a = 0$ and get for $b > 0$

$$d_H(0, b) = \log\left(\frac{1+b}{1-b}\right).$$

Indeed, by rotational symmetry, we get for any $z \in \mathbb{H}^2$,

$$d_H(0, z) = \log\left(\frac{1+|z|}{1-|z|}\right).$$

As a result,

$$d_H(0, z) \to \infty \qquad \text{as } |z| \to 1.$$

So the hyperbolic distance to the edge of B_1 is infinite.

Exercise 25.1 *Prove that any path that goes to the edge of B_1 has infinite hyperbolic length.*

§26 Geodesics in $\mathbb{H}^2$

References

- Weeks, Chapter 10.
- Hitchmann, section 5.2.
- Anderson, 2-3, 188 (but he does the upper half-space model first).
- Loustau, 12-16.

We say what all the geodesics of $\mathbb{H}^2$ are.

Recall that the x-axis is a minimizing geodesic (proven above). It follows that every Euclidean straight line through 0 is a minimizing geodesic.

Theorem 26.1 *The geodesics of $\mathbb{H}^2$ are precisely the portions of lines and circles in B_1 that meet ∂B_1 orthogonally. They are all minimizing.*

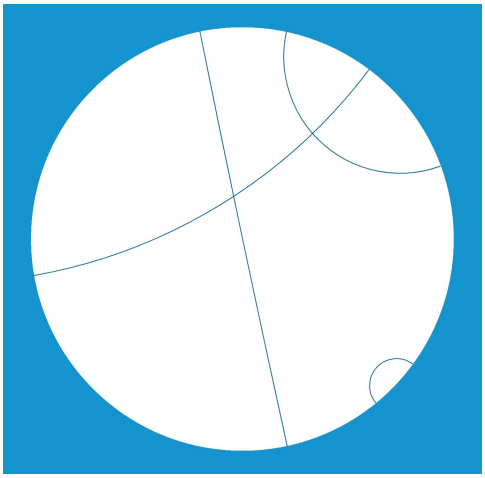

Figure 26.1: Hyperbolic geodesics (made with A. Zampa's Geogebra applet)

The proof will be given much later (§52). We will begin using this theorem immediately.

Euclid's parallel axiom

Geodesics are called parallel if they don't meet.[1]

[1]Some authors define parallel, in the hyperbolic case, to mean they have a common endpoint at infinity (see below). We won't do this.

Now, Euclid's geometry has five axioms (plus other assumptions that were later recognized to be axioms). See Loustau, pp. 12-16 for a nice discussion.

The fifth axiom (the famous *Parallel Postulate*) says "through each point not on a line there is exactly one parallel line".

This is true in $\mathbb{R}^2$ but false in S^2 and $\mathbb{H}^2$.

In S^2, there are no parallel geodesics.

In $\mathbb{H}^2$, there are infinitely many parallel geodesics.

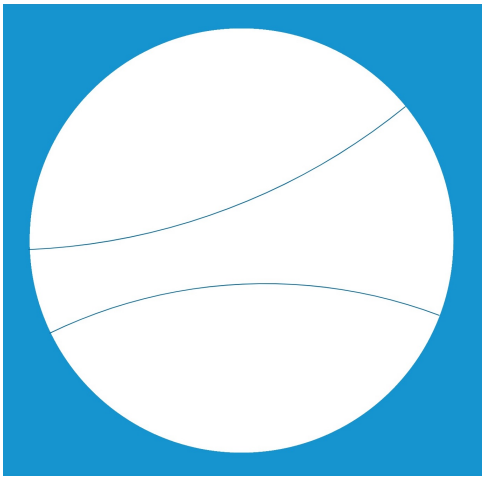

Figure 26.2: Many parallels through a given point (made with A. Zampa's Geogebra applet)

The other four Euclidean axioms are true in all three spaces.

Two kinds of parallel geodesic

Let β, γ be parallel geodesics in $\mathbb{H}^2$. They can be parallel in two ways.

If β and γ have a common endpoint on B_1 as γ, we say that β and γ are *limiting parallel*.

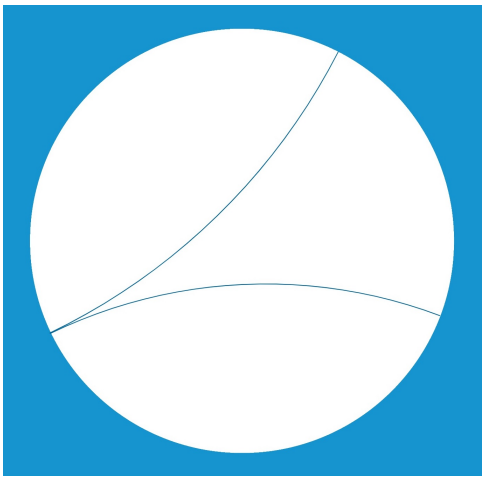

Figure 26.3: Limiting parallel (made with A. Zampa's Geogebra applet)

Otherwise, we say that β and γ are *ultraparallel*.

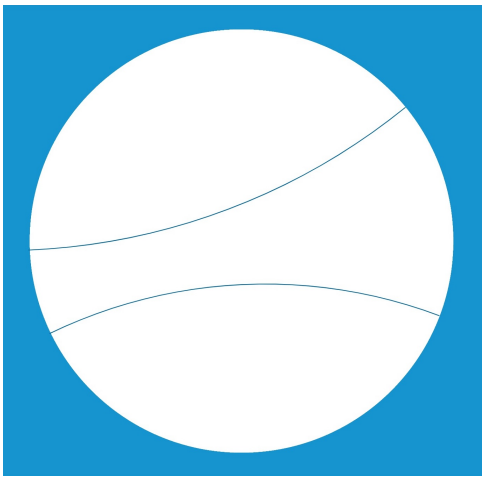

Figure 26.4: Ultraparallel (made with A. Zampa's Geogebra applet)

Since the intersection point at infinity does not lie in $\mathbb{H}^2$, it does not yield a measurement that ants living in $\mathbb{H}^2$ could make. It would be nice to have a characterization of "tangent at infinity" and "ultraparallel" that is instrinsic to hyperbolic space.

Exercise 26.1

(a) Show that ultraparallel geodesics move infinitely far apart at infinity.

(b) Show that limiting parallel geodesics approach each other exponentially at infinity.

Exercise 26.2 *A sportsman in the hyperbolic plane has a double-barreled shotgun. The two barrels are 0.01 unit apart and both are orthogonal to a bar that connects them.*

a) Find the distance apart of the two bullet trajectories as a function of distance. What is its growth rate?

b) Is it very easy or very hard to aim in hyperbolic space?

c) What is the easiest target to hit in the sphere?

10

Circumference and area of a hyperbolic disk

§27 Other expressions for the length

Recall

$$A = B_1 \cap (x\text{-axis}).$$

Proposition 27.1 *Let a, b be points in A, $-1 < a < b < 1$. Then*

$$d_H(a,b) = \log \frac{(1-a)(1+b)}{(1+a)(1-b)} \tag{27.1}$$

$$= 2\operatorname{arctanh}(b) - 2\operatorname{arctanh}(a) \tag{27.2}$$

$$= \operatorname{arccosh}\left(1 + \frac{2(a-b)^2}{(1-a^2)(1-b^2)}\right). \tag{27.3}$$

$$= \operatorname{arcsinh}\left(\frac{2|a-b|\sqrt{1+a^2b^2-2ab}}{(1-a^2)(1-b^2)}\right). \tag{27.4}$$

We leave the proof as an exercise. It is a routine algebraic manipulation of the definitions of arctanh, arccosh and arcsinh.

The usefulness of (27.1) is that it is easy to compute with.

The usefulness of (27.2) is that it is easy to visualize.

The usefulness of (27.3) is that it is analogous to the formula for spherical distance that we derived in §17.

The usefulness of (27.4) is to calculate the hyperbolic circumference of a circle in the next section.

So they are all useful!

Comparison to spherical distance

To see the analogy involving (27.3), recall the formula for spherical distance from Proposition 17.1, namely

$$d_S(z,w) = \arccos\left(1 - \frac{2|z-w|^2}{(1+|z|^2)(1+|w|^2)}\right), \qquad z,w \in \mathbb{C}.$$

Specialize this to the case where a, b lie in the x-axis. Obtain

$$d_S(a,b) = \arccos\left(1 - \frac{2(a-b)^2}{(1+a^2)(1+b^2)}\right), \qquad a,b \in x\text{-axis}.$$

This differs from (27.3) purely by reversing some signs!

§28 Circumference and area of a hyperbolic disk

Let's derive a relation between circumference and area of a hyperbolic disk. We restate the relationship we mentioned in §12 – but now as a theorem.

Theorem 28.1 *Let K_r be a circle of hyperbolic radius with center 0. Then it has hyperbolic circumference*

$$C_H(r) = 2\pi \sinh(r),$$

and hyperbolic area

$$A_H(r) = 2\pi(\cosh(r) - 1).$$

So the circumference and area grow exponentially as $r \to \infty$, as we previously said several times.

Exercise 28.1 *Argue that it is easy to get lost in hyprbolic space.*

Proof 1. By rotational symmetry of the hyperbolic metric about 0, a hyperbolic circle with center 0 is also a Euclidean circle with center 0, it's just that the hyperbolic radius is not the same as the Euclidean radius.

By Proposition 27.1 with $a = 0$, these two radii are related by

$$r = \operatorname{arcsinh}\left(\frac{2b}{1-b^2}\right),$$

where

$$r = \text{hyperbolic radius}, \qquad b = \text{Euclidean radius}.$$

So

$$\frac{2b}{1-b^2} = \sinh(r).$$

Now calculate

$$\begin{aligned} C_H(r) &= C_E(b)\, g_H(b) && \text{(Euclidean circumference} \times \text{stretch factor)} \\ &= (2\pi b)\left(\frac{2}{1-b^2}\right) \\ &= \frac{4\pi b}{1-b^2} \\ &= 2\pi \sinh(r), && \text{from above} \end{aligned}$$

as desired.

2. Now we integrate this to get the area $A_H(r)$. It's just like we did for the spherical area $A_S(r)$ in §12.

We fill the region between 0 and K_r by "parallel" circles

$$K_s, \qquad 0 < s \leq r.$$

The circles are equidistant from each other, and the distance between K_s and K_t is $|t - s|$. So we can compute the hyperbolic area of the enclosed disk by

$$\begin{aligned} A_H(r) &= \int_0^r L_H(K_s)\, ds \\ &= \int_0^r C_H(s)\, ds \\ &= \int_0^r 2\pi \sinh(s) \\ &= (2\pi \cosh(s)]_{s=0}^{s=r} \\ &= 2\pi \cosh(r) - 2\pi \cos(0) \\ &= 2\pi(\cosh(r) - 1). \end{aligned}$$

So the hyperbolic area of the disk is

$$A_H(r) = 2\pi(\cosh(r) - 1), \qquad r > 0,$$

as required.

□

§29 Visualization and resources

There are a number of visualization apps, a new book, and an upcoming event at ETH.

Hyperrogue

A great way to get intuition for hyperbolic space – especially its hugeness and lostness – is the Hyperrogue program, available at

- Hyperrogue: `https://roguetemple.com/z/hyper/`

You can play in the browser but it's much better to download the app. It's easy to get into, but under the surface it has extremely diverse features and many setting – a lot to explore.

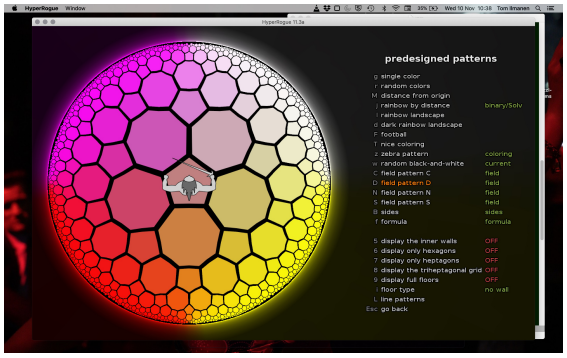

Figure 29.1: Hyperrogue (screenshot)

ZenoTheRogue

The same group that made Hyperrogue have produced a few dozen geometry videos on various topics including hyperbolic geometry. They can be found on Youtube.

- ZenoTheRogue :

  `https://www.youtube.com/channel/UCfCtbgiDxwFtlqrbEralvTw`

Geogebra

An applet for hyperbolic geometry can be found on the Geogebra website. You can use it to do ruler-and-compass constructions in hyperbolic geometry.[1]

[1] Zirkel und Lineal.

- Hyperbolic geometry on Geogebra:

  `https://www.geogebra.org/classic/tHvDKWdC`

Here is a screenshot.

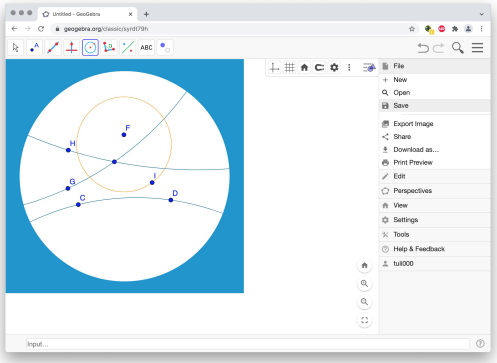

Figure 29.2: Geogebra (screenshot)

The general Geogebra website is:

- Geogebra: `https://www.geogebra.org`

It has applets for all different topics, including several for hyperbolic geometry. The one we've selected may be the best. There is also a "general" app where you can write your own applets.

Hitchman book

I want to recommend the online book by Michael Hitchman, which I just found. It is called *Geometry with an Introduction to Cosmic Topology*. It is about hyperbolic geometry, elliptic geometry, and cosmic topology.

- Hitchman book: `https://mphitchman.com/geometry/frontmatter.html`

Hyperbolic billiards

In addition to these, I had a great experience the other day. Jeff Weeks was in town and demonstrated his virtual reality software for hyperbolic billiards.

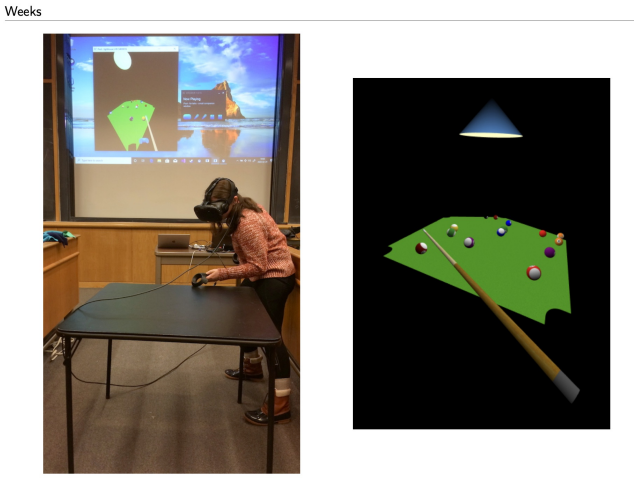

Figure 29.3: Hyperbolic billiards VR system (J. Weeks paper)

You can do billiards either in hyperbolic space, spherical space, or the torus. For example, in hyperbolic space there exists a polygon with

5 sides, 5 right angles.

Here is a picture:

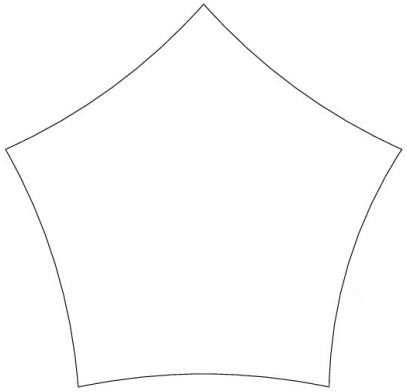

Figure 29.4: Pentagon with five right angles (Lixin Liu)

Such a pentagon is impossible in $\mathbb{R}^2$, but exists in $\mathbb{H}^2$. A regular pentagon in $\mathbb{R}^2$ has angle sum

$$\alpha + \alpha + \alpha + \alpha + \alpha = (5-2)\pi = 3\pi.$$

But this hyperbolic pentagon with 5 right angles has angle sum

$$\alpha + \alpha + \alpha + \alpha + \alpha = \pi/2 + \pi/2 + \pi/2 + \pi/2 + \pi/2 = (5/2)\pi.$$

Therefore it has angle defect

$$3\pi - (5/2)\pi = \pi/2.$$

Since the angle defect formula

$$(\text{angle defect}) = (\text{area})$$

is valid for any convex polygon in hyperbolic space (check!), we find that the pool table has area

$$A(P) = \pi/2.$$

Weeks explains the VR billiards system in

- J. Weeks, *Non-Euclidean Billiards in VR*

  `http://archive.bridgesmathart.org/2020/bridges2020-1.pdf`.

Kaleidotile

If we repeat the billiards table infinitely often in $\mathbb{H}^2$, we get a tiling[2] of $\mathbb{H}^2$ by 90-degree pentagon. 5-sided figure meet 4 to a corner, so it's called a $(5, 4)$ tiling.

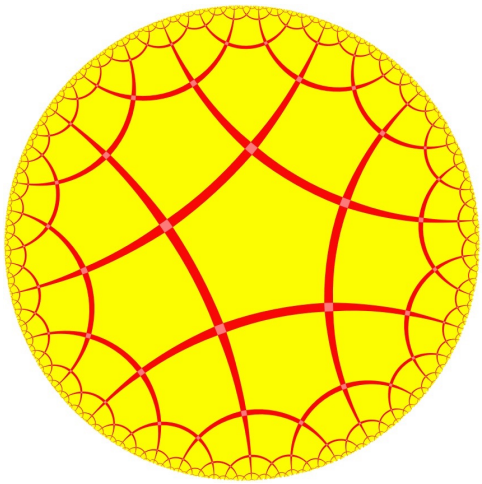

Figure 29.5: Hyperbolic $(5, 4)$ tiling (made with Kaleidotile)

Here is a $(7, 3)$ tiling.

[2] Tesselation, Parkettierung.

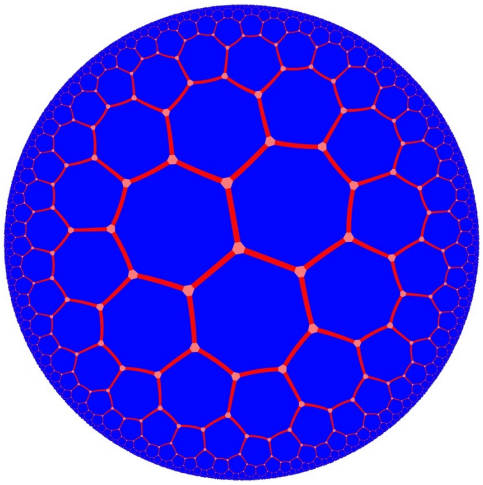

Figure 29.6: Hyperbolic $(7, 3)$ tiling (made with Kaleidotile)

Here is a more complicated kind of tiling. It has the same $(2, 3, 7)$ symmetry of $\mathbb{H}^2$ as the $(7, 3)$ tiling.

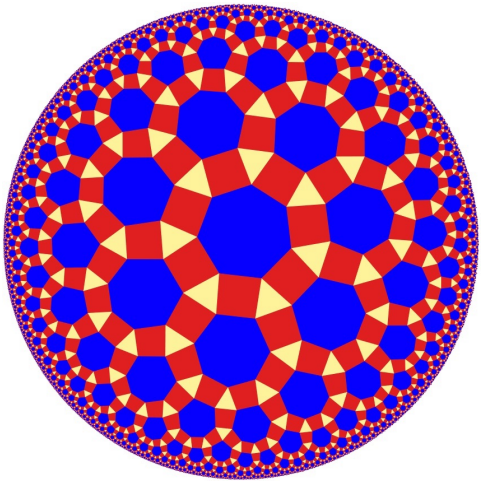

Figure 29.7: Hyperbolic tiling with $(2, 3, 7)$ symmetry (made with Kaleidotile)

These are produced by Kaleidotile of Jeff Weeks. It can be downloaded from

- Kaleidotile: `https://www.geometrygames.org/KaleidoTile`

Let's take a look at it on my computer. You can adjust the colors. You can move the control point in a 2-dimensional space of possibilities. This varies the geometry without changing the symmetry group.

You can move around in hyperbolic space with the mouse. It has momentum.

The program is very fast and smooth.

You can change the symmetry group. If you do that, the tiling will close up differently to produce either a hyperbolic space, a Euclidean plane, or a sphere.

Gomath

From 14 to 25 March 2022, there will be a 2 weeks math exhibition in the ETH main building called *The Shape of Space*. Jeff Weeks will be demonstrating his software (in particular the Hyperbolic Billiards VR experience) and giving talks, and other activities. See

- Gomath:
  `https://math.ethz.ch/news-and-events/events/gomath/gomath-2022.html`

11

The extended complex plane and clines

§30 The extended complex plane and the Riemann sphere

Recall that stereographic projection is a bijection

$$\sigma : S^2 \setminus \{N\} \to \mathbb{C}.$$

Inspired by this, we find it convenient to add a "point at infinity", called ∞, to the complex plane to produce the *extended complex plane*[1]

$$\hat{\mathbb{C}} := \mathbb{C} \cup \{\infty\}.$$

We then define a bijection

$$\hat{\sigma} : S^2 \to \hat{\mathbb{C}}$$

by

$$\hat{\sigma}(P) = \begin{cases} \sigma(P) & P \in S^2 \setminus \{N\} \\ \infty & P = N. \end{cases}$$

In this context, we call S^2 the *Riemann sphere*[2]

We define the bijection

$$\hat{\tau} : \hat{\mathbb{C}} \to S^2$$

as the inverse of $\hat{\sigma}$. It takes ∞ to $N = (0, 0, 1)$.

Higher dimensions

A similar procedure can be done in higher dimensions. There is higher-dimensional stereographic projection, an extended space $\widehat{\mathbb{R}^n} = \mathbb{R}^n \cup \{\infty\}$, and a theory of higher dimensional conformal maps. But we won't do this.

[1] Erweiterte komplexe Ebene.
[2] Riemannscher Zahlenkugel.

§31 Clines

We work in the extended complex plane $\hat{\mathbb{C}}$.

Definition 31.1

a) An *extended line*[3], is a line together with the point ∞.

b) A *cline*,[4] or *generalized circle*, is either a circle or an extended line.

The idea is that as a circle gets larger and larger, it can converge to an extended line. To complete the collection of circles, we need to include the extended lines.

When working in $\hat{\mathbb{C}}$, we will usually refer to an extended line simply as a line.

Let $\hat{\sigma} : S^2 \to \hat{\mathbb{C}}$ be the correspondence of the Riemann sphere to the extended complex plane.

Theorem 31.2 *Under $\hat{\sigma}$, circles in S^2 correspond to clines in $\hat{\mathbb{C}}$.*

In particular:

(a) Circles in S^2 that pass through N correspond to extended lines in $\hat{\mathbb{C}}$.

(b) Circles in S^2 that don't pass through N correspond to circles in $\mathbb{C}$.

Proof Let C be a circle in S^2. Then $C = S^2 \cap P$, where P is a plane in $\mathbb{R}^3$

(a) This is really easy. If P passes through N, then σ takes $C \setminus \{N\}$ to the line

$$L := P \cap \mathbb{C}.$$

So $\hat{\sigma}$ takes C to the extended line $\hat{L}$.

(b) Let C be a circle in S^2 that doesn't pass through N. So $\hat{\sigma}(C) = \sigma(C)$.

We can visualize $\sigma(C)$ by drawing all the lines that pass through N and an arbitrary point in C. They meet $\mathbb{C}$ in $\sigma(C)$. The union of these lines is a kind of cone.

In order to compute what $\sigma(C)$ is, we write the equation of P:

$$P: \qquad aX + bY + cZ = d,$$

where we must require that P meets S^2, P is not tangent to S^2, and $P \not\ni N$. This can be effected by

$$a^2 + b^2 + c^2 = 1, \qquad d < 1, \qquad c \neq d.$$

Then the equation of $\sigma(C)$ is

$$C: \qquad aX + bY + cZ = d, \quad X^2 + Y^2 + Z^2 = 1.$$

[3] Erweiterte Gerade.

[4] V-Kreis, oder verallgemeinerter Kreis.

We substitute $P = \tau(z) = \tau(x, y)$ into this to find the equation of $\tau(C)$. The equation $X^2 + Y^2 + Z^2 = 1$ is automatically fulfilled, so we get

$$\sigma(C): \qquad (a, b, c) \cdot \tau(x, y) = d.$$

which represents the intersection of the "cone" with $\mathbb{C}$. This becomes

$$(a, b, c) \cdot \frac{(2x, 2y, |z|^2 - 1)}{|z|^2 + 1} = d.$$

$$(a, b, c) \cdot (2x, 2y, x^2 + y^2 - 1) = d(x^2 + y^2 + 1).$$

$$2ax + 2by + c(x^2 + y^2 - 1) = d(x^2 + y^2 + 1).$$

We get the equation of $\sigma(C)$:

$$\sigma(C): \qquad (c - d)x^2 + 2ax + (c - d)y^2 + 2by = c + d$$

Using the conditions $c \neq d$, $d < 1$, and $a^2 + b^2 + c^2 = 1$, we can verify that the solution set is not empty, is not degenerate, and not just one point. So it is an ellipse. But the coefficients of x^2 and y^2 are the same, so it is a circle. So $\sigma(C)$ is a circle in $\mathbb{C}$.

□

12

Inversion

§32 Inversion in a cline

We will define a new kind of transformation, called *inversion in a cline*. It generalizes reflection in a line.

Later, we will see that inversion in a cline is conformal.

Work in $\hat{\mathbb{C}}$. We start with the unit circle S^1. Define *inversion in the unit circle*

$$J : \hat{\mathbb{C}} \to \hat{\mathbb{C}}$$

as follows.

1) Let $z \neq 0, \infty$. Let $R_z := \{tz : t > 0\}$ be the open ray from 0 through x. Define $J(z)$ to be the point on R_z with

$$|J(z)| = \frac{1}{|z|}.$$

The formula is

$$J(z) = \frac{z}{|z|^2}, \qquad z \neq 0.$$

2) Define $J(0) := \infty$, $J(\infty) := 0$.

This defines $J(z)$ for all $z \in \hat{\mathbb{C}}$.

The map J "flips" points inside S^1 outside, and points outside inside, while taking 0 to ∞ and ∞ to 0.

Relation to complex inverse

Let I be the complex inverse map

$$I(z) = 1/z.$$

Then

$$J(z) = \frac{z}{|z|^2} = \frac{z}{z\bar{z}} = \frac{1}{\bar{z}} = I(C(z)),$$

where $C(z) = \bar{z}$. So

$$J = I \circ C = C \circ I.$$

Also since $C^2 = 1$,

$$I = C \circ J = J \circ C.$$

That is, $1/z$ is inversion in the unit circle followed by conjugation.

Inversion in other circles

The map J fixes each point of the circle S^1 and reverses the inside and the outside. Also $J^2 = J \circ J = \text{id}$. We call a map whose square is the identity an *involution*.

Let K be any circle in $\mathbb{C}$. Let A_K be a similarity transformation that takes S^1 to K.

Define *inversion in* K to be

$$J_K := A_K \circ J \circ (A_K)^{-1}.$$

Note $J = J_{S^1}$. Like J, J_K fixes K, switches the inside and outside of K, has $(J_K)^2 = \text{id}$, and throws the center of K to infinity.

Inversion in extended lines

The mirror reflection across a line L has similar properties to the above. So wew define *inversion in* $\hat{L}$ to be essentially reflection across L:

$$J_{\hat{L}}(z) := \begin{cases} \text{reflection of } z \text{ across } L & z \in \mathbb{C} \\ \infty & z = \infty. \end{cases}$$

The only new piece of information is that $J_{\hat{L}}$ fixes ∞. So $J_{\hat{L}}$ fixes every point in $\hat{L}$.

We now can invert in any cline.

Properties of inversions

To summarize, we have for any cline E,

$$J_E \text{ fixes each point of } E,$$

$$J_E \text{ reverses the inside and outside of } E,$$

$$(J_E)^2 = \text{id}.$$

If E is a circle, then J_E swaps the center of E with ∞, whereas if E is an extended line, then ∞ lies on E and J_E fixes ∞.

§33 Transferring operations between $\hat{\mathbb{C}}$ and S^2

Transformations of $\hat{\mathbb{C}}$

We have defined the following tansformations (bijections) of $\hat{\mathbb{C}}$.

General ones:

$$\begin{aligned} M_a &= \text{ multiplication by } a = az \\ T_b &= \text{ translation by } b = z + b \\ J_E &= \text{ inversion in the cline } E \end{aligned}$$

Special ones:

$$\begin{aligned} I &= \text{ complex inverse } = \frac{1}{z} \\ C &= \text{ complex conjugation } = \bar{z} = J_{x\text{-axis}} \\ J &= \text{ inversion in } S^1 = J_{S^1} \end{aligned}$$

Orientation-preserving: M_a, T_b, I

Orientation-reversing: J_E, C, J

Similarities of $\hat{\mathbb{C}}$: $M_a, T_b, J_{\hat{L}}$ ($\hat{L}$ an extended line)

Corresponding transformations of S^2

Given a bijection f of $\hat{\mathbb{C}}$, we get a bijection $\tilde{f}$ of S^2 by

$$\tilde{f} := \sigma^{-1} \circ f \circ \sigma.$$

Sometimes we write W_f for $\tilde{f}$.

So we have $\tilde{M_a}$, $\tilde{T_b}$, $\tilde{J_E}$, $\tilde{I}$, $\tilde{J}$, $\tilde{C}$. What do they look like?

Here are some important examples:

1) Rotations about 0 become rotations of S^2 about the Z-axis. These (and others that look like them) are called *elliptic*.

Exercise 33.1 *What becomes of rotations about other points b in $\mathbb{C}$?*

2) Homothetic expansion of $\hat{\mathbb{C}}$ by a factor $\lambda > 0$ becomes an operation that moves all the points of S^2 along trajectories connecting the south pole to the

north pole. This creates an interesting new operation on S^2. These (and others that look like them) are called *hyperbolic*.

Exercise 33.2 *What is the effect on S^2 of λ-expansion about other points b in $\mathbb{C}$?*

Exercise 33.3 *What if we combine rotation and expansion? What does that look like on S^2?*

3) Translation of $\hat{\mathbb{C}}$ becomes a strange new operation on S^2 that fixes N and moves the rest of the points "sideways", with an interesting pattern of movement near N. These (or anything that looks like them) are called *parabolic*.

Exercise 33.4 *Study the pattern of movement of the transformation $\tilde{T}_b$ near N by moving N to a place where we can study it by formula. Namely, let $\rho : S^2 \to \hat{\mathbb{C}}$ be stereographic projection from the south pole, and let U_b be the transformation of $\mathbb{C}$ that corresponds to $\tilde{T}_b$ via ρ.*

(a) Verify that

$$U_b = \frac{z}{1+bz}.$$

(b) Let $b > 0$ and study the pattern made by this map near 0. You may use the fact that U_b takes clines to clines (indeed, you'll be able to prove his after the next section).

(c) In particular, if we fix $z \neq 0$, ehat does the orbit

$$\{U_b(z) : 0 < b < \infty\}$$

look like?

4) Inversion in S^1 of $\hat{\mathbb{C}}$ (the map J) becomes reflection of S^2 across the XY-plane. This is interesting because a nonlinear operation becomes an easy-to-understand linear operation.

Exercise 33.5

(a) Describe the operation of $\tilde{J}, C, I$ on S^2. It is remarkably simple

(b) The set $\{\mathrm{id}, I, J, C\}$ is closed under composition and taking the inverse function. Such a set of bijections is called a transformation group. This one is the famous Klein 4-group, with four elements.

Exercise 33.6 *Describe the effect of inversion in other circles of $\mathbb{C}$.*

§34 Inversion takes clines to clines and is conformal

Theorem 34.1

a) Inversion in a cline takes clines to clines.

b) Inversion in a cline is conformal.

Proof 1. If we invert in an extended line, then it is mirror reflection, so it obviously takes clines to clines and is conformal.

2. If we invert in a circle K, then the inversion can be written as

$$J_K := A \circ J \circ A^{-1},$$

where A is a similarity transformation and J is inversion in the unit circle. Now A and A^{-1} take clines to clines and are conformal, so it suffices to prove that J takes clines to clines and is conformal.

3. By the Exercise in the previous section,

$$J = \sigma^{-1} \circ S \circ \sigma$$

where S is reflection of S^2 in the XY-plane.

By Theorem 31.2, σ^{-1} takes clines to circles. Clearly S takes circles to circles. By Theorem 31.2 again, σ takes circles to clines. So J takes clines to clines.

By Corollary 20.1, σ^{-1} is conformal. Clearly S is conformal. By Corollary 20.1 again, σ is conformal. So J is conformal.

□

Remark 1. That J takes clines to clines can also be proven by a direct calculation analogous to the proof of Theorem 31.2. One substitutes $z = 1/\bar{w}$ in the equation of a circle, resulting in a new circle after some calculation. Lines must be treated as special cases.

Remark 2. That J is conformal can also be proven by a direct calculation analogous to the proof of Theorem 19.1 and Corollary 20.1. One shows that J is nearly a similarity transform on small open sets, which implies conformality.

This computation is done in the next section, for I. Use the formula $J = C \circ I$ to apply it to J.

Remark 3. It is no coincidence that the two statements have essentially the same proof. For taking clines to clines automatically implies conformality, at least heuristically, as follows.

If a map f preserves clines, then in particular f takes small circles centered at z to small circles around $f(z)$ (not necessarily centered at $f(z)$).

But that implies (or strongly suggests) that near z, f stretches nearly isotropically – the same in all directions. That is, f is nearly a similarity transformation on a small neighborhood. So f is conformal.

In fact, this can be taken as a (heuristic) *definition* of conformality: taking tiny round circles to tiny nearly round small circles. (Rather than ellipses!)

§35 The stretch factor of $1/z$

I didn't lecture on this

Write $Sh(z)$ for the stretch factor of a conformal map h at the point z. (This is not standard notation.)

Theorem 35.1 *$I(z) = 1/z$ is conformal for $z \neq 0, \infty$, with stretch factor*

$$Sf(z) = 1/|z|^2.$$

This result says that large places shrink and small places grow, as we expect from the action of $I(z) = 1/z$. $1/z$ takes large numbers to small numbers and vice versa.

Proof 1. Conformality of I follows by factoring

$$I = C \circ J$$

and observing that J is conformal by Theorem 34.1.

2. We will now give a second, computational proof of conformality that also gives us the stretch factor SI of I. It is similar to the proof of Theorem 19.1 and Corollary 20.1, so we will only sketch it without bothering with O-notation.

Let $z \neq 0$. Assume w is very close to z (much closer to z than to zero).

What is $I(w) - I(z)$ as a function of $w - z$ (roughly)?

Calculate

$$\begin{aligned} \delta I &= I(w) - I(z) \\ &= \frac{1}{w} - \frac{1}{z} \\ &= \frac{z - w}{wz} \\ &= -\frac{w - z}{wz} \\ &\approx -\frac{1}{z^2}(w - z), \end{aligned}$$

since $w \approx z$.

That is, when w and z are really close, the transition from

$$w - z \qquad \text{to} \qquad I(w) - I(z)$$

is nearly multiplication by

$$-\frac{1}{z^2}.$$

But multiplication by $-1/z^2$ is a similarity transformation. So this gives a computational proof that I is conformal.

3. Then

$$\begin{aligned} |I(w) - I(z)| &\approx \left| -\frac{1}{z^2}(w - z) \right| \\ &= \frac{1}{|z|^2}|w - z|. \end{aligned}$$

So the stretch factor at z is

$$SI(z) = \frac{1}{|z|^2}.$$

The stretch factor is independent of the direction of $w - z$, reflecting the conformality of I.

□

Remark. In terms of complex derivatives, we have proven

$$I'(z) = -1/z^2, \qquad z \neq 0.$$

The stretch factor is given by

$$SI(z) = |I'(z)| = |-1/z^2| = 1/|z|^2, \qquad z \neq 0.$$

13

Möbius transformations

§36 Möbius transformations

References

- Hitchman, @cite.
- Anderson, @cite.
- Ahlfors, 76-88.

Definition 36.1 A *Möbius transformation*, or *fractional linear transformation*,[1] is a function

$$f : \hat{\mathbb{C}} \to \hat{\mathbb{C}}$$

given by

$$f(z) = \frac{az+b}{cz+d}, \qquad z \in \hat{\mathbb{C}}, \qquad \text{(orientation-preserving)}$$

or

$$f(z) = \frac{a\bar{z}+b}{c\bar{z}+d}, \qquad z \in \hat{\mathbb{C}}, \qquad \text{(orientation-reversing)}$$

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

The first kind are orientation-preserving.[2] They include similarities and the complex inverse.

The second kind are orientation reversing.[1] They include inversion in clines.

The set of all Möbius transformation is called Möb. The set of all orientation-preserving Möbius transformation is called Möb_+.

Goals:

We wish to do the following:

@re-order

- Handle ∞ correctly
- Prove they are bijective
- Prove they form a group
- Factor them into elementary transformations
- Prove they are conformal
- Prove they preserve clines

[1] Gebrochene Lineartransformationen.
[2] We leave this to the reader to verify.

§37 Handling ∞ correctly

We wish to define the Möbius tranformations on all of $\hat{\mathbb{C}}$. To do this, we must handle the two special cases

$$z = \infty$$

which ordinarily "doesn't compute", and

$$z = -\frac{d}{c} \qquad \left(\text{resp. } \bar{z} = -\frac{d}{c} \right),$$

which corresponds to the denominator being zero. These cases were passed over in silence above.

The net result is a continuous map f defined on all of $\hat{\mathbb{C}}$, where $\hat{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$ is given the obvious topology.[3]

Here are the recipes. They are just common sense. We crucially use the fact that $ad - bc \neq 0$.

1) First consider the orientation-preserving case $f(z) = (az + b)/(cz + d)$.

Case 1 $c \neq 0$ and $z = \infty$.

We have for $z \neq 0, \infty$,

$$f(z) = \frac{az + b}{cz + d} = \frac{a + b/z}{c + d/z}, \qquad z \neq 0, \infty.$$

This motivates us to write

$$f(\infty) = \frac{a + b/\infty}{c + d/\infty} = \frac{a + 0}{c + 0} = \frac{a}{c},$$

which makes sense because $c \neq 0$. This can be justified by taking the limit $z \to \infty$.

Case 2 $c \neq 0$ and $z = -d/c$.

Write

$$f(-d/c) = \frac{a(-d/c) + b}{c(-d/c) + d} = \frac{-ad + bc}{c \cdot 0} = \infty,$$

which makes sense because $ad - bc \neq 0$.

Case 3 $c = 0$.

We have for $z \neq \infty$,

$$f(z) = \frac{az + b}{d} = \frac{a}{d} z + \frac{b}{d}$$

[3] Ignore this remark if in the first year.

This is a well-defined similarity transformation because $ad - bc \neq 0$, so $a, d \neq 0$. It give a bijection of $\mathbb{C}$ to $\mathbb{C}$.

So it makes sense to write

$$f(\infty) = \frac{a \cdot \infty + b}{d} = \frac{\infty}{d} = \infty.$$

To summarise:

i) When $c \neq 0$, then ∞ and $-d/c$ are distinct, and

$$f(\infty) = \frac{a}{c}, \qquad f\left(-\frac{d}{c}\right) = \infty.$$

ii) When $c = 0$, then $-d/c = \infty$, f is a similarity, and

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

2) For the orientation-reversing case $f(z) = (a\bar{z} + b)/(c\bar{z} + d)$, simply apply the first case to $\bar{z}$, where $\overline{\infty} = \infty$, to derive similar formulas.

§38 Möbius transformations are invertible

@ This will be edited to make the motivation more transparent

Recall that a Möbius transformation has $ad - bc \neq 0$.

Proposition 38.1 *Every Möbius transformation f is bijective, and the inverse is a Möbius transformation.*

1) The inverse of

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

is

$$f^{-1}(z) = \frac{dz-b}{-cz+a}.$$

2) The inverse of

$$f(z) = \frac{a\bar{z}+b}{c\bar{z}+d}$$

is

$$f^{-1}(z) = \frac{\bar{d}\bar{z}-\bar{b}}{-\bar{c}\bar{z}+\bar{a}}.$$

Proof 1. First let f be orientation-preserving, namely

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

We will invert f.

2. Let us motivate the formula. To invert f, solve

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

for z in terms of w. We get

$$\begin{aligned} az+b &= czw+dw \\ az-czw &= dw-b \\ z(a-cw) &= dw-b \\ z &= \frac{dw-b}{-cw+a} \end{aligned}$$

This motivates the formula, but does not prove it. The reason is that we have not been careful when dividing by zero; we cannot be sure that $-cw+a \neq 0$.

Also, we haven't used the condition $ad-bc \neq 0$, so we can't be done yet. In fact, the simple calculation above fails in a very subtle way when $ad-bc=0$.

Exercise 38.1 *Analyze what goes wrong when $ad - bc = 0$. Why isn't this calculation already a proof of invertibility?*

3. To get a complete proof, define $g(w)$ by

$$g(z) := \frac{dw - b}{-cw + a}.$$

We will verify concretely that

$$g \circ f = f \circ g = \mathrm{id}_{\hat{\mathbb{C}}}.$$

Compute for $z \neq -d/c, \infty$, where $-d/c = \infty$ in the case $c = 0$,

$$\begin{aligned} g(f(z)) &= \frac{d\left(\frac{az+b}{cz+d}\right) - b}{-c\left(\frac{az+b}{cz+d}\right) + a} \\ &= \frac{d(az+b) - b(cz+d)}{-c(az+b) + a(cz+d)} \\ &= \frac{(ad-bc)z}{ad-bc} \\ &= z, \end{aligned}$$

where we have used the fact that $ad - bc \neq 0$ in an essential way.

Using the special rules at the two omitted points, we verify

$$g(f(\infty)) = g(a/c) = \infty, \qquad g(f(-d/c)) = g(\infty) = -d/c.$$

This even works in the extra-special case $-d/c = \infty$. So

$$g \circ f = \mathrm{id}_{\hat{\mathbb{C}}}$$

on all of $\hat{\mathbb{C}}$. In a similar way we check that

$$g \circ f = \mathrm{id}_{\hat{\mathbb{C}}}$$

on all of $\hat{\mathbb{C}}$. So f is bijective with inverse g. This verifies the formula.

4. To see that f^{-1} is a Möbius transformation, observe that

$$da - (-b)(-c) = ad - bc \neq 0.$$

5. The orientation-reversing case easily follows from the orientation-preserving case.

If f is orientation-reversing, then $h := f \circ C$ is orientation preserving. Since h is bijective and C is bijective, f is bijective.

To get the formula, compute

$$f^{-1} = (h \circ C^{-1})^{-1} = C \circ h^{-1},$$

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

$$h^{-1}(z) = \frac{dz-b}{-cz+a}$$

$$f^{-1}(z) = \overline{\left(\frac{dz-b}{-cz+a}\right)} = \frac{\bar{d}\bar{z}-\bar{b}}{-\bar{c}\bar{z}+\bar{a}}.$$

□

14

The group of Möbius transformations

§39 Transformation groups

References

- T. Ilmanen, Geometrie 2020, Chap. 6-9 et al., `https://metaphor.ethz.ch/x/2020/hs/401-1511-00L/literatur/script.pdf`.
- D. Saracino, *Abstract Algebra: A First Course*.

We define transformation groups.

Definition 39.1 A *transformation group* is a collection G of bijections of a set X such that

1) $\mathrm{id}_X \in G$,
2) $f, g \in G \Longrightarrow f \circ g \in G$, (closed under composition)
3) $f \in G \Longrightarrow f^{-1} \in G$, (closed under inverses)

See Ilmanen, *Geometrie 2020*, §20. That script is full of information bout symmetry groups.

Our first example:

Proposition 39.2 *Let (X, d) be a metric space. Then* $\mathrm{Isom}(X)$ *is a transformation group.*

This was Exercise 8.1. The proof is obvious.

A set of bijections that preserve something is always a transformation group.

For example, let T be an equilateral triangle in the plane. The set

$$\mathrm{Sym}(T) := \{f : \mathbb{R}^2 \to \mathbb{R}^2 | f \text{ is an isometry of } \mathbb{R}^2, f(T) = T\},$$

is a transformation group with 6 elements. It is called the *symmetry group* of the triangle. See Ilmanen, *Geometrie 2020*, §1, §19.

For another example, the five Platonic solids figures each have a lot of symmetries. Here are three Platonic solids.

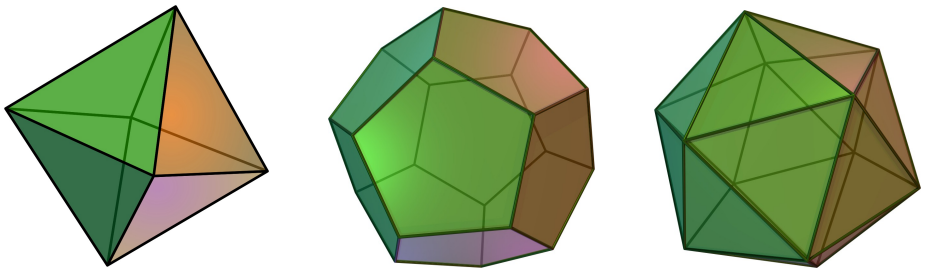

Figure 39.1: Octahedron, Dodecahedron, Icosahedron (Cyp, Wikipedia)

Exercise 39.1 *How many self-isometries do each of these Platonic solids have?*

See Ilmanen, *Geometrie 2020*, §6.

Subgroups

If G is a transformation group and

$$H \subseteq G$$

is a subset, then H is a transformation group in its own right if and only if it is nonempty and closed under composition and inverses. (Check this!) We say in this case that

$$H \text{ is a subgroup of } G,$$

and write

$$H \leq G$$

to indicate this. So $\mathrm{Sym}(T)$ is a subgroup of $\mathrm{Isom}(\mathbb{R}^2)$.

See Ilmanen, *Geometrie 2020*, §26.

Further examples:

$$\mathrm{Sym}(T) \leq \mathrm{Isom}(\mathbb{R}^2),$$

$$\mathrm{Möb}_+ \leq \mathrm{Möb}.$$

The latter will be true once we prove that Möb and $\mathrm{Möb}_+$ are groups.

Isomorphisms

Two groups G, H are called isomorphic if there is a bijection f between them that exactly preserves the group operations and the identity element. To wit:

$$F(gh) = F(g)F(h), \quad F(g^{-1}) = F(g)^{-1}, \quad F(\mathrm{id}_X) = \mathrm{id}_Y,$$

where G acts on X, H acts on Y. It is written

$$G \cong H.$$

The bijection f is called an *isomorphism*. See Ilmanen, *Geometrie 2020*, §21.

Main examples

Our main examples will be

$$\mathrm{Möb}, \quad \mathrm{Möb}_+, \quad \mathrm{Isom}(\mathbb{H}^2), \quad \mathrm{Isom}_+(\mathbb{H}^2),$$

and some subgroups. The + indicates an orientation-preserving subgroup. We will eventually show (or claim)

$$\mathrm{Isom}(\mathbb{H}^2) \cong \mathrm{M\ddot{o}b}(B_1), \qquad \mathrm{Isom}_+(\mathbb{H}^2) \cong \mathrm{M\ddot{o}b}_+(B_1),$$

where

$$\mathrm{M\ddot{o}b}(B_1) := \{f \in \mathrm{M\ddot{o}b} : f(B_1) = B_1\}.$$

Note that

$$\mathrm{M\ddot{o}b}(B_1) \leq \mathrm{M\ddot{o}b}.$$

There is also

$$\mathrm{Conf}(S^2) := \{\text{conformal transformations of } S^2\},$$

and subgroups. It is true that

$$\mathrm{Conf}(S^2) \cong \mathrm{M\ddot{o}b}, \qquad \mathrm{Conf}_+(S^2) \cong \mathrm{M\ddot{o}b}_+,$$

although we will (essentially) prove only one direction of this.

§40 The Möbius transformations form a group

Theorem 40.1 Möb *and* Möb_+ *are groups.*

Proof 1. We have already proven that they are closed inder inverses. Clearly the identity transformation

$$f(z) = \frac{z+0}{0 \cdot z + 1} = z$$

belongs to both.

2. So we must show that each is closed under multiplication. Let us start with Möb_+. Let

$$f(z) = \frac{az+b}{cz+d}, \qquad g(z) = \frac{ez+f}{gz+h}.$$

Then

$$\begin{aligned} f(g(z)) &= \frac{a(ez+f)/(gz+h)+b}{c(ez+f)/(gz+h)+d} \\ &= \frac{a(ez+f)+b(gz+h)}{c(ez+f)+d(gz+h)} \\ &= \frac{(ae+bg)z+(af+bh)}{(ce+dg)z+(cf+dh)} \end{aligned}$$

This has the form of a Möbius transformation, but we have to verify the nonzero "determinant". We get

$$\begin{aligned} &(ae+bg)(cf+dh)-(af+bh)(ce+dg) \\ &\qquad = aecf+aedh+bgcf+bgdh-afce-afdg-bhce-bhdg \\ &\qquad = aedh+bgcf-afdg-bhce \\ &\qquad = (ad-bc)(eh-fg) \\ &\qquad \neq 0. \end{aligned}$$

So

$$f \circ g \in \text{Möb}_+$$

and Möb_+ is a group.

3. For Möb, we have to take account of the various conjugates. All elements of Möb have the form

$$f \quad \text{or} \quad f \circ C$$

where $f \in \text{Möb}_+$. So we have four multiplications to check:

$$f \circ g, \quad f \circ (g \circ C), \quad (f \circ C) \circ g, \quad (f \circ C) \circ (g \circ C),$$

where $f, g \in \text{Möb}_+$.

By 2., $f \circ g \in \text{Möb}_+$, so the first two products lie in Möb. For the third and fourth, we take advantage of the identity

$$C \circ g = \tilde{g} \circ C,$$

where $\tilde{g}$ is the element

$$\tilde{g}(z) := \frac{\bar{a}z + \bar{b}}{\bar{c}z + \bar{d}}$$

when $g(z) = (az + b)/cz + d)$. So for the third product,

$$(f \circ C) \circ g = (f \circ \tilde{g}) \circ C \in \text{Möb}$$

and similarly for the fourth product. So Möb is closed under multiplication. So Möb is a group.

□

§41 Matrix multiplication and $\mathrm{Möb}_+$

The reader may have noticed the similarity of the composition rule for orientation-preserving Möbius transformations and matrix multiplixcation.

Let

$$F = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, \qquad G = \begin{bmatrix} e & f \\ g & h \end{bmatrix}.$$

From these, define

$$f(z) = \frac{az+b}{cz+d}, \qquad g(z) = \frac{ez+f}{gz+h}.$$

Then we calculate

$$FG = \begin{bmatrix} ae+bg & af+bh \\ ce+dg & cf+dh \end{bmatrix}.$$

and (see the formula in Step 2. of the proof above)

$$(f \circ g)(z) = \frac{(ae+bg)z + (af+bh)}{(ce+dg)z + (cf+dh)}.$$

Comparing these formulas, we see that multiplication of F and G implements composition of f and g.

Similiarly, matrix inversion

$$F^{-1} = \frac{\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}}{ad-bc} = \begin{bmatrix} d/(ad-bc) & -b/(ad-bc) \\ -c/(ad-bc) & a/(ad-bc) \end{bmatrix}.$$

yields the inverse of the map f, given by

$$f^{-1}(z) = \frac{dz-b}{-cz+a}$$

Note that the denominators go away when we turn F^{-1} into a Möbius transformation.

We define a map

$$U : F = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \to f(z) = \frac{az+b}{cz+d}.$$

This is a map

$$U : GL_2(\mathbb{C}) \to \mathrm{Möb}_+,$$

where

$$GL_2(\mathbb{C}) := \left\{ \begin{bmatrix} a & b \\ c & d \end{bmatrix} : ad-bc \neq 0 \right\}$$

is the *general linear group* of invertible complex 2x2 matrices, a transformation group of $\mathbb{C}^2$.

As we have seen above, U preserves multiplication and inverses. It also preserves identity element. Such a map between groups is called a *homomorphism*. For a homomorphism, we don't require bijectivity.

U is obviously surjective.

But U is *not* injective.

Indeed, we observe that

$$F = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

and

$$\lambda F = \begin{bmatrix} \lambda a & \lambda b \\ \lambda c & \lambda d \end{bmatrix}$$

yield the same Möbius transformation, because

$$\frac{\lambda a z + \lambda b}{\lambda c z + \lambda d} = \frac{az + b}{cz + d}.$$

Remark: Note that the representation as matrices works only for orientation-preserving Möbius transformations.

Exercise 41.1 *Express the following Möbius transformations as matrices:*

- *identity, translation, rotations, similarities, conjugation, complex inverse, inversion in* S^1

Expression as a quotient group

For those who know group theory, we go a little further.

Since F, λF go to the same element (when $\lambda \neq 0$), we have

$$\begin{aligned} \ker(U) &:= \{F \in GL_2(\mathbb{C}) : U(F) = \mathrm{id}_{\hat{\mathbb{C}}}\} \\ &= \mathbb{C}^* \cdot \mathrm{id}, \end{aligned}$$

where $\mathbb{C}^* := \mathbb{C} \setminus \{0\}$. It follows that Möb is the quotient group given by

$$\boxed{\mathrm{Möb}_+ \cong GL_2(\mathbb{C})/(\mathbb{C}^* \cdot \mathrm{id}).}$$

This expresses, in symbolic form, the ambiguity in choosing a matrix F to represent f.

In order to get rid of the ambiguity, we usually normalize by requiring that $ad - bc = 1$. Since we start with $ad - bc \neq 0$, this can be accomplished by multiplying the matrix by a suitable λ.

Exercise 41.2 *What is this* λ*?*

Define the *special linear group* of 2x2 complex matrices by

$$SL_2(\mathbb{C}) := \left\{ F = \begin{bmatrix} a & b \\ c & d \end{bmatrix} : ad - bc = 1. \right\}$$

It is a group because it is nonempty and closed under multiplication and inverses.

The map

$$U|SL_2(\mathbb{C}) : SL_2(\mathbb{C}) \to \text{Möb}_+$$

is still surjective.

But F and $-F$ have the same determinant. So this time, the kernel is

$$\ker(U) = \{\text{id}, -\text{id}\}.$$

The map is still not injective, but the ambiguity is smaller. It follows that

$$\boxed{\text{Möb}_+ \cong PSL_2(\mathbb{C})/\{\text{id}, -\text{id}) := SL_2(\mathbb{C})/\{\text{id}, -\text{id}\}.}$$

This is the *projective special linear group*.

15

Factoring Möbius transformations

§42 Factoring Möbius transformations

Clearly translation, rotations, similarities, the complex inverse $1/z$, complex conjugation, and inversion $1/\bar{z}$ are Möbius transformations.

We can show that any Möbius transformation can be expressed as a product of these.

1) Orientation-preserving case:

If $c = 0$, then because $ad - bc \neq 0$, we have $d \neq 0$ and f is a similarity transformation. Compute

$$\begin{aligned} f(z) &= \frac{az+b}{d} \\ &= \frac{a}{d}z + \frac{b}{d}, \end{aligned}$$

so

$$\boxed{f = T_{b/d} \circ M_{a/d},} \tag{42.1}$$

if $c = 0$.

Note that the factors are bijections because $d \neq 0$.

If $c \neq 0$, then

$$\begin{aligned} f(z) &= \frac{az+b}{cz+d} \\ &= \frac{(a/c)(cz+d)}{cz+d} + \frac{(az+b)-(a/c)(cz+d)}{cz+d} \\ &= \frac{a}{c} + \frac{b-ad/c}{cz+d} \\ &= \frac{a}{c} + \frac{bc-ad}{c}\frac{1}{cz+d}. \end{aligned}$$

From this we get the factorization

$$f = T_{a/c} \circ M_{(bc-ad)/c} \circ I \circ T_d \circ M_c.$$

This is what we did in class.

But we can do better. We continue the calculation with

$$f(z) = \frac{a}{c} + \frac{bc-ad}{c^2}\frac{1}{z+d/c}$$

so

$$\boxed{f = T_{a/c} \circ M_{(bc-ad)/c^2} \circ I \circ T_{d/c},} \tag{42.2}$$

if $c \neq 0$.

Note that all the factors are bijections because $ad - bc \neq 0$, $c \neq 0$.

Exercise 42.1 *Use the factorization to give a new proof that Möbius transformations are invertible.*

2) Orientation-reversing case:

Let

$$f(z) = \frac{a\bar{z} + b}{c\bar{z} + d}.$$

Factor f as

$$\boxed{f = T_{b/d} \circ M_{a/d} \circ C,} \tag{42.3}$$

if $c = 0$.

Factor f as

$$\boxed{f = T_{b/d} \circ M_{(bc-ad)/c^2} \circ I \circ T_{d/c} \circ C.} \tag{42.4}$$

if $c \neq 0$.

Exercise 42.2 *Show that inversion in any cline is a Möbius transformation.*

The upshot is the following theorem:

Theorem 42.1

1) The group Möb_+ *is generated by multiplication operators, translation, and the complex inverse.*

2) The group Möb *is generated by multiplication operators, translation, the complex inverse, and complex conjugation.*

§43 Möbius transformations are conformal and preserve clines

As a corollary to §42 we have:

Theorem 43.1 *Every Möbius transformation is conformal and preserves clines.*

Proof By the factorization formulas (42.1)-(42.4), all Möbius transformations can be factored as translations, multiplications by nonzero constants, the complex inverse, and complex conjugation.

We already know that each of these special types of transformation is conformal and preserves clines.

(For the complex inverse: Recall that the complex inverse $I(z) = 1/z$ can be written $I = J \circ C$, where $J(z) = z/|z|^2 = 1/\bar{z}$, $C(z) = \bar{z}$. We already know that J and C are conformal and preserve clines. So I is conformal and preserves clines.)

It follows that all Möbius transformations are conformal and preserve clines.

□

§44 Relation of Möb to Conf(S^2)

Here are some remarks that go a bit beyond the class. I will skip the details.

The sphere S^2 can be identified with the extended plane $\hat{\mathbb{C}}$ by stereographic projection σ. This leads to a map

$$W : \{\text{bijections of } \hat{\mathbb{C}}\} \to \{\text{bijections of } S^2\}$$

given by

$$W_f := \sigma^{-1} \circ f \circ \sigma.$$

Obviously W is bijective and is a group isomorphism.

So we can regard each $f \in$ Möb as acting on S^2 via W_f.

By Corollary 20.1, σ is conformal for $z \neq N := (0, 0, 1)$. By Theorem 43.1, f is conformal for $z \neq \infty, -d/c$.

So W_f is conformal except possibly at the images $\sigma^{-1}(\infty)$, $\sigma^{-1}(\infty)$ of the points ∞ and $-d/c$.

At the exceptional points, it can be checked by hand that W_f is conformal. We will skip this.

So W_f is a conformal bijection of S^2.

So we get the subgroup relation

$$W(\text{Möb}) \leq \text{Conf}(S^2),$$

where Conf(S^2) is the group of all conformal transformations of S^2.

Informally, regarding $\hat{\mathbb{C}}$ and S^2 as identical, we can write this as

$$\text{Möb} \leq \text{Conf}(S^2).$$

Is it equality? That is, can every conformal transformation of S^2 be written as a Möbius transformation?

The answer is yes.

But it requires methods of complex analysis, especially the Liouville Theorem.

The net result is (identifying $\hat{\mathbb{C}}$ with S^2)

$$\text{Möb} = \text{Conf}(S^2).$$

16

Properties of Möbius transformations

§45 Möbius transformations are 3-transitive

Theorem 45.1

1) The action of Möb_+ *on* $\hat{\mathbb{C}}$ *is triply transitive, meaning that for each triple*

$$z_1, z_2, z_3 \in \hat{\mathbb{C}}$$

of distinct points, and each triple

$$w_1, w_2, w_3 \in \hat{\mathbb{C}}$$

of distinct points, there is a transformation $f \in \text{Möb}_+$ *with*

$$f(z_1) = w_1, \qquad f(z_2) = w_2, \qquad f(z_3) = w_3.$$

2) f *is unique.*

Proof 1. The points z_1, z_2, z_2 and w_1, w_2, w_3 are given. We must find f to make the equations work.

Wlog we may assume

$$z_1 = 0, \qquad z_2 = 0, \qquad z_3 = \infty.$$

For then we can find f_1, f_2 such that

$$f_1(0) = z_1, \qquad f_1(1) = z_2, \qquad f_1(\infty) = z_3,$$

$$f_2(0) = w_1, \qquad f_2(1) = w_2, \qquad f_2(\infty) = w_3,$$

and then $f := f_2 \circ (f_1)^{-1}$ satisfies

$$f(z_1) = w_1, \qquad f(z_2) = w_2, \qquad f(z_3) = w_3.$$

2. So set

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

We must find a, b, c, d so that

$$f(0) = w_1, \qquad f(1) = w_2, \qquad f(\infty) = w_3.$$

We proceed as follows.

Case 1: $w_3 = \infty$

Then $f(\infty) = \infty$. So $c = 0$. So

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

Dividing all coefficients by d, wlog $d = 1$. We get

$$f(z) = az + b,$$

an OE similarity transformation. Since $z_1 \neq z_2$, $w_1 \neq w_2$, it is then geometrically clear that there is a similarity transformation such that

$$f(0) = w_1, \qquad f(1) = w_2.$$

It is unique because of the OE requirement. Algebraically, we get $b = w_1$, $a = w_2 - w_1$. These values are forced, so f is unique.

Case 2: $w_3 \neq \infty$.

Then $c \neq 0$. Dividing all coefficients by c, wlog $c = 1$. We require

$$f(0) = w_1, \qquad f(1) = w_2, \qquad f(\infty) = w_3,$$

where all w_i are finite. This becomes

$$\frac{b}{d} = w_1, \qquad \frac{a+b}{1+d} = w_2, \qquad a = w_3,$$

i.e.

$$b = dw_1, \qquad a + b = w_2 + dw_2, \qquad a = w_3,$$

Substituting the first and last equations into the middle one, we get

$$w_3 + dw_1 = w_2 + dw_2$$

so

$$d = -\frac{w_2 - w_3}{w_2 - w_1}$$

so

$$b = -w_1 \frac{w_2 - w_3}{w_2 - w_1}.$$

Summarizing,

$$a = w_3, \qquad b = -w_1 \frac{w_2 - w_3}{w_2 - w_1}, \qquad c = 1, \qquad d = -\frac{w_2 - w_3}{w_2 - w_1}.$$

These values are forced (under the normalization $c = 1$), so f is unique. To prove existence, it suffices to back-substitute and verify

$$f(0) = w_1, \qquad f(1) = w_2, \qquad f(\infty) = w_3,$$

as can easily be done.

□

For the next application, we note the following fact.

Proposition 45.2 *Through every three distinct points of $\hat{\mathbb{C}}$ runs a unique cline.*

Proof *Case 1:* One of the points is ∞.

Then the unique extended line through the other two points is the unique cline through all three points.

Case 2: The points are finite and collinear.

Then the unique extended line through all three points is the unique cline through all three points.

Case 3: The points are finite and not collinear.

Then a classic theorem of geometry allows us to construct a unique circle through the three points with ruler and compass. This circle is then the unique cline through the three points.

□

As a corollary to the 3-transitivity, we now obtain the following transitivity on clines.

Proposition 45.3 *The orientation-preserving Möbius translations take any cline to any other cline.*

Proof Let E, F be two clines in $\hat{\mathbb{C}}$. Pick 3 distinct points z_1, z_2, z_3 on E and three distinct points w_1, w_2, w_3 on F. Let $f \in \text{Möb}_+$ take $z_1 \to w_1$, $z_2 \to w_2$, $z_3 \to w_3$. Then $f(E)$ is a cline through w_1, w_2, w_3, so $f(E) = F$ by the uniqueness statement of the previous proposition.

□

§46 The cross ratio and its symmetries

We now come to an all-important invariant of Möb_+ transformations called the cross ratio.

Definition 46.1 Let $z_1, z_2, z_3, z_4 \in \hat{\mathbb{C}}$ be distinct. We define their *cross ratio* by

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

If one argument is ∞, we cross out the factors where ∞ appears. For example:

$$\frac{(\infty - z_1)(z_4 - z_2)}{(\infty - z_2)(z_4 - z_1)} = \frac{z_4 - z_2}{z_4 - z_1}$$

which has only finite numbers.

The cross ratio is sort of miraculous, but it will take some investigation to see this.

The symmetries of the cross ratio

The cross ratio has a lot of symmetries.

Proposition 46.2

$$[A, B; C, D] = [B, A; D, C] = [C, D; A, B] = [D, C; B, A].$$

The proof is trivial.

The proposition says: We can switch the first two and the last two, or the first two *with* the last two. More precisely:

i) Switch positions $1 \leftrightarrow 2$, $3 \leftrightarrow 4$, no change.

ii) Switch $1 \leftrightarrow 3$, $2 \leftrightarrow 4$, no change.

This pattern explains the location of the semicolon.

If we define $\lambda := [A, B; C, D]$, then we have further (table from Wikipedia)

Proposition 46.3

$$[A,B;C,D] = [B,A;D,C] = [C,D;A,B] = [D,C;B,A] = \lambda$$
$$[A,B;D,C] = [B,A;C,D] = [C,D;B,A] = [D,C;A,B] = \frac{1}{\lambda}$$
$$[A,C;B,D] = [B,D;A,C] = [C,A;D,B] = [D,B;C,A] = 1-\lambda$$
$$[A,C;D,B] = [B,D;C,A] = [C,A;B,D] = [D,B;A,C] = \frac{1}{1-\lambda}$$
$$[A,D;B,C] = [B,C;A,D] = [C,B;D,A] = [D,A;C,B] = \frac{\lambda-1}{\lambda}$$
$$[A,D;C,B] = [B,C;D,A] = [C,B;A,D] = [D,A;B,C] = \frac{\lambda}{\lambda-1}.$$

The proof is trivial. The special case where one of the inputs is ∞ must be done separately.

So the 24 permutations of A, B, C, D fall into 6 groups of 4, each of which has the same value. So there are only 6 possible values for the cross ratio of A, B, C, D, depending on the order, and they all can be calculated from each other.

The new rules can be summarized as follows:

iii) Switch $1 \leftrightarrow 2$ or $3 \leftrightarrow 4$, send λ to $1/\lambda$.

iv) Switch $2 \leftrightarrow 3$ or $1 \leftrightarrow 4$, send λ to $1-\lambda$

Together with i)-ii), these yield the first three rows of the table; the final three rows follow by combining them.

Possible values of the cross ratio

Proposition 46.4

$$[z_1, z_2; z_3, z_4] \neq 0, 1, \infty.$$

Proof

1. It is obvious that $[z_1, z_2; z_3, z_4] \neq 0, \infty$ because the numbers are all distinct.

2. If

$$[z_1, z_2; z_3, z_4] = 1,$$

then by Proposition 46.3, first and third lines,

$$[z_1, z_3; z_2, z_4] = 1 - 1 = 0,$$

which is impossible by 1. So $[z_1, z_2; z_3, z_4] = 1$ is impossible.

□

Are there other forbidden numbers like these?

Indeed, the set $\{0, 1, \infty\}$ is invariant under the six operations

$$\lambda, \qquad 1/\lambda, \qquad 1-\lambda, \qquad 1/(1-\lambda), \qquad (\lambda-1)/\lambda, \qquad \lambda/(\lambda-1).$$

So we won't get any other forbidden numbers by applying the symmetries of Proposition 46.3.

§47 The cross ratio is preserved

References

- Loustau pp. 124-125

Theorem 47.1 *Let*

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

be an element of Möb_+. *Let* z_1, z_2, z_3, z_4 *be distinct points of* $\hat{\mathbb{C}}$. *Then*

$$[f(z_1), f(z_2), f(z_3), f(z_4)] = [z_1, z_2, z_3, z_4].$$

Proof

1. If we insert f directly into the formula for the cross-ratio, we have to do a long calculation. So let's find another way.

Recall that

$$T_a, M_b, I$$

generate Möb_+. So it suffices to check that each of these conserve the cross ratio.

2. Let $f = T_a$, $a \in \mathbb{C}$. Then

$$\begin{aligned}[f(z_1), f(z_2), f(z_3), f(z_4)] &= \frac{((z_3+a)-(z_1+a))((z_4+a)-(z_2+a))}{((z_3+a)-(z_2+a))((z_4+a)-(z_1+a))} \\ &= \frac{(z_3-z_1)(z_4-z_2)}{(z_3-z_2)(z_4-z_1)} \\ &= [z_1, z_2, z_3, z_4].\end{aligned}$$

The case where one of the z_i is ∞ is included above by striking the appropriate factors on top and bottom.

3. Let $f = R_b$, $b \neq 0$. Then

$$\begin{aligned}[f(z_1), f(z_2), f(z_3), f(z_4)] &= \frac{(bz_3-bz_1)(bz_4-bz_2)}{(bz_3-bz_2)(bz_4-bz_1)} \\ &= \frac{(z_3-z_1)(z_4-z_2)}{(z_3-z_2)(z_4-z_1)} \\ &= [z_1, z_2, z_3, z_4].\end{aligned}$$

The case where one of the z_i is ∞ is included above by striking the appropriate factors on top and bottom.

4. Let $f = I$. We have to do several cases.

Assume first that none of the z_i are 0 or ∞. Then

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

To handle the cases of 0 and/or ∞, it suffices to check the following four identities by hand. The remaining cases can be reduced to one of these via Proposition 46.2. The variables z_2, z_3, z_4 are assumed to be distinct and not equal to 0 or infty. Prove:

$$\begin{aligned}[I(0), I(z_2), I(z_3), I(z_4)] &= [0, z_2, z_3, z_4], \\ [I(\infty), I(z_2), I(z_3), I(z_4)] &= [\infty, z_2, z_3, z_4], \\ [I(0), I(\infty), I(z_3), I(z_4)] &= [0, \infty, z_3, z_4], \\ [I(0), I(z_2), I(\infty), I(z_4)] &= [0, z_2, \infty, z_4].\end{aligned}$$

Each is trivial. We leave them to the reader.

□

§48 When the cross ratio is real

Proposition 48.1 *Let z_1, z_2, z_3, z_4 be distinct points of $\hat{\mathbb{C}}$. Then*

$$[z_1, z_2, z_3, z_4] \in \mathbb{R}$$

precisely when

$$z_1, z_2, z_3, z_4 \quad \textit{lie on a common cline.}$$

We begin with

Observation: $[z, 1; 0, \infty] = z$.

(Proof trivial.)

Proof of Proposition 48.1

Let z_1, z_2, z_3, z_4 be distinct points in $\hat{\mathbb{C}}$. By triple transitivity, select f in Möb_+ such that

$$f(z_2) = 1, \qquad f(z_3) = 0, \qquad f(z_4) = \infty.$$

Define $z = f(z_1)$. We then calculate:

$$\begin{aligned} z &= [z, 1; 0, \infty] && \text{by the observation} \\ &= [f(z_1), f(z_2), f(z_3), f(z_4)] \\ &= [z_1, z_2, z_3, z_4] && \text{by invariance} \end{aligned}$$

($\Longrightarrow$) Suppose $[z_1, z_2, z_3, z_4] \in \mathbb{R}$. Then $z \in \mathbb{R}$. So $z, 1, 0, \infty$ lie on a common cline, namely the real axis. That is, $f(z_1), f(z_2), f(z_3), f(z_4)$ lie on a common cline. But f^{-1} takes clines to clines. So z_1, z_2, z_3, z_4 lie in a common cline.

($\Longleftarrow$) Suppose z_1, z_2, z_3, z_4 lie on a common cline. Since f takes clines to clines, $f(z_1), f(z_2), f(z_3), f(z_4)$ lie on a common cline. That is, $z, 1, 0, \infty$ lie on a common cline. But any cline through $1, 0, \infty$ must be the real axis. So $z \in \mathbb{R}$. But then $[z_1, z_2, z_3, z_4] = z \in \mathbb{R}$.

□

Here is the interpretation of the cross ratio in light of the Proposition and the Observation.

The cross ratio $[z_1, z_2, z_3, z_4]$ expresses a *relationship* between z_1 and the reference points z_2, z_3, z_4.

The points z_2, z_3, z_4 are like markers, or guideposts, on the celestial sphere $S^2 = \hat{\mathbb{C}}$ that play the role of $1, 0, \infty$. The cline E through z_2, z_3, z_4 plays the

role of the real axis. The entire sky is painted with tiny numbers, giving the value

$$z = [z_1, z_2, z_3, z_4]$$

at the variable point z_1.

z_1 lies on the cline E if and only if z is real. More generally, the relationship of z_1 to z_2, z_3, z_4 (in some mysterious "complex projective" sense) is the same as the relationship of z to $1, 0, \infty$.

17

The elements of $\mathrm{M\ddot{o}b}_+(B_1)$

§49 Some elements of $\mathrm{Möb}_+(B_1)$

We define

$$\mathrm{Möb}_+(B_1) := \{f \in \mathrm{Möb}_+ : f(B_1) = B_1\}$$

for the orientation-preserving Möbius transformations of B_1. Then

$$f|B_1 : B_1 \to B_1$$

is a bijection. We will usually abbreviate this as f.

What are all the elements of $\mathrm{Möb}_+$?

Our goal in this section is to list certain useful ones.

1) Rotations about 0:

Obviously we have

$$\boxed{R_\theta = M_{e^{i\theta}} \in \mathrm{Möb}_+(B_1).}$$

2) Movements along the real axis:

Let $-1 < t < 1$. Define

$$K_t(z) := \frac{z+t}{tz+1}, \qquad z \in B_1.$$

Earlier in the course, you proved in an exercise (Serie 4 Aufgabe 3):

$$K_t(B_1) = B_1. \tag{49.1}$$

As a result, we get

$$\boxed{K_t \in \mathrm{Möb}_+(B_1).}$$

For the record, let's prove (49.1).

Proof of (49.1) 1. First of all, let us show

$$K_t(B_1) \subseteq B_1.$$

Let $z < 1$. We must show $K_t(z) < 1$. That is, show

$$\left|\frac{z+t}{tz+1}\right| < 1.$$

We successively reduce this as follows

$$\begin{aligned} |z+t| &< |tz+1| \\ |z+t|^2 &< |tz+1|^2 \\ (z+t)(\bar{z}+t) &< (tz+1)(t\bar{z}+1) \\ |z|^2 + zt + \bar{z}t + t^2 &< t^2|z|^2 + tz + t\bar{z} + 1 \\ |z|^2 + t^2 &< t^2|z|^2 + 1 \\ 0 &< (1-t^2)(1-|z|^2). \end{aligned}$$

This latter is true because $|t| < 1$, $|z| < 1$. So all the previous inequalities are true. So we get

$$K_t(B_1) \subseteq B_1,$$

as claimed.

2. Replacing t by $-t$ we get

$$K_{-t}(B_1) \subseteq B_1.$$

Now (as may easily be checked), K_t is the inverse of K_t, so applying K_t, we get

$$B_1 \subseteq K_t(B_1).$$

Combining this with the inclusion in 1., we get

$$K_t(B_1) = B_1,$$

as was desired.

□

Let us visualize K_t. Let $L := x\text{-axis} \cap B_1$. Obviously

$$K_t(L) = L.$$

We also have

$$K_t(-1) = -1, \qquad K_t(1) = 1.$$

and

$$K_t(-t) = 0, \qquad K_t(0) = t.$$

We can ask:

a) Where does 0 go under repeated applications of K_t, and where does it come from?

Let us suppose that $0 < t < 1$.

We calculate where 0 goes:

$$K_t(0) = t, \qquad K_t(K_t(0)) = \frac{2t}{t^2+1}, \qquad K_t(K_t(K_t(0))) = \dots$$

The expressions get more and more complicated, but they keep increasing. Indeed, one can confirm this by showing that for $-1 < s < 1$, $t > 0$,

$$K_t(s) > s.$$

To see where 0 comes from under K_t, we apply the inverse K_{-t}:

$$K_{-t}(0) = -t, \qquad K_{-t}(K_{-t}(0)) = \frac{-2t}{t^2+1}, \qquad K_{-t}(K_{-t}(K_{-t}(0))) = \dots$$

These points can be plotted on the real number line and we find

$$\dots < K_{-t}^3(0) < K_{-t}^2(0) < K_{-t}(0) < 0 < K_t(0) < K_t^2(0) < K_t^3(0) < \dots$$

This gives us a picture of the action of K_t on the segment from -1 to $+1$: it moves everything to the right in a nonlinear way. -1 is a source, $+1$ is a sink.

b) Where does the y-axis go under repeated applications of K_t, and where does it come from?

We actually are interested in the segment $M := y\text{-axis} \cap B_1$.

Since M is a cline normal to S^1 (intersected with B_1), and S^1 goes to itself under K_t, $K_t(M)$ must be a cline normal to S^1 (intersected with B_1). Similarly, $K_t(M)$ will be normal to the x-axis.

Similar considerations apply to $K_t^2(M)$, $K_t^3(M)$, etc., and to $K_{-t}(M)$, $K_{-t}^2(M)$, etc. So the sequence

$$\dots, K_{-t}^2(M), \quad K_{-t}(M), \quad M, \quad K_t(M), \quad K_t^2(M), \dots$$

are a sequence of circle arcs progressing from left to right, from the sources -1 to the sink $+1$, giving a striped effect.

c) What happens with clines through -1 and $+1$?

Let E be a cline through -1 and $+1$. I claim: $K_t(E) = E$.

First, $K_t(E)$ is a cline. Since -1, $+1$ are fixed points of K_t, $K_t(E)$ is a cline through -1, $+1$.

Now we would like to prove that $K_t(E) = E$ and not some other cline through -1, $+1$.

@ How?

So we get

$$K_t(E) = E$$

Now K_t moves every point of E along E rightwards, just as it does in the special case $E = x\text{-axis} \cap B_1$. -1 is the source and $+1$ is the sink.

The entire action of $K - t$ can be described in pictures by filling B_1 with flow arrows that preserve the x-axis, preserve each circle from -1 to $+1$, and preserve the upper and lower boundary semicircles of B_1.

3) Movements along any line through 0:

Fix $b \in B_1$. Define in analogy to K_t,

$$K_b(z) := \frac{z+b}{\bar{b}z+1}, \qquad z \in B_1.$$

If $b \neq 0$, define

$$L_b := \left\{ s\frac{b}{|b|} : -1 < s < 1 \right\},$$

the intersection of B_1 with the line determined by 0 and b.

Then we may easily prove by essentially the same proof as for K_t

$$K_b(B_1) = B_1.$$

(The proof needs the $\bar{b}$ coefficient in the denominator.) So

$$\boxed{K_b \in \text{Möb}_+(B_1).}$$

We can verify

$$K_t(L_b) = L_b,$$

$$K_t\left(-\frac{b}{|b|}\right) = -\frac{b}{|b|}, \qquad K_t\left(\frac{b}{|b|}\right) = \frac{b}{|b|},$$

$$K_b(-b) = 0, \qquad K_b(0) = b.$$

Factorization of K_b

Now, it is possible to write K_b in terms of K_t and rotations as follows. Assume $b \neq 0$ (the $b = 0$ case is just the identity). Write $b = |b|e^{i\theta}$, where $\theta \in \mathbb{R}$, $|b| > 0$. Calculate

$$\begin{aligned} K_b(z) &= \frac{z+b}{\bar{b}z+1} \\ &= \frac{z+|b|e^{i\theta}}{|b|e^{-i\theta}z+1} \\ &= e^{i\theta}\,\frac{e^{-i\theta}z+|b|}{|b|e^{-i\theta}z+1} \\ &= (R_\theta \circ K_{|b|} \circ R_{-\theta})(z). \end{aligned}$$

So

$$K_b = R_\theta \circ K_{|b|} \circ R_{-\theta}.$$

This implies two things.

First, it gives a second proof (besides the one mentioned above) that K_b is a bijection of B_1 to itself.

Second, it shows that the way K_b acts on B_1 is just the same as the way $K_{|b|}$ acts on B_1, but rotated by θ. In particular, we can get a picture of the action of K_b by rotating the $K_{|b|}$ picture by θ.

.

§50 Factoring elements of $\mathrm{M\ddot{o}b}_+(B_1)$

We have seen that $R_\theta, K_b \in \mathrm{M\ddot{o}b}_+(B_1)$. We are now in a position to prove that they generate the group.

Theorem 50.1

(a)

$$\mathrm{M\ddot{o}b}_+(B_1) = \left\{ \frac{ax+b}{\bar{b}z+\bar{a}} : a, b \in \mathbb{C}, |a| > |b| \right\}$$

(b) As a consequence, every element of $\mathrm{M\ddot{o}b}_+(B_1)$ *can be written in the form*

$$f(z) = e^{i\theta}\frac{z+b}{\bar{b}z+1}, \qquad z \in B_1,$$

that is,

$$f = R_\theta \circ K_b.$$

We omit the proof because there is no time. The easy step is going from (a) to (b). The harder step is (a). The theorem is used in the next section.

18

Isometries, geodesics, and distances in $\mathbb{H}^2$

§51 Isometries of $\mathbb{H}^2$

Theorem 51.1 *The isometries of $\mathbb{H}^2$ are exactly the Möbius transformations that preserve B_1. That is,*

$$\mathrm{Isom}(\mathbb{H}^2) = \mathrm{Möb}(B_1), \qquad \mathrm{Isom}_+(\mathbb{H}^2) = \mathrm{Möb}_+(B_1).$$

We now have the ingredients to prove this. But we'll only sketch the proof – and only for the orientation-preserving case. The full proof takes a few pages.

($\supseteq$)

R_θ is a hyperbolic isometry. With some effort, using stretch factors and the hyperbolic length elements at z and at $K_b(z)$, K_b is a hyperbolic isometry.

By Theorem 50.1, these generate $\mathrm{Möb}_+(B_1)$. So every element of $\mathrm{Möb}_+(B_1)$ is a hyperbolic isometry. So

$$\mathrm{Isom}_+(\mathbb{H}^2) \supseteq \mathrm{Möb}_+(B_1).$$

($\subseteq$)

Let f be an orientation-preserving hyperbolic isometry. By setting $g := K_b \circ R_\theta \circ f$ for suitable isometries R_θ and K_b, we may arrange that g fixes 0, and indeed that g fixes every point on the real segment $L := x\text{-axis} \cap B_1$.

It follows by considering distances to points on L that for each $z \in \mathbb{H}^2$, $g(z) = z$ or $g(z) = \bar{z}$. Since g is orientation-preserving, $g(z) = z$. So $g = \mathrm{id}$.

So $f = R_{-\theta} \circ K_{-b}$ is an orientation-preserving Möbius transformation. So

$$\mathrm{Isom}_+(\mathbb{H}^2) \subseteq \mathrm{Möb}_+(B_1).$$

§52 Geodesics of $\mathbb{H}^2$

We now can prove Theorem 52.1, which identified the geodesics in $\mathbb{H}^2$. Let us restate it.

Theorem 52.1 *The geodesics of $\mathbb{H}^2$ are precisely the curves of the form*

$$E \cap B_1,$$

where E is a cline that meets ∂B_1 orthogonally. They are all minimizing.

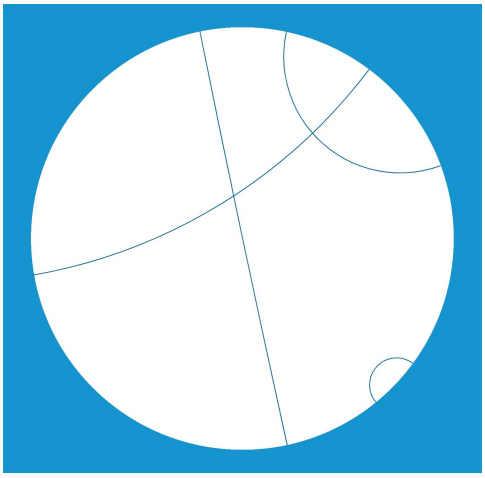

Figure 52.1: Hyperbolic geodesics (made with A. Zampa's Geogebra applet)

Proof 1. Let $\alpha = E \cap B_1$, where E is a cline that meets ∂B_1 orthogonally. Let A and B be the endpoints of α, lying on B_1. Let C be a point on ∂B_1 distinct from A and B such that

$$A, C, B$$

are counterclockwise around S^1.

Select an orientation-preserving Möbius transformation f that takes

$$A, C, B \to -1, -i, 1.$$

So $f(S^1) = S^1$. But f is a homeomorphism (bijective and continuous in both directions). So (assuming some topology methods) either

$$f(B_1) = B_1$$

or

$$f(B_1) = \hat{\mathbb{C}} \setminus \bar{B}_1.$$

2. We claim that the fact that f preserves the ccl order of A, C, B around S^1 implies that $f(B_1) = B_1$.

For if $f(B_1) = \hat{\mathbb{C}} \setminus \bar{B}_1$, then $(I \circ f)(B_1) = B_1$. But $I \circ f$ takes

$$A, B, C \to -1, i, 1,$$

that is, it is an orientation-preserving bijection of $\bar{B}_1$ to $\bar{B}_1$ that takes ccl to cl. This is impossible.

Therefore $f(B_1) = B_1$. So $f \in \text{Möb}_+(B_1)$. So f is a hyperbolic isometry.

3. Now f takes $\alpha = E \cap B_1$ to $f(E) \cap B_1$. Then $f(E)$ is a cline. It is orthogonal to S^1 because f preserves angles. So $f(E) = \hat{\mathbb{R}}$. So $f(\alpha) = \mathbb{R} \cap B_1$.

But $\mathbb{R} \cap B_1$ is a minimizing geodesic. Since f is an isometry, α is a minimizing geodesic.

□

We have the following two corollaries arising from the proof of this theorem.

Corollary 52.2 *The group $Isom_+(\mathbb{H}^2)$, extended to act on S^1, takes any three counterclockwise points on S^1 to any three counterclockwise points on S^1.*

Compare this to Theorem 45.1.

A *directed geodesic* is a geodesic with a direction specified along the geodesic.

Corollary 52.3 *The group $Isom_+(\mathbb{H}^2)$ takes any directed geodesic in $\mathbb{H}^2$ to any directed geodesic in $\mathbb{H}^2$, preserving the direction of the geodesic.*

Compare this to Proposition 45.3.

§53 Cross ratio formula for distance

We have a remarkable formula for hyperbolic distance in terms of the cross-ratio.

Theorem 53.1 *Let z, w be distinct points in $\mathbb{H}^2$. Let α be the hyperbolic geodesic through them. Let A, B be the endpoints of α on S^1 in such a way that*

$$A, z, w, B$$

are in order along α. Then

$$\boxed{d_H(z,w) = \log(|[B, A; z, w]|).}$$

Effectively, this theorem generalizes Theorem 25.1, which says

$$d_H(a,b) = \log \frac{(1-a)(1+b)}{(1+a)(1-b)},$$

for points $a < b$ on $L := \mathbb{R} \cap B_1$.

Proof By Corollary 52.3, there is a hyperbolic isometry f such that

$$f(\alpha) = L, \qquad f(A) = -1, \qquad f(B) = +1.$$

Then $f(z)$, $f(w)$ lie in L and are real. The order is retained and

$$f(z) < f(w).$$

So by Theorem 47.1 and Theorem 25.1,

$$\begin{aligned} d_H(z,w) &= d_H(f(z), f(w)) \\ &= \log \left| \frac{(1-f(z))(1+f(w))}{(1+f(z))(1-f(w))} \right| \\ &= \log |[1, -1; f(z), f(w)]| \\ &= \log |[B, A; z, w]|, \end{aligned}$$

as required.

□

§54 Arccosh formula for distance

A disadvantage of the cross ratio formula is that is can be awkward to find the endpoints A, B.

Here is a formula for distance in $\mathbb{H}^2$ that only depends on $|z|$, $|w|$, and $|z-w|$.

Theorem 54.1 *Let $z, w \in \mathbb{H}^2$. Then*

$$d_H(z,w) = \operatorname{arccosh}\left(1 + \frac{2|z-w|^2}{(1-|z|^2)(1-|w|^2)}\right), \qquad z, w \in B_1.$$

It generalizes (27.3), which says the same thing, but only for points in $L := \mathbb{R} \cap B_1$.

We're now in a position to prove it, but we've run out of time.

Part IV

End

19

Bibliography

§55 Books

Last year's script:

- T. Ilmanen, Geometrie 2020, https://metaphor.ethz.ch/x/2020/hs/401-1511-00L/literatur/script.pdf.
  The topics were different, but the older script has more about group theory. It also has many pictures and audiovisuals.

Very accessible:

- J. R. Weeks, *The Shape of Space,* CRC press, 2019.
- M. Hitchman, *Geometry with an Introduction to Cosmic Topology*, https://mphitchman.com/geometry/frontmatter.html.

Classical:

- E. A. Abbott, *Flatland,* Dover Publications, 1884.
- D. Burger, *Sphereland: A Fantasy About Curved Spaces and an Expanding Universe*, 1957.

For fractals:

- K. J. Falconer, *The geometry of fractal sets,* Cambridge Univ. Press, 1985, https://www.cambridge.org/core/books/geometry-of-fractal-sets/7ECAB3C918C66E62AB673246B2CDE6FA.

For group theory:

- D. Saracino, *Abstract Algebra: A First Course,* Waveland Press, 2008, https://www.waveland.com/browse.php?t=483, pp 1-132.
- J. J. Rotman, *An Introduction to the Theory of Groups,* Springer, 1984, https://www.springer.com/gp/book/9780387942858.

For linear algebra:

- K. Jänich, *Lineare Algebra*, Springer, 11th ed., 2010.
- G. Fischer, *Lineare Algebra: Eine Einführung für Studienanfänger*, 18th ed., Springer, 2014.

For complex analysis:

- L. Ahlfors, *Complex Analysis*, 1979, pp 18-20, 76-88. Looking for a more available reference.

For hyperbolic geometry:

- J. W. Anderson, *Hyperbolic Geometry*, Springer, 2005.
- W. P. Thurston, *Three-dimensional Geometry and Topology, vol. I*, Princeton Univ. Press, 1997, pp. 3-42, 43-.
- B. Loustau, *Hyperbolic geometry*, online notes, https://arxiv.org/abs/2003.11180, 2020.
- A. F. Beardon, *The Geometry of Discrete Groups*, Springer, 1983, pp.

56-82, 126-187.

Riemannian geometry:

- J. M. Lee, *Introduction to Riemannian Manifolds*, 2nd ed., Springer, 2018.

Mathematical symbols:

- *Liste mathematischer Symbole,*
  `https://de.wikipedia.org/wiki/Liste_mathematischer_Symbole`

Mathematical dictionaries:

- G. Eisenreich, R. Sube, *Dictionary of Mathematics; Wörterbuch Mathematik*, Verlag Harry Deutsch, 1987.

§56 Articles, blogs, and references

Articles:

- T. M. Apostol & M. A. Mnatsakanian, *A fresh look at the method of Archimedes*, Math. Assoc. of America Monthly 111, 2004.
- J. Weeks, *Non-Euclidean billiards in VR*, `http://archive.bridgesmathart.org/2020/bridges2020-1.pdf`.

Blogs:

-

Wikipedia:

- *Conformal map projection*
- *Equal-area map*
- *Euclidean group*
- *Lambert cylindrical equal-area projection*
- *Mercator projection*
- *Stereographic map projection*
- *Stereographic projection*
- *Taxicab geometry*
- *Uniform tilings in hyperbolic plane*

Names and properties of concrete groups:

- T. Dokchitser, interactive list of groups of small order, `https://people.maths.bris.ac.uk/~matyd/GroupNames/`
- J. Jones, interactive calculator for groups of small order, `https://hobbes.la.asu.edu/groups/groups.html`.
- X. Lee, wallpaper groups, `http://xahlee.info/Wallpaper_dir/c5_17WallpaperGroups.html`

§57 Software, visualization, and activities

Gomath exhibition (14-15 March 2022):

- https://math.ethz.ch/news-and-events/events/gomath/gomath-2022.html

Jeff Weeks geometry apps:

- Flying in curved space (iOS, macOS, Windows)
  http://www.geometrygames.org/CurvedSpaces
- Kaleidotile (iOS, macOS, Windows)
  http://www.geometrygames.org/KaleidoTile
- Crystal flight (iOS, macOS)
  http://www.geometrygames.org/CrystalFlight

Geogebra:

- https://www.geogebra.org/m/tHvDKWdC

Hyperrogue:

- https://roguetemple.com/z/hyper

Youtube videos:

- ZenoTheRogue
  https://www.youtube.com/channel/UCfCtbgiDxwFtlqrbEralvTw

List of Figures

3.1 Surjective . . . 13
3.2 Injective . . . 14
3.3 Bijective . . . 14
3.4 Commutative diagram . . . 15
3.5 Equivalence relation . . . 16
4.1 Complex numbers . . . 17
4.2 Grid (Mathematica) . . . 20
4.3 Exponential image of grid (Mathematica) . . . 20

5.1 The 2-sphere . . . 22
5.2 A geodesic arc . . . 23
5.4 Order-4 bisected pentagonal tiling of the hyperbolic plane (Rocchini, Wikipedia) . . . 24
5.5 A tessellated hyperbolic space (J. Weeks' Curved Spaces app, `https://www.geometrygames.org/CurvedSpaces`) . . . 25
6.1 Point x in $\mathbb{R}^n$ . . . 26
6.2 Distance between x and y . . . 26

7.1 Triangle inequality . . . 30
7.2 A three-point metric space . . . 30
7.3 The Sierpinski gasket . . . 32
8.1 Distance is preserved . . . 33
8.2 An isometry of an **R** to another **R** . . . 33
8.3 The Sierpinski gasket . . . 34
9.1 Taxi metric . . . 36
9.2 Many pathways of the same length . . . 37
9.3 Set of points that saturate the triangle inequality . . . 37
9.4 Sup metric . . . 38
10.1 The Sierpinski gasket . . . 40
10.2 Koch snowflake (Wxs, Wikipedia, made in Inkscape with the L-system effect, `https://commons.wikimedia.org/wiki/File:KochFlake.svg`) . . . 41
10.3 Isometric embedding . . . 42

11.1 Two-sphere . . . 45
11.2 Path with velocity vector . . . 45

11.3 Central angle is length . . . 46
11.4 Arc versus chord . . . 46
11.5 The red path is the long way around . . . 47
11.6 Many paths . . . 47
12.1 Circumference of a disk . . . 49
12.2 Euclidean radius of circle . . . 50
12.3 Circumference of a circle: $\mathbb{H}^2$, $\mathbb{R}^2$ and S^2 (Mathematica) . . . 51
12.4 Area of the disk by integration of shells . . . 51
12.5 Area of a circle: $\mathbb{H}^2$, $\mathbb{R}^2$ and S^2 (Mathematica) . . . 53
12.6 The two sectors have equal areas . . . 53
12.7 A sphere inscribed in a cylinder . . . 54
12.8 A spherical cap compared to a short cylinder . . . 54

13.1 Triangle in $\mathbb{R}^2$ . . . 57
13.2 Triangle in S^2 . . . 57
13.3 Half an orange slice . . . 58
13.4 Triangle in $\mathbb{H}^2$ . . . 60
13.5 Three types of triangles (J. Weeks, *The shape of space*, Fig. 3.7 60

15.1 Mercator map (Strebe, Wikipedia, CC BY-SA 3.0 license) . . 64
15.2 Lambert cylindrical equal area projection (Strebe, Wikipedia, CC BY-SA 3.0 license) . . . 66
15.3 How Lambert is done (KoenB, Schuyler Erle, Wikipedia, made with Blender, public domain) . . . 66
16.1 Stereographic projection (Che Che, Mark.Howison, Wikipedia, CC BY-SA 4.0 International license) . . . 69
16.2 Stereographic projection (Strebe, Wikipedia, CC BY-SA 3.0 license) . . . 70

22.1 Order-4 bisected pentagonal tiling of the hyperbolic plane (Rocchini, Wikipedia) . . . 88

23.1 The blue path is minimizing. The red one is not. . . . 92
26.1 Hyperbolic geodesics (made with A. Zampa's Geogebra applet) 97
26.2 Many parallels through a given point (made with A. Zampa's Geogebra applet) . . . 98
26.3 Limiting parallel (made with A. Zampa's Geogebra applet) . . 99
26.4 Ultraparallel (made with A. Zampa's Geogebra applet) 99

29.1 Hyperrogue (screenshot) . . . 105
29.2 Geogebra (screenshot) . . . 106
29.3 Hyperbolic billiards VR system (J. Weeks, *Non-Euclidean Billiards in VR*, `http://archive.bridgesmathart.org/2020/bridges2020-1.pdf`) . . . 107
29.4 Pentagon with five right angles (Lixin Liu) . . . 107
29.5 Hyperbolic $(5, 4)$ tiling (made with Kaleidotile) . . . 108

29.6 Hyperbolic $(7, 3)$ tiling (made with Kaleidotile) 109
29.7 Hyperbolic tiling with $(2, 3, 7)$ symmetry (made with Kaleidotile) 109

39.1 Octahedron, Dodecahedron, Icosahedron (Cyp, Wikipedia, license `https://creativecommons.org/licenses/by-sa/3.0`)) 132

52.1 Hyperbolic geodesics (made with A. Zampa's Geogebra applet) 165