Geometric Structures in Dimension two

Bjørn Jahren

August 11, 2015

Contents

Chapter 1

Hilbert's axiom system for plane geometry; A short introduction

Euclid's "Elements" introduced the axiomatic method to geometry, and for more than 2000 years this was the main textbook for students of geometry. But the 19th century brought about a revolution both in the understanding of geometry and of logic and axiomatic method, and it became more and more clear that Euclid's system was incomplete and could not stand up to the modern standards of rigor. The most famous attempt to rectify this was made by the great German mathematician *David Hilbert*, who published a new system of axioms in his book "Grundlagen der Geometrie" in 1898. Here we will give a short presentation of Hilbert's axioms with some examples and comments, but with no proofs. For more details, we refer to the rich literature in this field — e. g. the books "Euclidean and non-Euclidean geometries" by M. J. Greenberg and "Geometry: Euclid and beyond" by R. Hartshorne.

Hilbert also treats geometry in 3-space, but we will only consider the 2-dimensional case. The basic objects of our study are then *points* and *lines* in a *plane*. At the outset the plane is just a set S where the elements P are called points. The lines are, or can at least be naturally identified with certain subsets l of S, and the fundamental relation is the *incidence relation* $P \in l$, which may or may not be satisfied by a point P and a line l. But we also introduce two additional relations: *betweenness*, enabling us to talk about points lying between two given points, and *congruence*, which is needed when we want to compare configurations in different parts of the plane. Hilbert formulated three sets of axioms for these relations: *incidence axioms, betweenness axioms* and *congruence axioms.* In addition to these we also need an *axiom of continuity* to make sure that lines and circles have "enough" points to intersect as they should, and of course the *axiom of parallels.* As we introduce Hilbert's axioms, we will gradually put more and more restrictions on these ingredients, and in the end they will essentially determine the geometry of the Euclidean plane uniquely.

Note that although *circles* also are important objects of study in classical plane geometry, we do not have to postulate them, since, as we shall see, they can be defined in terms of the other notions.

Before we start, maybe a short remark about language is in order: An axiom system is a formal matter, but the following discussion will not be very formalistic. After all, the goal is to give a firm foundation for matters that we all have a clear picture of in our minds, and as soon as we have introduced the various formal notions, we will feel free to discuss them in more common language. For example, although the relation $P \in l$ should, strictly speaking, be read: "P and l are incident", we shall use "l contains P", "P lies on l" or any obviously equivalent such expression.

We are now ready for the first group of axioms, the *incidence axioms:*

I1: *For every pair of distinct points A and B there is a unique line l containing A and B.*

I2: *Every line contains at least two points.*

I3: *There are at least three points that do not lie on the same line.*

We let $\overline{AB}$ denote the unique line containing A and B.

These three axioms already give rise to much interesting geometry, so-called "incidence geometry". Given three points A, B, C, for example, any two of them span a unique line, and it makes sense to talk about the *triangle ABC.* Similarly we can study more complicated configurations. The Cartesian model $\mathbb{R}^2$ of the Euclidean plane, where the lines are the sets of solutions of nontrivial linear equations $ax+by=c$, is an obvious example, as are the subsets obtained if we restrict a, b, c, x, y to be rational numbers ($\mathbb{Q}^2$), the integers ($\mathbb{Z}^2$), or in fact any fixed subring of $\mathbb{R}$ (requiring now also that lines are nonempty). However, *spherical geometry,* where S is a sphere and the *lines* are great circles, is not an example, since any pair of antipodal points lies on infinitely many great circles — hence the uniqueness in I1 does not hold. This can be corrected by identifying every pair of antipodal points on the sphere. Then we obtain an incidence geometry called *the (real) projective plane* $\mathbb{P}^2$. One way to think about the points of $\mathbb{P}^2$ is as lines through the origin in $\mathbb{R}^3$. If the sphere has center at the origin, such a line determines and is determined by the antipodal pair of points of intersection between the line and the sphere. A "line" in $\mathbb{P}^2$ can then be thought of as a plane through the origin in $\mathbb{R}^3$, since such a plane intersects the sphere precisely in a great circle. Notice that in this interpretation the incidence relation $P \in l$ corresponds to the relation "the line l is contained in the plane P".

There are also *finite* incidence geometries — the smallest has exactly three points where the lines are the three subsets of two elements.

The next group of axioms deals with the relation "B lies between A and C". In Euclidean geometry this is meaningful for three points A, B and C lying on the same straight line. The finite geometries show that it is not possible to make sense of such a relation on every incidence geometry, so this is a new piece of structure, and we have to declare the properties we need. We will use the notation $A * B * C$ for "B lies between A and C".

Hilbert's *axioms of betweenness* are then:

B1: *If $A*B*C$, then A, B and C are distinct points on a line, and $C*B*A$ also holds.*

B2: *Given two distinct points A and B, there exists a point C such that $A * B * C$.*

B3: *If A, B and C are distinct points on a line, then one and only one of the relations $A * B * C$, $B * C * A$ and $C * A * B$ is satisfied.*

B4: *Let A, B and C be points not on the same line and let l be a line which contains none of them. If $D \in l$ and $A * D * B$, there exists an E on l such that $B * E * C$, or an F on l such that $A * F * C$, but not both.*

If we think of A, B and C as the vertices of a triangle, another formulation of B4 is this: *If a line l goes through a side of a triangle but none of its vertices, then it also goes through exactly one of the other sides.* This formulation is also called *Pasch's axiom.* The uniqueness (the word "exactly") is actually not necessary here, as it can be shown to be a consequence of the other axioms. Note that B4 does not hold in $\mathbb{R}^n$ for $n \geq 3$. Hence I3 and B4 together define the geometry as '2–dimensional'.

In the standard Euclidean plane (and in other examples we shall study later) we can use the concept of *distance* to define betweenness. Namely, we can then define $A * B * C$ to mean that A, B are C are distinct and $d(A, C) = d(A, B) + d(B, C)$, where $d(X, Y)$ is the distance between X and Y. (Check that B1-4 then hold!) This way $\mathbb{Q}^2$ also becomes an example, but not $\mathbb{Z}^2$, since B4 is not satisfied. (Exercise 3.)

Observe also that every open, convex subset K of $\mathbb{R}^2$ (e. g. the interior of a circular disk) satisfies all the axioms so far, if we let the "lines" be the nonempty intersections between lines in $\mathbb{R}^2$ and K, and betweenness is defined as in $\mathbb{R}^2$. (This example will be important later.) The projective plane, however, can not be given such a relation. The reason is that in the spherical model for $\mathbb{P}^2$, the "lines" are great circles where antipodal points have been identified, and these identification spaces can again naturally be identified with circles. But if we have three distinct points on a circle, each of them can equally well be said to lie "between" the others. Therefore B3 can not be satisfied.

The betweenness relation can be used to define the *segment* AB as the point set consisting of A, B and all the points between A and B:

$$AB = \{A, B\} \cup \{C | A * C * B\}.$$

Similarly we can define the *ray* $\overrightarrow{AB}$ as the set

$$\overrightarrow{AB} = AB \cup \{C | A * B * C\}.$$

If A, B and C are three point not on a line, we can then define the *angle* $\angle BAC$ as the pair consisting of the two rays $\overrightarrow{AB}$ and $\overrightarrow{AC}$.

$$\angle BAC = \{\overrightarrow{AB}, \overrightarrow{AC}\}.$$

Note also that $\overline{AB} = \overrightarrow{AB} \cup \overrightarrow{BA}$.

Betweenness also provides us with a way to distinguish between the two *sides* of a line l. We say that two points A and B are *on the same side of* l if $AB \cap l = \emptyset$. It is not difficult to show, using the axioms, that this is an equivalence relation on the complement of l, and that there are exactly two equivalence classes: the two sides of l. (Exercise 4.) Similarly we say that a point D is *inside* the angle $\angle BAC$ if B and D are on the same side of $\overline{AC}$, and C and D are on the same side of $\overline{AB}$. This way we can distinguish between points inside and outside a triangle. We also say that the angles $\angle BAC$ and $\angle BAD$ are on the same (resp. opposite) side of the ray $\overrightarrow{AB}$ if C and D are on the same (resp. opposite) side of the line $\overline{AB}$.

The same idea can also be applied to distinguish between the points on a line on either side of a given point. Using this, one can define a linear ordering of all the points on a line. Therefore the axioms of betweenness are sometimes called "axioms of order".

We have now introduced some of the basic concepts of geometry, but we are missing an important ingredient: we cannot yet compare two different configurations of points and lines. To achieve this, we need the fundamental notion of *congruence*. Intuitively, we may think of two configurations as congruent if there is some kind of "rigid motion" which moves one onto the other. In the Euclidean plane $\mathbb{R}^2$ this can be defined in terms of angle measures and distances, such that two configurations are congruent if all their ingredients are "of the same size". However, this has no meaning on the basis of just the incidence- and betweenness axioms. Hence congruence has to be introduced as yet another piece of structure — a relation whose properties then must must be defined by additional axioms.

There are two basic notions of congruence — congruence of segments and congruence of angles. Congruence of more general configurations can then be defined as a one-one correspondence between the point sets involved such that all corresponding segments and angles are congruent. We use the notation $AB \cong CD$ for "the segment AB is congruent to the segment CD", and similarly for angles or more general configurations. Hilbert's axioms for congruence of segments are:

C1: *Given a segment AB and a ray r from C, there is a uniquely determined point D on r such that $CD \cong AB$.*

C2: *$\cong$ is an equivalence relation on the set of segments.*

C3: *If $A * B * C$ and $A' * B' * C'$ and both $AB \cong A'B'$ and $BC \cong B'C'$, then also $AC \cong A'C'$.*

If betweenness is defined using a distance function (as in the Euclidean plane) we can define $AB \cong CD$ as $d(A, B) = d(C, D)$. C2 and C3 are then automatically satisfied, and C1 becomes a stronger version of B2.

Even without a notion of distance we can use congruence to compare "sizes" of two segments: we say that AB is shorter than CD ($AB < CD$) if there exists a point E such that $C * E * D$ and $AB \cong CE$.

We can now also define what we mean by a *circle:* Given a point O and a segment AB, we define the circle with center O and radius (congruent to)

AB as the point set $\{C \in S \,|\, OC \cong AB\}$. Note that this set is nonempty: C1 implies that any line through O intersects the circle in two points.

The axioms for congruence of angles are:

C4: *Given a ray $\overrightarrow{AB}$ and an angle $\angle B'A'C'$, there are angles $\angle BAE$ and $\angle BAF$ on opposite sides of $\overline{AB}$ such that $\angle BAE \cong \angle BAF \cong \angle B'A'C'$.*

C5: $\cong$ *is an equivalence relation on the set of angles.*

C6: *Given triangles ABC and $A'B'C'$. If $AB \cong A'B'$, $AC \cong A'C'$ and $\angle BAC \cong \angle B'A'C'$, then the two triangles are congruent — i. e. $BC \cong B'C'$, $\angle ABC \cong \angle A'B'C'$ and $\angle BCA \cong \angle B'C'A'$.*

C4 and C5 are the obvious analogues of C1 and C2, but note that C4 says that we can construct an arbitrary angle on *both* sides of a given ray. C6 says that a triangle is determined up to congruence by any angle and its adjacent sides. This statement is often referred to as the "SAS" (side–angle–side) congruence criterion.

In the Euclidean plane $\mathbb{R}^2$ we define congruence as equivalence under actions of the *Euclidean group* of transformations of $\mathbb{R}^2$. This is generated by rotations and translations, and can also be characterized as the set of transformation of $\mathbb{R}^2$ which preserve all distances. It is quite instructive to prove that the congruence axioms hold with this definition.

These three groups contain the most basic axioms, and they are sufficient to prove a large number of propositions in book I of "Elements". However, when we begin to study circles and "constructions with ruler and compass", we need criteria saying that circles intersect (have common points) with other circles or lines when our intuition tells us that they should. The next axiom provides such a criterion.

First a couple of definitions:

Definition: Let Γ be a circle with center O and radius OA. We say that a point B is *inside* Γ if $OB < OA$ and *outside* if $OA < OB$.

We say that a line or another circle is *tangent to* Γ if they have exactly one point in common with Γ.

We can now formulate Hilbert's axiom E:

E: *Given two circles Γ and Δ such that Δ contains points both inside and outside Γ. Then Γ and Δ have common points. (They "intersect".)*

(It follows from the other axioms that they will then intersect in exactly *two* points.) This is an example of what we call a *continuity axiom.* The following variation is actually a consequence of axiom E:

E': *If a line l contains points both inside and outside the circle Γ, then l and Γ will intersect. (Again in exactly two points.)*

Hilbert gives the *Axiom of parallels* the following formulation — often called "Playfair's axiom" (after John Playfair i 1795, although apparently it goes back to Proclus in the fifth century):

P: *(Playfair's axiom) Given a line l and a point P not on the line. Then there is at most one line m through P which does not intersect l.*

If the lines m and l do not intersect, we say that they are *parallel*, and we write $m \parallel l$. The *existence* of a line m through P parallel to l can be shown to follow from the other axioms, so the real content of the axiom is the uniqueness.

With these axioms we are able to prove all the results in Euclid's "Elements" I–IV, but they do not yet determine the Euclidean plane uniquely. The standard plane ($\mathbb{R}^2$ with the structure defined so far) is an example, and it is an instructive exercise to prove this in detail, but we obtain other examples by replacing the real numbers by another ordered field where every element has a square root! For uniqueness we need a stronger continuity axiom, as for instance *Dedekind's axiom:*

D: *If a line l is a disjoint union of two subset T_1 and T_2 such that all the points of T_1 are on the same side of T_2 and vice versa, then there is a unique point $A \in l$ such that if $B_1 \in T_1$ and $B_2 \in T_2$, then either $A = B_1$, $A = B_2$ or $B_1 * A * B_2$.*

This is a completeness axiom with roots in Dedekind's definition of the real numbers, and an important consequence is that the geometry on any line can be identified with the geometry on $\mathbb{R}$. One can show that it implies axiom E, and together with the groups of axioms I*, B*, C* and P it does determine Euclidean geometry completely.

Finally we mention that axiom D also implies another famous continuity axiom, the *Axiom of Archimedes:*

A: *Given two segments AB and CD, we can find points $C = C_0, \ldots, C_n$ on $\overrightarrow{CD}$, such that $C_iC_{i+1} \cong AB$ for every $i < n$ and $CD < CC_n$.*

("Given a segment AB, then every other segment can be covered by a finite number of congruent copies of AB".)

Using this axiom we can introduce notions of distance and length such that AB has length one, say, and a geometry with the axioms I*, B*, C* P, E and A can be identified with a subset of the standard Euclidean plane.

Exercises.

1. Find all incidence geometries with four or five points.

2. Let V be a vector space of dimension at least 2 over a field F. Show that V satisfies I1-3, if we define *lines* to be sets of the form $\{A+tB|t \in F\}$, where $A, B \in V$, $B \neq 0$.

3. Show that $\mathbb{Q}^2$ satisfies axioms B1–4, but $\mathbb{Z}^2$ does not.

4. Prove that 'being on the same side of the line ℓ' is an equivalence relation on the complement of ℓ, with exactly two equivalence classes.

5. Let A and B be distinct points in a geometry satisfying axioms I1–3 and B1–4. Show that we can find a point C such that $A * C * B$.

6. Consider a triangle ABC and a line ℓ not containing any of the vertices. Show that ℓ cannot intersect all three sides AB, BC and AC.

   Why does this prove uniqueness in *Pasch's axiom* (B4)?

7. Assume $A * B * C$ and $B * C * D$. Show that $A * B * D$ and $A * C * D$.

8. Show that $\mathbb{Q}^2$ does not satisfy C1. Try to determine conditions on an algebraic extension F of $\mathbb{Q}$ such that F^2 will satisfy C1.

9. Show that the center of a circle is uniquely determined.

10. Discuss which axioms are needed in order to bisect a given segment.

11. Which axioms are satisfied by $\overline{\mathbb{Q}}^2$, where $\overline{\mathbb{Q}}$ is the algebraic closure of $\mathbb{Q}$?

12. Show that the axiom of Archimedes can be used to define a length function on segments.

13. Suppose given a geometry with incidence, betweenness and congruence, and let r be a ray with vertex $\mathcal{O}$. Let ℓ be the unique line containing r.

    Show that we can give ℓ the structure of an ordered abelian group, with $\mathcal{O}$ as neutral element and such that $\mathcal{P} \geqslant \mathcal{O}$ if and only if $\mathcal{P} \in r$.

    Show that two different rays give rise to isomorphic ordered groups.

14. Give the vector space $\mathbb{R}^2$ its standard inner product. Show that a map $\phi : \mathbb{R}^2 \to \mathbb{R}^2$ preserves distances if and only if it can be written $\phi(x) = Ax + b$, where A is an orthogonal 2×2 matrix and b is a vector.

Chapter 2

An introduction to hyperbolic geometry

Introduction.

Among Euclid's axioms, the parallel axiom has always been the one causing the most trouble. Already from the beginning it was recognized as less obvious than the other axioms, and during more than two thousand years of fascinating mathematical history, geometers were trying to either prove it from the other axioms, or replace it by something more obvious but with the same consequences. Today we know that the reason they did not succeed, is that there exist geometries where the axiom is not satisfied, but where the remaining axioms are still valid. One may wonder why this was not realized earlier, but we must remember that geometry throughout all this time was concerned with a description of the world "as it is", and in the real world a statement like the parallel axiom must either be true or not true. Euclid's axioms do not *define* geometry; they describe more precisely what kind of arguments we are allowed to use when proving new results about the geometry of the world around us.

But in the 19th century the development of mathematics and mathematical thinking finally brought freedom from this purely descriptive approach to geometry, allowing mathematicians like Lobachevski, Bolyai and Gauss to realize that one might construct perfectly valid geometries where all the other axioms of Euclid hold, but where the parallel axiom fails.

The first concrete such models were constructed by *Beltrami* in 1868, and most of the models we shall present here are due to him, even if some of them have names after other mathematicians.[1] The term *hyperbolic geometry* seems to have been introduced by Felix Klein in 1871.

Instead of Euclid's axiom system we shall use Hilbert's axioms, and we define a *hyperbolic geometry* to be an incidence geometry with betweenness and congruence where Hilbert's axioms hold, except that the parallel axiom is replaced by

H: Given a line l and a point $P \notin l$, there are at least two lines through P which do not intersect l.

We start with a heuristic discussion that may, hopefully, serve as a motivation for our models for hyperbolic geometry. Discussing Hilbert's axiom system we observed that an open, convex subset K of the Euclidean plane is a candidate for such a geometry if we define 'lines' to be intersections between K and Euclidean lines (i. e. open *chords*), and betweenness is defined as in $\mathbb{R}^2$. The *Beltrami–Klein model* $\mathbb{K}$ of the hyperbolic plane utilizes a particularly simple such convex set: the interior of the unit disk. Then the incidence- and betweenness axioms will remain satisfied, as will Dedekind's axiom. However, there will clearly exist *infinitely many* lines parallel to a line l through a point P outside the line, hence axiom H will hold instead of axiom P. (Recall that we call two lines parallel if they do not intersect.)

The only missing ingredient of a hyperbolic geometry is therefore a notion of *congruence* satisfying Hilbert's axioms C1–C6. Clearly the usual, Euclidean definition of congruence does not work, since the fundamental axiom C1 breaks down. (Although C2–C6 are still satisfied!) But, inspired by the Euclidean definition of congruence as equivalence under the action of the Euclidean group of transformations, it is natural to see if there is an analogous group of homeomorphisms of the unit disk that might work.

An absolutely essential property these homeomorphisms should have is that they should map all chords to chords. This property is rather difficult to study directly, but there is a geometric trick that will enable us to find sufficiently many such maps, using some elementary results of complex function theory! The trick is to map $\mathbb{K}$ to open subsets of $\mathbb{C}$ by certain homeomorphisms mapping chords to circular arcs. Then the problem is reduced to finding homeomorphism mapping circles to circles, and this is much simpler, leading to the theory of *Möbius transformations*. The Beltrami–Klein model $\mathbb{K}$ is obtained by transporting back the resulting congruence notion.

However, since the theory is computationally (as well as in other respects) much simpler in the homeomorphic models in $\mathbb{C}$ — *the Poincaré*

[1] The exception is the *hyperboloid model* constructed in exercise 2A.6

disk $\mathbb{D}$ and *Poincaré's upper half–plane* $\mathbb{H}$ — these are the models mostly studied. They will be models for hyperbolic geometry where the "lines" are circular arcs (and certain straight lines) in $\mathbb{C}$. The Beltrami–Klein model will only be used for geometric motivation, except for a discussion in the appendix.

Here is an overview of the contents of this chapter. The preparatory Section 1 discussed the transformation of $\mathbb{K}$ into the other models and relations between them. The Möbius transformations — especially those preserving the upper half–plane — are introduced and studied in some depth in Sections 2 and 3. These transformations can be used to define a congruence relation, giving $\mathbb{H}$ the structure of a hyperbolic plane. This is verified in Section 4. In Sections 5 and 6 we define distance and angle measures in $\mathbb{H}$, and in Section 7 we translate everything done so far to the disk model $\mathbb{D}$. Each of the models has its own advantages, and this is exploited in the remaining sections, where we study arc length and area (Section 8) and trigonometry (Section 9).

Some notation: In the different models we are going to introduce ($\mathbb{K}$, $\mathbb{B}$, $\mathbb{D}$, $\mathbb{H}$), the 'lines' of the geometry will be different types of curves. We shall call these curves $\mathbb{K}$–lines, $\mathbb{B}$–lines etc., or simply *hyperbolic* lines if the model is understood or if it doesn't matter which model we use. For example, the $\mathbb{K}$–lines are the open chords in the interior of the unit circle in the Euclidean plane. Similarly, many of our constructions will take place in standard Euclidean $\mathbb{R}^2$ and $\mathbb{R}^3$, and then 'lines', 'circles' etc. will refer to the usual Euclidean notions.

2.1 Stereographic projection.

As a set, $\mathbb{K}$ is just the interior of the unit disk in $\mathbb{R}^2$:

$$\mathbb{K} = \{(x, y) \in \mathbb{R}^2 \,|\, x^2 + y^2 < 1\}.$$

Consider $\mathbb{R}^2$ as the subspace of $\mathbb{R}^3$ where the last coordinate is 0, and let $\mathbb{B}$ be the lower open hemisphere

$$\mathbb{B} = \{(x, y, z) \in R^3 \,|\, x^2 + y^2 + z^2 = 1, z < 0\}.$$

Vertical projection then defines a homeomorphism $\mathbb{K} \approx \mathbb{B}$, mapping the chords in $\mathbb{K}$ onto (open) semi–circles in $\mathbb{B}$ meeting the boundary curve $\{(x, y) \in \mathbb{R}^2 \,|\, x^2 + y^2 = 1\}$ orthogonally. (Perhaps the easiest way to see this is to consider the image of a chord as the intersection between $\mathbb{B}$ and the plane which contains the chord and is parallel to the z-axis. Then, by symmetry, the image is half of the intersection of this plane with the sphere.) Defining such half–circles as $\mathbb{B}$–lines, we obtain another model $\mathbb{B}$ with the same properties as $\mathbb{K}$.

We now use *stereographic projection* to map $\mathbb{B}$ back to $\mathbb{R}^2$. The version of stereographic projection that we shall use here is the homeomorphism $S^2 - (0,0,1) \approx \mathbb{R}^2$ defined as follows: If P is a point in $S^2 - (0,0,1)$, there is a uniquely determined straight line in $\mathbb{R}^3$ through P and (0,0,1), and this line meets $\mathbb{R}^2$ in a unique point. This defines a map $\Phi : S^2 - (0,0,1) \to \mathbb{R}^2$ which clearly is both injective and surjective. A simple argument using similar triangles (see fig.1) shows that Φ is given by the formula

$$\Phi(x,y,z) = \left(\frac{x}{1-z}, \frac{y}{1-z}\right), \tag{2.1.1}$$

and the inverse map is given by

$$\Phi^{-1}(u,v) = \left(\frac{2u}{u^2+v^2+1}, \frac{2v}{u^2+v^2+1}, \frac{u^2+v^2-1}{u^2+v^2+1}\right). \tag{2.1.2}$$

These maps are both continuous, hence inverse homeomorphisms.

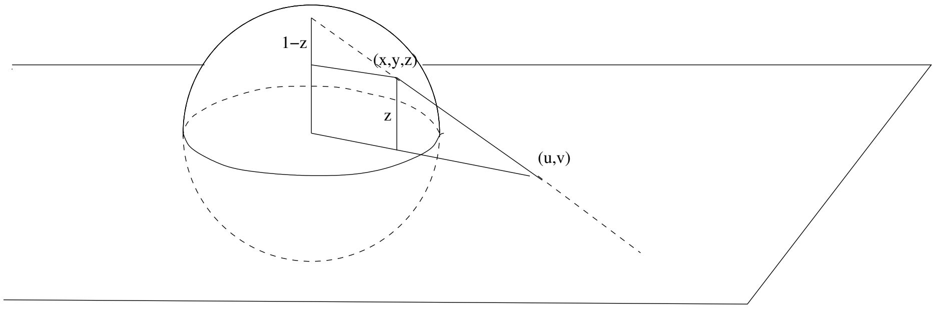


Fig. 2.1.1: Stereographic projection

In the following two Lemmas we state some important properties of stereographic projection.

Lemma 2.1.1. *Let $\mathcal{C}$ be a circle on S^2.*

(i) *If $(0,0,1) \notin \mathcal{C}$, then $\Phi(\mathcal{C})$ is a circle in $\mathbb{R}^2$.*

(ii) *If $(0,0,1) \in \mathcal{C}$, then $\Phi(\mathcal{C} - (0,0,1))$ is a straight line in $\mathbb{R}^2$*

Proof. The circle $\mathcal{C}$ is the intersection between S^2 and a plane defined by an equation $ax + by + cz = d$, say, and $(0, 0, 1) \in \mathcal{C}$ if and only if $c = d$. Now substitute x, y and z from formula (2.1.1 for Φ and get

$$\frac{2au}{u^2 + v^2 + 1} + \frac{2bv}{u^2 + v^2 + 1} + \frac{c(u^2 + v^2 - 1)}{u^2 + v^2 + 1} = d\,.$$

Clearing denominators and collecting terms then yields

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

This is the equation of a line if $c = d$ and a circle if $c \neq d$. □

To formulate the next Lemma, recall that to give a curve an *orientation* is to choose a sense of direction along the curve. In all cases of interest to us, this can be achieved by choosing a nonzero tangent vector at every point, varying continuously along the curve. The *angle* between two oriented curves intersecting in a point P is then the angle between the tangent vectors at P.

Lemma 2.1.2. *Φ preserves angles — i. e. if $\mathcal{C}$ and $\mathcal{C}'$ are oriented circles on S^2 intersecting in a point P at an angle θ, then their images under Φ intersect in $\Phi(P)$ at the same angle.*

Remark. Here we are only interested in unoriented angles, i. e. we do not distinguish between the angle between $\mathcal{C}$ and $\mathcal{C}'$ and the angle between $\mathcal{C}'$ and $\mathcal{C}$. Then we can restrict to angles between 0 and π, and this determines θ uniquely. Note that in this range θ is also determined by $\cos(\theta)$. (With appropriate choices of orientations of S^2 and $\mathbb{R}^2$ the result is also true for oriented angles, but we shall not need this.)

Proof. By rotational symmetry around the z-axis we may assume that the point P lies in a fixed meridian, so we assume that $P = (0, y, z)$ with $y \geq 0$. Furthermore, it clearly suffices to compare each of the circles with this meridian, i. e. we may assume $\mathcal{C}'$ is the circle $x = 0$, with oriented tangent direction $(0, -z, y)$ at the point $(0, y, z)$. The image of this circle under Φ is the y-axis with tangent direction $(0, 1)$.

Observe that Φ can be extended to $\{(x, y, z) \in \mathbb{R}^3 | z < 1\}$ by the same formula (2.1.1), and that tangential curves will map to tangential curves (by the chain rule). Hence we can replace $\mathcal{C}$ by any curve in $\mathbb{R}^3$ with the same oriented tangent as $\mathcal{C}$ in P — e. g. a straight line. This line can be parametrized by

$$\theta(t) = (0, y, z) + t(\alpha, \beta, \gamma) = (t\alpha, y + t\beta, z + t\gamma),$$

where $\alpha^2+\beta^2+\gamma^2=1$ and $(\alpha,\beta,\gamma)\cdot(0,y,z)=\beta y+\gamma z=0$. The angle u between this line and the meridian $\mathcal{C}'$ is determined by

$$\cos u=(\alpha,\beta,\gamma)\cdot(0,-z,y)=-\beta z+\gamma y\,.$$

Now consider the image of this line under Φ. This is parametrized by

$$\omega(t)=\Phi(\theta(t))=\left(\frac{t\alpha}{1-z-t\gamma},\frac{y+t\beta}{1-z-t\gamma}\right),$$

(Restrict t such that $1-z-t\gamma>0$.) It is geometrically obvious that this is again a straight line, and to see this from the formula for $\omega(t)$, note that

$$\begin{aligned}\omega(t)-\omega(0)=\omega(t)-\Phi(P)&=\left(\frac{t\alpha}{1-z-t\gamma},\frac{y+t\beta}{1-z-t\gamma}-\frac{y}{1-z}\right)\\&=\frac{t}{1-z-t\gamma}\left(\alpha,\frac{\beta-\beta z+\gamma y}{1-z}\right).\end{aligned}$$

(Straightforward calculation.) But this has constant direction given by the vector $V=(\alpha,\dfrac{\beta-\beta z+\gamma y}{1-z})$.

It remains to check that the angle v between V and the positive y-axis is equal to u, or, equivalently, that $\cos u=\cos v=\dfrac{V\cdot(0,1)}{||V||}$.

Recall that $\cos u=-\beta z+\gamma y$ and $\beta y+\gamma z=0$. Then

$$z\cos u=-\beta z^2+\gamma yz=-\beta z^2-\beta y^2=-\beta,$$

since $y^2+z^2=1$. Similarly, $y\cos u=\gamma$. It follows that

$$\frac{\beta-\beta z+\gamma y}{1-z}=\frac{-z\cos u+\cos u}{1-z}=\cos u\,,$$

hence $V=(\alpha,\cos u)$. Moreover, $\beta^2+\gamma^2=z^2\cos^2 u+y^2\cos^2 u=\cos^2 u$, so $||V||=\alpha^2+\cos^2 u=\alpha^2+\beta^2+\gamma^2=1$. But then $\cos v=\cos u$. □

Figure 2.1.2 illustrates a geometric proof of Lemma 2.1.2. N is the "north pole", P' is a point on S^3 and $P=\Phi(P)$. The lines m and n in $\mathbb{R}^2$ intersect in P. The crucial observation is that the image under Φ^{-1} of a line is a circle — the intersection between S^3 and the plane through the line and N. Moreover, the tangent line at a point Q of this circle is the intersection between the plane and the tangent plane of S^3 at Q.

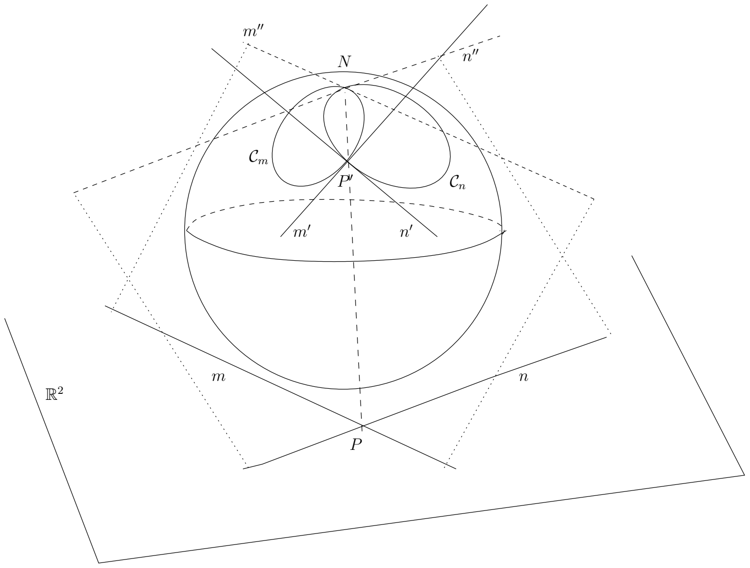


Fig. 2.1.2: Stereographic projection preserves angles

Accordingly, let $\mathcal{C}_m$ be the circle $\Phi^{-1}(m)$, and let m' and m'' be the tangents to $\mathcal{C}_m$ at P' and N. $\mathcal{C}_n$, n', n'' are defined similarly from n. Then by symmetry the angle between m' and n' is the same as the angle between m'' and n''. But since the tangent plane of S^3 at N is parallel to $\mathbb{R}^2$, the angle m'' is parallel to m and n'' is parallel to n. Hence the angle between m'' and n'' is the same as the angle between m and n.

There are, of course, many ways to identify $\mathbb{B}$ with an open hemisphere of S^2, and we obtain a homeomorphic image in the plane by stereographic projection as long as we avoid the point (0,0,1). The Lemmas above imply that every such model will be bounded by a circle or a straight line in $\mathbb{R}^2$, and the 'lines' will correspond to circular arcs or straight lines meeting the bounding curve orthogonally. We make will use of two models obtained this way, both being named after the great French mathematician Henri Poincaré:

1. *Poincaré's disk model* $\mathbb{D}$ is obtained by choosing $\mathbb{B}$ to be the lower hemisphere, as before. Then the image is again the interior of the unit circle, but the $\mathbb{D}$-lines are now either diameters or circular arcs perpendicular to the boundary circle.

2. The *Poincaré upper half–plane* $\mathbb{H} = \{(x, y) \mid y > 0\}$ is the image of the

open hemisphere $\{(x, y, z) \in S^2 \mid y > 0\}$. The $\mathbb{H}$–lines are either semicircles with center on the x-axis or straight lines parallel to the y-axis.

When we analyze these models we shall henceforth identify $\mathbb{R}^2$ with the complex plane $\mathbb{C}$ and make use of the extra structure and tools we have available there (complex multiplication, complex function theory, etc.). Thus, as sets we make the identifications $\mathbb{D} = \{z \in \mathbb{C} \mid |z| < 1\}$ and $\mathbb{H} = \{z \in \mathbb{C} \mid \operatorname{Im} z > 0\}$.

$\mathbb{D}$ is the most symmetric of the two models and therefore often the one best suited for geometric arguments. But we will see that $\mathbb{H}$ is better for analyzing and describing the notion of congruence. Therefore this is where we begin our analysis.

Notation: In both models the hyperbolic lines have natural extensions to the boundary curve. (The unit circle for $\mathbb{D}$ and the real line for $\mathbb{H}$.) These extension points we refer to as *endpoints* of the hyperbolic lines, although they are not themselves points on the lines. Analogously, we also say that ∞ is an endpoint of a vertical $\mathbb{H}$–line.

Exercises for 2.1

1. Derive the formulas for Φ and its inverse.

2. We can also define stereographic projection from $(0, 0, -1)$ instead of $(0, 0, 1)$. Let Φ_- be the resulting map.

   Determine the map $\Phi_- \circ \Phi^{-1}$. (We identify $\mathbb{R}^2$ with $\mathbb{C}$.)

3. If F is an identification between the two hemispheres we use in the definitions of $\mathbb{H}$ and $\mathbb{D}$, the map $\Phi \circ F \circ \Phi^{-1}$ will be a homeomorphism between the two models. Find a formula for such a homeomorphism. (Choose F as simple as possible.)

4. (a) Show that $z \mapsto z^{-1} : \mathbb{C} - \{0\} \to \mathbb{C} - \{0\}$ corresponds to a rotation of S^2 via stereographic projection.

   (b) Which self–map of $\mathbb{C} - \{0\}$ does the antipodal map $x \mapsto -x$ on $S^2 - \{0, 0, \pm 1\}$ correspond to?

2.2 Congruence in $\mathbb{H}$; Möbius transformations

As noted before, we define congruence In Euclidean geometry as equivalence under the *Euclidean group* $E(2)$ of "rigid movements", generated by orthogonal linear transformations $x \mapsto Ax$ and translations $x \mapsto x + b$. This means that the congruence relation $\cong$ is defined by

Segments: $AB \cong A'B' \iff$ there is a $g \in E(2)$ such that $g(AB) = A'B'$.
Angles: $\angle BAC \cong \angle B'A'C' \iff$ there is a $g \in E(2)$ such that $g(\overrightarrow{AB}) = \overrightarrow{A'B'}$ and $g(\overrightarrow{AC}) = \overrightarrow{A'C'}$.

Observe that

> Every element of $E(2)$ maps straight lines to straight lines, and if A and A' are points on the lines l and l', there is a $g \in E(2)$ such that $g(l) = l'$ and $g(A) = A'$.

We now wish to do something similar in the case of $\mathbb{H}$. Motivated by the Euclidean example, we will look for a group G of bijections of $\mathbb{H}$ to itself such that

> Every element of G maps $\mathbb{H}$–lines to $\mathbb{H}$–lines, and if A and A' are points on the $\mathbb{H}$-lines l and l', there is a $g \in G$ such that $g(l) = l'$ and $g(A) = A'$.

We will show that there exists such a group, consisting of so–called *Möbius transformations* preserving the upper half–plane.

From complex function theory we know that meromorphic functions with at most poles at ∞ can be thought of as functions $f : \overline{\mathbb{C}} \to \overline{\mathbb{C}}$, where $\overline{\mathbb{C}}$ is the *extended complex plane* or the *Riemann sphere* $\mathbb{C} \cup \{\infty\}$, with a topology such that stereographic projection extends to a homeomorphism $\overline{\Phi} : S^2 \approx \overline{\mathbb{C}}$. Lemma 2.1.1 says that $\overline{\Phi}$ maps circles to curves in $\overline{\mathbb{C}}$ that are either circles in $\mathbb{C}$ or of the form $l \cup \{\infty\}$, where l is a (real) line in $\mathbb{C}$. It is convenient not to have to distinguish between the two cases, so we call all of these curves $\overline{\mathbb{C}}$*–circles.*

Similarly, we let $\overline{\mathbb{R}}$ denote $\mathbb{R} \cup \infty$, considered as a subspace of $\overline{\mathbb{C}}$.

By Lemma 2.1.2 the angle between oriented such circles at an intersection point in $\mathbb{C}$ is the same as between the corresponding circles on S^2. Moreover, if they intersect in two points, the two angles will be the same. Hence we can also define the angle at ∞ between two circles intersecting there — i. e. two lines in $\mathbb{C}$. If they intersect in a point $P \in \mathbb{C}$, the angle of intersection at ∞ is the same as the angle at P. If they are parallel, the angle of intersection at ∞ is 0. With this definition, Φ is also angle preserving at ∞.

For a meromorphic function $f : \overline{\mathbb{C}} \to \overline{\mathbb{C}}$ to be a homeomorphism, it must have exactly one pole and one zero — hence it must have the form

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

where $a, b, c, d \in \mathbb{C}$. Solving the equation $w = f(z)$ with respect to z we get

$$z = g(w) = \frac{-dw+b}{cw-a},$$

and (formally) substituting back again:

$$f(g(w)) = \frac{(ad-bc)w}{ad-bc} \quad \text{and} \quad g(f(z)) = \frac{(ad-bc)z}{ad-bc}.$$

Therefore f is invertible with g as inverse if $ad - bc \neq 0$. If $ad - bc = 0$ these expressions have no meaning, but it is easy to see that in that case $f(z)$ is constant. Hence we have:

> *The function* $f(z) = \dfrac{az+b}{cz+d}$ *defines a homeomorphism of* $\overline{\mathbb{C}}$ *if and only if* $ad - bc \neq 0$.

Such a function is called a *fractional linear transformation* — FLT for short. Here are some crucial properties of FLT's:

Lemma 2.2.1. (i) *An FLT maps* $\overline{\mathbb{C}}$*–circles to* $\overline{\mathbb{C}}$*–circles.*
(ii) *An FLT preserves angles between* $\overline{\mathbb{C}}$*–circles.*

Proof. (i) Note that we may write the equations of both circles and straight lines in $\mathbb{C}$ as $\lambda(x^2+y^2) + \alpha x + \beta y + \gamma = 0$, where $\lambda, \alpha, \beta, \gamma$ are real numbers and $z = x + iy$; $\lambda = 0$ for straight lines and $\lambda \neq 0$ for circles. Using that $x^2 + y^2 = z\bar{z}$, $x = (z+\bar{z})/2$ and $y = (z - \bar{z})/2i = (\bar{z} - z)i/2$, we can write the equation as

$$\lambda z\bar{z} + \mu z + \bar{\mu}\bar{z} + \gamma = 0,$$

where $\mu = (\alpha - i\beta)/2$. Hence we need to show that if z satisfies such an equation and $w = f(z)$ for an FLT f, then w satisfies a similar equation. This can be checked by writing $z = f^{-1}(w) = \dfrac{aw+b}{cw+d}$ and substituting:

$$\lambda \frac{aw+b}{cw+d} \cdot \overline{\frac{aw+b}{cw+d}} + \mu \frac{aw+b}{cw+d} + \overline{\mu}\overline{\frac{aw+b}{cw+d}} + \gamma = 0.$$

If we multiply this equation by $(cw + d)(\overline{cw + d})$ and simplify, we end up with an expression just like the one we want.

(ii) This is a consequence of a general fact in complex function theory. We say that a differentiable map is angle–preserving, or *conformal*, if it maps two intersecting curves to curves meeting at the same angle. It then follows from the geometric interpretation of the derivative that a complex function is conformal in a neighborhood of any point where it is analytic with nonzero derivative.

For a more direct argument in our case, see Exercise 3. □

Remark 2.2.2. As in Lemma 2.1.2 it is not difficult to show that the same result is true for oriented angles (with a suitable notion of orientation that also applies to $\infty \in \overline{\mathbb{C}}$), but we do not need that. An angle–preserving map (in the orientable sense) is called *conformal*, and it follows from the geometric interpretation of the derivative that a complex function is conformal in a neighborhood of any point where it is analytic with nonzero derivative.

The oriented version of Lemma 2.1.2 says that stereographic projection also is conformal.

The word 'linear' in FLT is related to the following remarkable observation:

Let $f(z) = \dfrac{az+b}{cz+d}$ and $g(z) = \dfrac{a'z+b'}{c'z+d'}$. A little calculation gives

$$(f \circ g)(z) = f(g(z)) = \frac{a\,g(z)+b}{c\,g(z)+d} = \frac{(aa'+bc')z+(ab'+bd')}{(ca'+dc')z+(cb'+dd')}.$$

This formula tells us two things. First, it means that the composition of two FLT's is a new FLT. We showed earlier that the inverse of an FLT is an FLT, and the identity map is trivially also an FLT. ($z = \frac{1z+0}{0z+1}$.) Hence the set of fractional linear transformations forms a *group* under composition. This group will be denoted *Möb*$^+(\mathbb{C})$. (The "even complex Möbius transformations".)

Secondly, it is possible to calculate with FLT's as with *matrices:* Evidently the matrix $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ determines f completely, and the condition $ad - bc \neq 0$ simply means that this matrix is invertible. In the same way g is determined by $\begin{bmatrix} a' & b' \\ c' & d' \end{bmatrix}$, and the calculation above shows that $f \circ g$ is determined by the product matrix $\begin{bmatrix} a & b \\ c & d \end{bmatrix} \cdot \begin{bmatrix} a' & b' \\ c' & d' \end{bmatrix}$.

The set of invertible 2×2–matrices over $\mathbb{C}$ forms a group — *the general linear group* $GL_2(\mathbb{C})$ — and we have shown that there is a surjective group homomorphism from $GL_2(\mathbb{C})$ onto $M\ddot{o}b^+(\mathbb{C})$. This homomorphism is not injective, since if $k \neq 0$, then $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ and $\begin{bmatrix} ka & kb \\ kc & kd \end{bmatrix} = k \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ will determine the same map. However, this is the only ambiguity (Exercise 4), and we get an *isomorphism* between the group $M\ddot{o}b^+(\mathbb{C})$ of fractional linear transformations and the quotient group $PGL_2(\mathbb{C}) = GL_2(\mathbb{C})/D$ (the *projective* linear group), where $D = \{uI | \, u \in \mathbb{C} - \{0\}\}$ and I is the identity matrix.

(A more conceptual explanation of the connection between FLT's and linear algebra belongs to *projective geometry* — see Exercise 11).

The next Lemma and its Corollary tell us exactly what freedom we have in prescribing values of fractional linear transformations. In fact, it provides us with a method of constructing FLT's with prescribed values.

Lemma 2.2.3. *Given three distinct points* z_1, z_2 *and* z_3 *in* $\overline{\mathbb{C}}$. *Then there exists a uniquely determined FLT* f *such that* $f(z_1) = 1$, $f(z_2) = 0$ *and* $f(z_3) = \infty$.

Proof. Existence: Suppose first that none of the z_i's is at ∞. Then we define

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

In the three other cases:

$$\text{if } z_1 = \infty: \qquad f(z) = \frac{z - z_2}{z - z_3},$$

$$\text{if } z_2 = \infty: \qquad f(z) = \frac{z_1 - z_3}{z - z_3},$$

$$\text{if } z_3 = \infty: \qquad f(z) = \frac{z - z_2}{z_1 - z_2}.$$

Uniqueness: Suppose $g(z)$ has the same properties and consider the composition $h = g \circ f^{-1}$. This is a new FLT with $h(1) = 1$, $h(0) = 0$ and $h(\infty) = \infty$. The last condition implies that h must have the form $h(z) = az + b$, and the first two conditions then determine $a = 1$ og $b = 0$. Thus $(g \circ f^{-1})(z) = z$ for all z, hence $f = g$. □

Definition 2.2.4. The element $f(z) \in \overline{\mathbb{C}}$ depends on the four variables (z, z_1, z_2, z_3) and is denoted $[z, z_1, z_2, z_3]$. It is defined as an element of $\overline{\mathbb{C}}$ whenever z_1, z_2 and z_3 are distinct points of $\overline{\mathbb{C}}$ and it has the following geometric interpretation:

If z_2 and z_3 span a Euclidean segment S, every point in S will divide it in two and we can compute the ratio between the lengths of the pieces. If we do this for two points z and z_1 in S, then $|[z, z_1, z_2, z_3]|$ is the quotient of the two ratios we obtain. Because of this, $[z, z_1, z_2, z_3]$ is traditionally called the *cross–ratio* of the four points, and it plays a very important role in geometry. We shall meet it again later, and some of its properties are given below, in Proposition 2.2.10.

Corollary 2.2.5. *Given two triples (z_1, z_2, z_3) og (w_1, w_2, w_3) of distinct points in $\overline{\mathbb{C}}$. Then there exists a unique FLT f such that $f(z_i) = w_i$, $i = 1, 2, 3$. If all six points lie in $\overline{\mathbb{R}}$, then f may be expressed with real coefficients — i. e. $f(z) = \dfrac{az+b}{cz+d}$ with a, b, c, d all real.*

Proof. By Lemma 2.2.3 we can find unique FLT's h and g such that $h(z_1) = 1$, $h(z_2) = 0$, $h(z_3) = \infty$, and $g(w_1) = 1$, $g(w_2) = 0$ $g(w_3) = \infty$. Let $f = g^{-1}h$. Then $f(z_i) = w_i$, $i = 1, 2, 3$.

Suppose also f' maps z_i to w_i. Then gf and gf' are both FLT's as in Lemma 2.2.3, and because of the uniqueness $gf = gf'$. Consequently $f = f'$.

The final assertion of the Corollary follows from the formulas in the proof of 2.2.3. They show that h and g have real coefficients, hence so does $f = g^{-1}h$. □

Remark 2.2.6. The *existence* of such f means that the group *Möb*$^+(\mathbb{C})$ acts *transitively* on the set of such triples. The *uniqueness* says that if two fractional linear transformations have the same values at three points, then they are equal. In particular, an FLT fixing three points is the identity map.

Note that $f(z)$ is characterized by the equation

$$[f(z), w_1, w_2, w_3] = [z, z_1, z_2, z_3]\,.$$

Our next observation is that *Möb*$^+(\mathbb{C})$ also acts transitively on the set of $\overline{\mathbb{C}}$*–circles.* The reason for this is that three distinct points in $\overline{\mathbb{C}}$ determine a unique $\overline{\mathbb{C}}$–circle containing all of them.

Corollary 2.2.7. *Given two circles $\mathcal{C}_1$ and $\mathcal{C}_2$ in $\overline{\mathbb{C}}$. Then there exists a fractional linear transformation f such that $f(\mathcal{C}_1) = \mathcal{C}_2$.*

Proof. Choose three distinct points (z_1, z_2, z_3) on $\mathcal{C}_1$ and (w_1, w_2, w_3) on $\mathcal{C}_2$, and let f be as in the Corollary above. Then $f(\mathcal{C}_1)$ is a $\overline{\mathbb{C}}$–circle which contains w_1, w_2 and w_3 — i. e. $\mathcal{C}_2$. □

We now want to determine the fractional linear transformations f which restrict to homeomorphisms of the upper half–plane. Such an f is characterized by $f(\overline{\mathbb{R}}) = \overline{\mathbb{R}}$, and $\operatorname{Im} f(z) > 0$ if $\operatorname{Im} z > 0$. Here $\overline{\mathbb{R}} = \mathbb{R} \cup \{\infty\} \subset \overline{\mathbb{C}}$.

Proposition 2.2.8. *A fractional linear transformation restricts to a homeomorphism of* $\mathbb{H}$ *if and only if it can be written on the form* $f(z) = \dfrac{az+b}{cz+d}$, *where* a, b, c, d *are real and* $ad - bc = 1$.

Such FLT's map $\mathbb{H}$*–lines to* $\mathbb{H}$*–lines.*

Proof. Corollary 2.2.5 says that if $f(z) = \dfrac{az+b}{cz+d}$ restricts to a homeomorphism of $\mathbb{H}$, then a, b, c, d can be chosen to be real, since $f(\overline{\mathbb{R}}) = \overline{\mathbb{R}}$. Conversely, $f(\overline{\mathbb{R}}) = \overline{\mathbb{R}}$ if a, b, c, d are real.

A short calculation gives

$$\begin{aligned} f(z) &= \frac{(az+b)(c\bar{z}+d)}{(cz+d)(c\bar{z}+d)} \\ &= \frac{ac|z|^2 + (ad+bc)\mathrm{Re}\, z + bd}{|cz+d|^2} + \frac{(ad-bc)\mathrm{Im}\, z}{|cz+d|^2}\, i\,. \end{aligned} \tag{2.2.1}$$

It follows that f preserves the upper half–plane if and only if $ad - bc > 0$. Hence, if we multiply a, b, c and d by $1/\sqrt{ad-bc}$, f is as asserted.

The last claim follows immediately from the fact that fractional linear transformations preserve $\overline{\mathbb{C}}$–circles and angles between them. Every $\mathbb{H}$–line determines a $\overline{\mathbb{C}}$–circle which meets $\mathbb{R}$ orthogonally, and since f preserves angles and $f(\overline{\mathbb{R}}) = \overline{\mathbb{R}}$, the images of these circles will also meet $\mathbb{R}$ orthogonally. □

The fractional linear transformations restricting to homeomorphisms of $\mathbb{H}$ form a subgroup of of $\textit{Möb}^+(\mathbb{C})$ denoted $\textit{Möb}^+(\mathbb{H})$.It can also be described using matrices, as follows:

Let $SL_2(\mathbb{R})$ be *the special linear group* — the group of real 2×2–matrices with determinant 1. The only multiples of the identity matrix in $SL_2(\mathbb{R})$ are $\pm I$, hence, arguing as before, we get an isomorphism between $\textit{Möb}^+(\mathbb{H})$ and the quotient group $PSL_2(\mathbb{R}) = SL_2(\mathbb{R})/(\pm I)$.

$\textit{Möb}^+(\mathbb{C})$ does not contain all circle–preserving homeomorphisms of $\overline{\mathbb{C}}$. *Complex conjugation* is also circle–preserving but not even complex analytic. (Define $\overline{\infty} = \infty$.) We define the group of *complex Möbius transformations,* $\textit{Möb}(\mathbb{C})$,to be the group of homeomorphisms of $\overline{\mathbb{C}}$ generated by the fractional linear transformations and complex conjugation.

Proposition 2.2.9. (1) *Every complex Möbius transformation can be written on exactly one of the forms*

$$f(z) = \frac{az+b}{cz+d} \quad \textit{or} \quad f(z) = \frac{a\bar{z}+b}{c\bar{z}+d}\,, \quad \textit{where} \;\; a,\, b,\, c,\, d \in \mathbb{C} \;\textit{and}\; ad-bc = 1\,.$$

(2) *The complex Möbius transformations preserving* $\mathbb{H}$ *can be written either as*

(i) $f(z)=\dfrac{az+b}{cz+d}\,,\quad$ *where* $a,\, b,\, c,\, d \in \mathbb{R}$ *and* $ad-bc=1\,,$ *or as*

(ii) $f(z)=\dfrac{a\bar{z}+b}{c\bar{z}+d}\,,\quad$ *where* $a,\, b,\, c,\, d \in \mathbb{R}$ *og* $ad-bc=-1\,.$

Proof. (1) Let S be the subset of the set of homeomorphisms of $\mathbb{C}$ given by such expressions. Clearly $S \subseteq$ *Möb*$(\mathbb{C})$, and S contains *Möb*$^+(\mathbb{C})$ and complex conjugation. Therefore it suffices to check that S is closed under composition and taking inverses, and this is an easy calculation. Moreover, the second expression is not even complex differentiable, hence no function can be written both ways.

To obtain determinant 1 we divide numerator and denominator by a square root of $ad-bc$.

(2) $f(z)$ can be written in one of the two types in (1). In the first case, the result is given in Proposition 2.2.8. If $f(z) = \dfrac{a\bar{z}+b}{c\bar{z}+d}$, then $g(z) = -\overline{f(z)}$ can be written as in (i). But then $f(z) = -\overline{g(z)}$ automatically has the form given in (ii). □

The representations in Proposition 2.2.9 is not unique, but it follows from the result in Exercise 4 that it is unique up to multiplication of $(a,\, b,\, c,\, d)$ by ± 1.

Let *Möb*$(\mathbb{H})$ be the group of Möbius transformations restricting to homeomorphisms of $\mathbb{H}$. We have shown that every element in *Möb*$(\mathbb{H})$ can be written as one of the two types in (2) of Proposition 2.2.9, and therefore we call these elements the *real Möbius transformations.* Note that complex conjugation, i. e. reflection in the real axis, is not in *Möb*$(\mathbb{H})$, but $f(z) = -\bar{z}$, reflection in the imaginary axis, is. *Möb*$(\mathbb{H})$ is generated by this reflection and *Möb*$^+(\mathbb{H})$.

We use the notation *Möb*$^-(\mathbb{H})$ for the elements in *Möb*$(\mathbb{H})$ of type (ii). These do not form a subgroup, but *Möb*$(\mathbb{H})$ is the disjoint union of *Möb*$^+(\mathbb{H})$ and *Möb*$^-(\mathbb{H})$. In fact, *Möb*$^+(\mathbb{H}) \subset$ *Möb*$(\mathbb{H})$ is a normal subgroup of index two, and *Möb*$^-(\mathbb{H})$ is the coset containing $-\bar{z}$.

$M\ddot{o}b(\mathbb{H})$ is the group we shall use to define congruence in $\mathbb{H}$, but before we show that Hilbert's axioms hold, we will analyze the elements in $M\ddot{o}b(\mathbb{H})$ further (next section) and show that they can be classified into a few very simple standard types.

We end this section with some properties satisfied by the cross ratio.

Proposition 2.2.10. *Assume that z, z_1, z_2 and z_3 are four distinct points in $\overline{\mathbb{C}}$, and let $\rho = [z, z_1, z_2, z_3]$. Then*

(i) $[z_1, z, z_2, z_3] = [z, z_1, z_3, z_2] = \dfrac{1}{\rho}$, and $[z, z_2, z_1, z_3] = 1 - \rho$.

(ii) z, z_1, z_2 and z_3 all lie on the same $\overline{\mathbb{C}}$–circle if and only if the cross–ratio $[z, z_1, z_2, z_3]$ is real.

(iii) $[g(z), g(z_1), g(z_2), g(z_3)] = [z, z_1, z_2, z_3]$ if g is a fractional linear transformation.

Proof. (i) Recall that the mapping $w \mapsto f(w) = [w, z_1, z_2, z_3]$ is the fractional linear transformation which is uniquely determined by its values $1, 0$ and ∞ at the points z_1, z_2 and z_3, respectively. Then the identities $[w, z_1, z_3, z_2] = \dfrac{1}{f(w)}$ and $[w, z_2, z_1, z_3] = 1 - f(w)$ follow easily by inspection. Setting $w = z$ proves two of the identities.

Note that since $f(z_2) = 0$ and $f(z_3) = \infty$, ρ is not 0 or ∞. Therefore $g(w) = \dfrac{1}{\rho} f(w)$ defines a new fractional linear transformation. But $g(z) = 1$, $g(z_2) = 0$ and $g(z_3) = \infty$ — hence $g(w) = [w, z, z_2, z_3]$. Consequently,

$$[z_1, z, z_2, z_3] = g(z_1) = \frac{1}{\rho}[z_1, z_1, z_2, z_3] = \frac{1}{\rho}\,.$$

(ii) Let $\mathcal{C}$ be the unique $\overline{\mathbb{C}}$–circle containing the three points z_1, z_2 and z_3. Then $f(\mathcal{C})$ must be the unique $\overline{\mathbb{C}}$–circle containing $1, 0$ and ∞ — i. e. $\overline{\mathbb{R}}$. Likewise, $f^{-1}(\overline{\mathbb{R}}) = \mathcal{C}$. Thus $z \in \mathcal{C}$ if and only if $f(z) \in \overline{\mathbb{R}}$. But since $z \neq z_3$, $f(z) \in \overline{\mathbb{R}}$ means $f(z) \in \mathbb{R}$.

(iii) Let $h(w) = [g(w), g(z_1), g(z_2), g(z_3)]$. This is a composition of two FLT's — hence h is also an FLT. By inspection, $h(z_j) = [z_j, z_1, z_2, z_3]$ for $j = 1, 2, 3$. Therefore $h(w) = [w, z_1, z_2, z_3]$ for all w, by uniqueness. □

Remark 2.2.11. (1) The three transpositions (1,2), (2,3) and (3,4) generate the whole group S_4 — the group of permutations of four letters. Hence (i) can be used to determine the cross ratio of any permutation of the points z, z_1, z_2 and z_3. For examples, see Exercise 8.

It follows that $[z, z_1, z_2, z_3]$ can be defined as long as *three* of the points z, z_1, z_2 and z_3 are distinct, and it can be considered as a fractional linear transformation in each of the variables separately. This observation will be used repeatedly later without any further comment.

(2) The identity in (iii) is not valid for *all* Möbius transformations g. For example, if $g(z) = \bar{z}$, then $[g(z), g(z_1), g(z_2), g(z_3)] = \overline{[z, z_1, z_2, z_3]}$. (Exercise 9.)

Exercises for 2.2

1. Discuss what conditions λ, μ, γ must satisfy for $\lambda z\bar{z} + \mu z + \bar{\mu}\bar{z} + \gamma = 0$ to define an $\mathbb{H}$–line.

2. Let the circle $\mathcal{C}$ be given by the equation $|z - z_0| = r$, and let f be the function $f(z) = 1/z$. When is $f(\mathcal{C})$ a straight line $\cup\{\infty\}$?

3. Show that any FLT can be written as a composition of maps of the following three simple types:

   (i) Translations $z \mapsto z + b,\ b \in \mathbb{C}$,

   (ii) Linear maps $z \mapsto kz,\ k \in \mathbb{C} - \{0\}$,

   (iii) Taking inverse $z \mapsto \dfrac{1}{z}$.

   Use this to give another proof of Lemma 2.2.1. (Hint: You may find Exercise 2.1.4a useful.)

4. Assume $ad - bc \neq 0$. Show that $\dfrac{az + b}{cz + d} = \dfrac{a'z + b'}{c'z + d'}$ for every z if and only if there exists a $k \neq 0$ such that $\begin{bmatrix} a' & b' \\ c' & d' \end{bmatrix} = k \begin{bmatrix} a & b \\ c & d \end{bmatrix}$.

5. Use the method of Corollary 2.2.5 to find explicit fractional linear transformations mapping $\mathbb{H}$ onto $\mathbb{D}$ and vice versa. (Compare with Exercise 2.1.3.)

6. Describe all the elements in *Möb*$(\mathbb{H})$ that map the imaginary axis to itself.

7. (a) Show that $M\ddot{o}b(\mathbb{H})$ acts transitively on the set of all triples of distinct points in $\overline{\mathbb{R}}$. Deduce that $M\ddot{o}b(\mathbb{H})$ acts transitively on the set of pairs of lines in $\mathbb{H}$ with one common endpoint.

   (b) Let ℓ and ℓ' be two $\mathbb{H}$-lines with endpoints z, w and z', w', resp., and let $p \in \ell$ and $p' \in \ell'$ be given points on the lines. Show that there is a unique $f \in M\ddot{o}b^+(\mathbb{H})$ such that $f(z) = z'$, $f(w) = w'$, and $f(p) = p'$.

   (c) Show that $M\ddot{o}b(\mathbb{H})$ does not act transitively on the set of pairs of distinct points in $\mathbb{H}$.

8. Using Remark 2.2.11, show that if $[z_1, z_2, z_3, z_4] = \rho$, then

   $[z_3, z_4, z_1, z_2] = \rho$ and $[z_3, z_2, z_1, z_4] = \dfrac{\rho}{1-\rho}$.

9. Show that $[g(z), g(z_1), g(z_2), g(z_3)] = \overline{[z, z_1, z_2, z_3]}$ for all $g \in M\ddot{o}b(\mathbb{C})$.

10. Give a geometric explanation for Prop. 2.2.10(ii).

    Suppose $z_1, z_2, z_3 \in \overline{\mathbb{R}}$ and consider the function $g(z) = [z, z_1, z_2, z_3]$. Discuss when we have $g \in M\ddot{o}b^+(\mathbb{H})$.

11. Show that $M\ddot{o}b(\mathbb{H})$ is isomorphic to the group $PGL_2(\mathbb{R}) = GL_2(\mathbb{R})/D$, where $D = \{uI | u \in \mathbb{R} - \{0\}\}$.

12. Let $CP^1 = (\mathbb{C}^2 - \{0\})/\sim$, where $\sim$ is the equivalence relation which identifies v and λv, for all $v \in \mathbb{C}^2 - \{0\}$ and $\lambda \in \mathbb{C} - \{0\}$. ($CP^1$ is called *the complex projective line.*)

    Show that multiplication by a matrix in $GL_2(\mathbb{C})$ induces a bijection of CP^1 with itself.

    Verify that $(z_1, z_2) \mapsto z_1/z_2$ defines a bijection $CP^1 \approx \overline{\mathbb{C}}$, and show that via this bijection, multiplication with the matrix $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ corresponds to the fractional linear transformation $\dfrac{az+b}{cz+d}$.

    Why does it now follow immediately that this correspondence is a group homomorphism $GL_2(\mathbb{C}) \to M\ddot{o}b^+(\mathbb{C})$?

2.3 Classification of real Möbius transformations

Since the Möbius transformations play such an important rôle in the theory, we would like to know as much as possible about them, both geometrically and algebraically. The results of this section can be interpreted in both directions. Geometrically we show that up to coordinate shifts, Möbius transformations can be given one of three possible "normal forms", from which it is easy to get a good picture of how they act on $\mathbb{H}$. Algebraically this translates into a classification into conjugacy classes in $\mathit{M\ddot{o}b}(\mathbb{H})$.

Since $\mathit{M\ddot{o}b}(\mathbb{H})$ is isomorphic to a matrix group, this classification could be done completely with tools from linear algebra. What we will do is equivalent to this, but interpreted in our geometric language.

The key to the classification of matrices is the study of eigenvectors, and it is not difficult, using Exercise 2.2.12, to see that in $\mathit{M\ddot{o}b}(\mathbb{C})$ this corresponds to analyzing the *fixpoints* of the transformations, i. e. the solutions in $\overline{\mathbb{C}}$ of the equation $z = f(z)$. By a "change of coordinates" we reduce to a situation where the fixpoint set is particularly nice. Then we can more easily read off the properties of f.

Let us first consider the subgroup $\mathit{M\ddot{o}b}^+(\mathbb{H}) \subset \mathit{M\ddot{o}b}(\mathbb{H})$. We have seen that an element here can be written $f(z) = \dfrac{az+b}{cz+d}$, where $ad - bc = 1$. We assume from now on that f has this form and is not the identity.

Observe that as a map $\overline{\mathbb{C}} \to \overline{\mathbb{C}}$, f has ∞ as fixpoint if and only if $c = 0$. Then $f(z) = a^2z + ab$. If also $a = d$, or $a^2 = 1$, this is the only fixpoint — otherwise we have one more, namely $z = -b/(a-d) = ab/(1-a^2)$, which is a real number. In other words, if $c = 0$ we have either one or two fixpoints, and they lie in $\overline{\mathbb{R}}$.

If $c \neq 0$, the equation $z = \dfrac{az+b}{cz+d}$ is equivalent to the equation

$$cz^2 - (a-d)z - b = 0\,,$$

with roots

$$z = \frac{a - d \pm \sqrt{(a-d)^2 + 4bc}}{2c}\,.$$

Using that $ad - bc = 1$, we can simplify the square root and write

$$z = \frac{a - d \pm \sqrt{(a+d)^2 - 4}}{2c}\,.$$

We see that we should distinguish between three cases:

- $(a+d)^2 = 4$: Exactly one real root
- $(a+d)^2 > 4$: Two real roots
- $(a+d)^2 < 4$: Two complex roots

The number $a+d$ is the *trace* of the matrix $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$, and it is invariant under conjugation by elements of $GL_2(\mathbb{C})$. The trace of $-\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ is $-(a+d)$, so it follows that the number

$$\tau(f) = (a+d)^2$$

is invariant under conjugation of f by elements in $\textit{Möb}^+(\mathbb{H})$. In fact, it is also invariant under conjugation by $z \mapsto -\bar{z}$, hence invariant under conjugation by *every* element of $\textit{Möb}(\mathbb{H})$.

Note that if $c = 0$, then $ad = 1$ and $(a+d)^2 = (a-d)^2 + 4 \geq 4$, with equality if and only if $a = d$. Taking into account the discussion above of the case $c = 0$, we can distinguish between the following three cases, regardless of the value of c:

- Exactly one fixpoint in $\overline{\mathbb{R}}$, when $\tau(f) = 4$. We then say that f is of *parabolic type.*
- Two fixpoints in $\overline{\mathbb{R}}$, when $\tau(f) > 4$. f is of *hyperbolic type.*
- Two fixpoints in $\mathbb{C}$, when $\tau(f) < 4$. f is of *elliptic type.*

Let us consider more closely each of the three cases:

Case (1) : *One fixpoint in* $\overline{\mathbb{R}}$; f is of *parabolic* type.

If the fixpoint is at ∞, we must have $c = 0$, and $f(z)$ has the form $f(z) = z + \beta$, i. e. f is a translation parallel to the x–axis.

If the fixpoint is $q \in \mathbb{R}$, we can find an $h \in \textit{Möb}^+(\mathbb{H})$ mapping q to ∞. (Choose e. g. $h(z) = -1/(z-q)$.) Then the composition $g = h \circ f \circ h^{-1}$ is also an element of $\textit{Möb}^+(\mathbb{H})$, and g has ∞ as unique fixpoint. Hence, as above, g has the form

$$g(z) = z + \gamma$$

for a real number γ.

In fact, we can do even better than this: $\gamma \neq 0$, so we can conjugate g by $k(z) = \dfrac{z}{|\gamma|}$:

$$k \circ g \circ k^{-1}(z) = z \pm 1.$$

Hence any parabolic transformation is conjugate to a translation of the form $z+1$ or $z-1$. These two translations are conjugate in $M\ddot{o}b(\mathbb{H})$, but not in $M\ddot{o}b^+(\mathbb{H})$. (See Exercise 5.)

We think of such a conjugation as a change of coordinates: writing $f = h^{-1} \circ g \circ h$, we see that $f(z)$ is obtained by first moving z to $h(z)$, then applying g and finally moving back again by h^{-1}.

g fixes the point ∞ and translates horizontally all straight lines orthogonal to the real axis, i. e. $\mathbb{H}$–lines ending in ∞. Since h and h^{-1} both map $\mathbb{H}$–lines to $\mathbb{H}$–lines, we see that f must map $\mathbb{H}$–lines ending in q to $\mathbb{H}$–lines of the same type.

Figure 2.3.1 illustrates this in more detail. If g translates the vertical lines horizontally, it must also preserve the horizontal lines (dashed lines in the left figure). Mapped back by h^{-1} these become circles, but these circles are now *tangent* to the x–axis, as in the figure to the right. It follows that f also must preserve such circles.

Remark 2.3.1. Such (Euclidean) circles, tangent to $\overline{\mathbb{R}}$ at a point p, are called *horocircles* at p. They can also be characterized by the property that they are orthogonal to all hyperbolic lines ending at the point of tangency (See Exercise 3). Another characterization is discussed in Exercise 2.7.5.

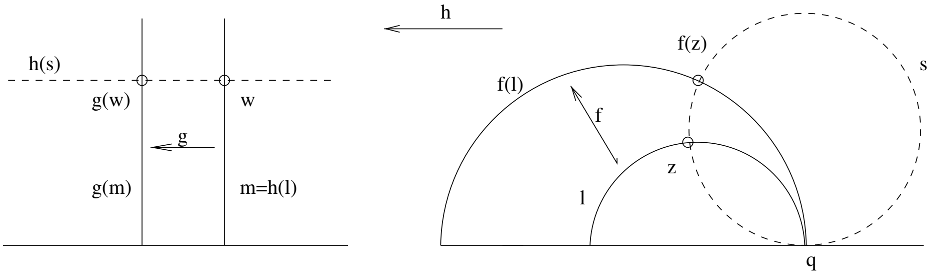


Fig. 2.3.1: Parabolic Möbius transformation

Case (2): f is of *hyperbolic* type, i. e. $\tau(f) = (a+d)^2 > 4$ and f has *two fixpoints in* $\overline{\mathbb{R}}$.

Let $h \in \textit{Möb}^+(\mathbb{H})$ be a real FLT mapping the two fixpoints to 0 and ∞. Then $g = h \circ f \circ h^{-1} \in \textit{Möb}^+(\mathbb{H})$ has 0 and ∞ as fixpoints. It also maps the imaginary axis to an $\mathbb{H}$–line l, But since the end points are fixed, the end points of l are again 0 and ∞, so l must also be the imaginary axis. In particular there is a positive real number η such that $g(i) = \eta i$, and since we know that g is uniquely determined by the values at the three points 0, i and ∞, we must have

$$g(z) = \eta z$$

for all z. Thus, up to a change of coordinates, f is just multiplication by a real number.

Write $\eta = \lambda^2$, such that $g(z) = \dfrac{\lambda z}{\lambda^{-1}}$, where we also may assume $\lambda > 0$. Since we must have $\tau(g) = \tau(f)$, λ satisfies the equation

$$\lambda + \frac{1}{\lambda} = |a + d|\,.$$

This equation has two roots λ and $1/\lambda$, and the corresponding functions $\lambda^2 z$ and z/λ^2 are conjugate by the transformation $z \mapsto -1/z$ — hence they are both realized by different choices of h. It follows that if we also choose $\eta > 1$, $g(z) = \eta z$ is uniquely determined. Moreover, this $g(z)$ is invariant under conjugation by $-\bar{z}$. Therefore there is a one–one correspondence between conjugacy classes of hyperbolic elements and real numbers > 1, in both $\textit{Möb}^+(\mathbb{H})$ and $\textit{Möb}(\mathbb{H})$.

Hyperbolic transformations behave as in figure 2.3.2. g preserves straight lines through 0, and these are mapped by h^{-1} to circular arcs or straight lines through the fixpoints of f, but the image of the imaginary axis is the only such curve meeting the x–axis orthogonally. Hence this is an $\mathbb{H}$–line between the two fixpoints of f, and it is mapped to itself by f. We call this $\mathbb{H}$–line the *axis* of f.

An element of $\textit{Möb}^+(\mathbb{H})$ of hyperbolic type with axis l is often called a "translation along l".

Case (3) : The final case is when we have *two complex fixpoints* — the *elliptic* case, when $\tau(f) < 4$.

Since the two fixpoints are the roots of a real, quadratic equation, they are complex conjugate. In particular, there is exactly one in the upper half–plane and none in $\overline{\mathbb{R}}$. Let p be the fixpoint in $\mathbb{H}$ and set $h(z) = \dfrac{z - \operatorname{Re} p}{\operatorname{Im} p}$, such that $h \in \textit{Möb}^+(\mathbb{H})$ and $h(p) = i$. This time $g = h \circ f \circ h^{-1}$ will be a real fractional linear transformation fixing i.

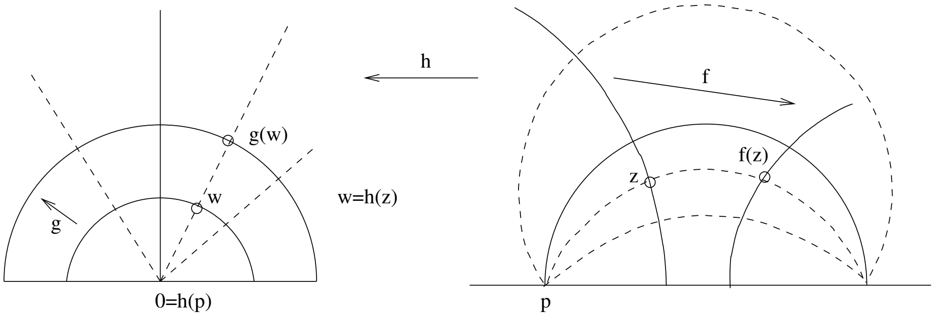


Fig. 2.3.2: Hyperbolic Möbius transformation

Let us analyze this g. Write g as $g(z) = \dfrac{\alpha z + \beta}{\gamma z + \delta}$, where α, β, γ, $\delta \in \mathbb{R}$. $g(i) = i$ means that $\alpha i + \beta = -\gamma + \delta i$ — i. e. $\alpha = \delta$ and $\beta = -\gamma$. If we substitute this into the equation $\alpha\delta - \beta\gamma = 1$, we see that $\alpha^2 + \beta^2 = 1$. Then we may write $\alpha = \cos(\theta)$, $\beta = \sin(\theta)$ for some real number θ, and we have

$$g(z) = g_\theta(z) = \frac{\cos(\theta)z + \sin(\theta)}{-\sin(\theta)z + \cos(\theta)} .$$

The matrix $\begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix}$ describes a (clock-wise) rotation by the angle θ in $\mathbb{R}^2$. We can think of g also as a kind of rotation, since it keeps i fixed and maps the $\mathbb{H}$–lines through i to lines of the same type. Figure 2.3.3 shows the image of the imaginary axis. (Note that $g_\theta(0) = \tan\theta$.)

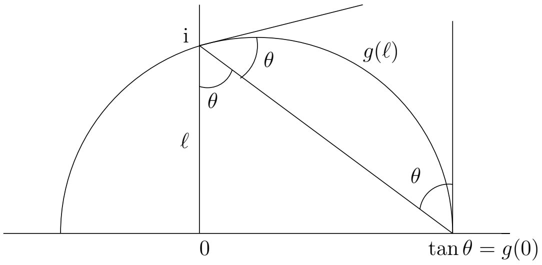


Fig. 2.3.3: Elliptic Möbius transformation

From the picture we see that the imaginary axis is rotated by an angle 2θ. But an easy calculation shows that $g_\theta g_{\theta'} = g_{\theta+\theta'}$ — hence g_θ will rotate *any* $\mathbb{H}$–line through i by an angle 2θ. In particular, g_π =id, and $g_{\theta+\pi} = g_\theta$.

Later, after we have introduced Poincaré's disk–model, the analogy with Euclidean rotations will become even clearer. See also Exercise 6.

If g_θ is conjugate to $g_{\theta'}$ in $Möb^+(\mathbb{H})$, then $g_\theta = g_{\theta'}$. The reason for this is that if $h^{-1}g_\theta h$ has i as fixpoint, then g_θ has $h(i)$ as fixpoint — hence $h(i) = i$. But then $h = g_\phi$ for some ϕ; thus g_θ and h commute. It follows that there is a one–one correspondence between conjugacy classes in $Möb^+(\mathbb{H})$ of elliptic elements and angles $\theta \in (0, \pi)$.

On the other hand, if $h(z) = -\bar{z}$, then $h^{-1}g_\theta h = g_{-\theta} = g_{\pi-\theta}$, so the conjugacy classes in $Möb(\mathbb{H})$ are in one–one correspondence with $\theta \in (0, \pi/2]$.

Let us sum up what we have done so far:

Proposition 2.3.2. *Suppose* $f(z) = \dfrac{az+b}{cz+d}$, *where* a, b, c, $d \in \mathbb{R}$ *and* $ad - bc = 1$, *and let* $\tau(f) = (a+d)^2$. *Assume that* f *is not the identity map. Then, as an element of* $Möb^+(\mathbb{H})$, f *is of*

- *Parabolic type, conjugate to* $z \mapsto z+1$ *or* $z-1$, *if* $\tau(f) = 4$,
- *Hyperbolic type, conjugate to exactly one* $z \mapsto \eta z$ *with* $\eta > 1$, *if* $\tau(f) > 4$,
- *Elliptic type, conjugate to a unique* $z \mapsto g_\theta(z) = \dfrac{\cos(\theta)z + \sin(\theta)}{-\sin(\theta)z + \cos(\theta)}$ *with* $\theta \in (0, \pi)$, *if* $\tau(f) < 4$.

In $Möb(\mathbb{H})$ the only differences are that there is only one conjugacy class of elements of parabolic type (see Exercise 5), and g_θ is conjugate to $g_{\pi-\theta}$.

We say that an element of $Möb^+(\mathbb{H})$ is given on *normal form* if it is written as a conjugate of one of these standard representatives.

We now move on to $Möb^-(\mathbb{H})$ and consider a transformation of the form $f(z) = \dfrac{a\bar{z}+b}{c\bar{z}+d}$, where a, b, c, $d \in \mathbb{R}$ and $ad - bc = -1$.

Again we will look for fixpoints of $f(z)$. As before, $z = \infty$ is a fixpoint if and only if $c = 0$. If $z \neq \infty$, the equation $f(z) = z$ is now equivalent to $c|z|^2 + dz - a\bar{z} - b = 0$, or

$$c(x^2+y^2) - (a-d)x - b = 0, \tag{2.3.1}$$

$$(a+d)y = 0. \tag{2.3.2}$$

We consider the two cases $a + d = 0$ and $a + d \neq 0$ separately.

First, let $a+d=0$. In this case equation (2.3.2) is trivially satisfied, so we have only one equation (2.3.1). This describes an $\mathbb{H}$–line: the vertical line $x = \dfrac{b}{d-a}$ when $c=0$ and the semi–circle with center $(a/c,0)$ and radius $1/|c|$ if $c \neq 0$. (Note that $a \neq d$ if $c=0$, since then $ad=-1$.) Hence f fixes an entire $\mathbb{H}$–line and interchanges the two components of its complement in $\mathbb{H}$. More precisely, by a suitable conjugation as above, we may assume that the fixed H–line is the imaginary axis. Then $c=b=0$ and $a=-d=\pm 1$ — hence $f(z)=-\overline{z}$. Thus all such transformations are conjugate to the horizontal reflection in the imaginary axis.

If the fixpoint set is another vertical line l, we can choose h to be a horizontal translation. Hence f must be horizontal reflection in l.

If $c \neq 0$ we can also write

$$f(z) = \frac{a}{c} + \frac{1/c}{c\bar{z}+d} = \frac{a}{c} + \frac{1/c^2}{\overline{z}-a/c}. \tag{2.3.3}$$

This has the general form

$$g(z) = m + \frac{r^2}{\overline{z-m}} = m + r^2 \frac{z-m}{|z-m|^2}\,.$$

If $\mathcal{C}$ is the circle (completion of an H-line) with center m and radius r, $g(z)$ maps points outside $\mathcal{C}$ to points inside and vice versa, and it leaves the circle itself fixed. More precisely, we see that $g(z)$ lies on the (Euclidean) ray from m through z, and such that the product $|g(z)-m||z-m|$ is equal to r^2. This is a very important geometric construction called "inversion in the circle $\mathcal{C}$".

By analogy we will also call the horizontal reflection in a vertical line l "inversion in l". Thus inversions in $\mathbb{H}$-lines are precisely the transformations in $M\ddot{o}b^-(\mathbb{H})$ such that $a+d=0$, and all inversions are conjugate. There are two particularly simple representatives for this conjugacy class: the horizontal reflection $z \mapsto -\overline{z}$ and the map $z \mapsto 1/\bar{z}$, which is inversion in the circle $|z|=1$. Either of these could be considered a "normal form" of such maps, but note that they all are inversions *a priori*, not only after a change of coordinates.

Next, assume $a+d \neq 0$. Then $y=0$ by (2.3.2), so there are no fixpoints in $\mathbb{H}$. In equation (2.3.1) we distinguish between the cases $c=0$ and $c \neq 0$.

If $c=0$, we get $x=b/(d-a)$, but then we also have the fixpoint (in $\overline{\mathbb{C}}$) $z=\infty$, so f must map the vertical line $x=b/(d-a)$ to itself. (But without fixpoints.) Note that we cannot have $a=d$, since $ad=-1$.

If $c \neq 0$, (2.3.1) has two solutions $x = \dfrac{a - d \pm \sqrt{(a+d)^2 + 4}}{2c}$ in $\mathbb{R}$. Hence $f(z)$ has two fixpoints on the real axis, and f must map the $\mathbb{H}$–line with these two points as endpoints to itself.

Thus, in both cases f preserves an $\mathbb{H}$–line ℓ, fixing the endpoints, and conjugating with a transformation mapping ℓ to the imaginary axis, we obtain a function of the form $k(z) = -\lambda^2\overline{z}$ — a composition of the reflection (inversion) in the y–axis and a hyperbolic transformation with the same axis.

Note that as $\lambda^2(-\overline{z}) = -\overline{(\lambda^2 z)}$, these two transformations *commute.* Conjugating back, we see that we have written f as a composition of two commuting transformations — a hyperbolic transformation h and an inversion g in the axis of h. Observe also that if $\lambda^2 \neq 1$, the imaginary axis is the only $\mathbb{H}$–line mapped to itself by $k(z) = -\lambda^2\overline{z}$. It follows that the line ℓ above is the only line such that $f(\ell) = \ell$. This is used in the proof of the following proposition:

Proposition 2.3.3. *Let $f \in$Möb$^-(\mathbb{H})$ have the form $f(z) = \dfrac{a\bar{z} + b}{c\bar{z} + d}$, where a, b, c, $d \in \mathbb{R}$ and $ad - bc = -1$.*

- *If $a + d = 0$, then f an inversion, conjugate to reflection in the imaginary axis.*
- *If $a + d \neq 0$, f can be written $f = gh$, where g is an inversion and g and h commute. Moreover, this decomposition is unique and h is of hyperbolic type and with axis equal to the line of inversion of g.*

Proof. It only remains to prove the uniqueness statement. So, suppose $f = gh = hg$, where g is inversion in a line ℓ. If $z \in \ell$ we have

$$g(h(z)) = h(g(z) = h(z)\,,$$

i. e. $h(z)$ is a fixpoint for g. Hence $h(z) \in \ell$. It follows that

$$f(\ell) = h(g(\ell)) = h(\ell) = \ell\,.$$

By as we just observed, this determines ℓ, hence also g. It is now clear that g and h are uniquely determined as the two transformations constructed above. □

It is worth pointing out that if we do not require that the components commute, there are many ways of decomposing an element of *Möb*$^-(\mathbb{H})$ into a product of an inversion and an element of $Möb^+(\mathbb{H})$. Trivial such decompositions are given by the formulas

$$\frac{a\bar{z}+b}{c\bar{z}+d} = \frac{(-a)(-\bar{z})+b}{(-c)(-\bar{z})+d} = -\overline{\left(\frac{(-a)z+(-b)}{cz+d}\right)}\,.$$

More interesting, perhaps; if $c \neq 0$, we can generalize (2.3.3) and write (using $ad-bc=-1$):

$$\frac{a\bar{z}+b}{c\bar{z}+d} = \frac{a}{c} + \frac{1/c}{c\bar{z}+d} = \frac{a+d}{c} + \left(-\frac{d}{c} + \frac{1}{c^2}\frac{z-(-d/c)}{|z-(-d/c)|^2}\right)\,. \qquad (2.3.4)$$

This is a composition of an inversion and a *parabolic* transformation. For more on decompositions of Möbius transformations, see exercises 9 and 10.

Note the following, which is implicit in what we have done:

- An element in $Möb^-(\mathbb{H})$ is an inversion if and only if its trace $a+d$ is 0. (This condition is independent of whether we have normalized the coefficients or not.)
- An element in $Möb^-(\mathbb{H})$ is an inversion if and only if it has a fixpoint in $\mathbb{H}$.
- An inversion in an $\mathbb{H}$–line l is characterized, as an element of $Möb(\mathbb{H})$, by having all of l as fixpoint set.

Exercises for 2.3

1. Classify the following maps and write them explicitly as conjugates of mappings on normal form.

$$\frac{4z-3}{2z-1}\,, \qquad -\frac{1}{z-1}\,, \qquad \frac{z}{z+1}\,.$$

2. Discuss the classification of Möbius transformations in terms of matrix representations, without assuming determinant 1.
3. Show geometrically that the horocircles at a point $p \in \overline{\mathbb{R}}$ are orthogonal to all $\mathbb{H}$-lines with p as one endpoint.

4. Explain what a hyperbolic transformation f does to the horocircles at the endpoints of the axis of f, and also to the other $\mathbb{H}$–lines sharing the same endpoint.

5. Show that all parabolic transformations are conjugate in $M\ddot{o}b(\mathbb{H})$. Show that the translations $z \mapsto z+1$ and $z \mapsto z-1$ are *not* conjugate in $M\ddot{o}b^+(\mathbb{H})$.

6. Fix a z in $\mathbb{H}$, $z \neq i$. Show that as θ varies, the points $g_\theta(z)$ all lie on the same circle in $\mathbb{C}$.

   (Hint: if $\cos\theta \neq 0$,write $g_\theta(z) = \dfrac{\tan\theta + z}{-\tan\theta\, z + 1}$, and think of this as a function of $\tan\theta$.)

7. Assume h_1 and h_2 are two nontrivial elements of $M\ddot{o}b^+(\mathbb{H})$ satsifying $h_1 h_2 = h_2 h_1$. Show that they have the same fixpoints, hence are also of the same type.

8. Show that an inversion in a circle $\mathcal{C} \subset \mathbb{C}$, considered as a map on $\mathbb{C}$ minus the center of $\mathcal{C}$, has the following properties:

   (a) It maps straight lines outside $\mathcal{C}$ to circles inside $\mathcal{C}$ and through its center.

   (b) Circles intersecting $\mathcal{C}$ orthogonally are mapped to themselves.

   (These are important results about inversions that are usually proved by geometric arguments. Her they should follow quite easily from what we now know about Möbius transformations.) .

9. Show that every element in $M\ddot{o}b^+(\mathbb{H})$ may be written as the composition of two inversions.

10. Show that $M\ddot{o}b(\mathbb{H})$ is generated by inversions, and show that $M\ddot{o}b^+(\mathbb{H})$ ($M\ddot{o}b^-(\mathbb{H})$) consists of those elements that can be written as a composition of an even (odd) number of inversions.

2.4 Hilbert's axioms and congruence in $\mathbb{H}$

We are now ready to prove that the upper half-plane provides a model for the hyperbolic plane satisfying the rest of Hilbert's axioms, with congruence based on the action of $\mathit{M\ddot{o}b}(\mathbb{H})$.

Recall that, using a combination of orthogonal and stereographic projections, we have identified the open unit disk $\mathbb{K} \subset \mathbb{R}^2$ with the upper half–plane $\mathbb{H} \subset \mathbb{C}$, such that chords in $\mathbb{K}$ correspond to what we have called $\mathbb{H}$-lines — vertical lines or semicircles with center on the real axis in $\mathbb{C}$. $\mathbb{K}$ inherits incidence and betweenness relations from $\mathbb{R}^2$, hence we obtain corresponding relations in $\mathbb{H}$. Automatically all of Hilbert's axioms I1–3 and B1–4 for these relations hold, as does Dedekind's axiom. In this section we introduce a congruence relation and show that it satisfies Hilbert's axioms C1–6. Since the parallel axiom has been replaced by the hyperbolic axiom, we will then have completed the construction of a hyperbolic geometry.

Remark 2.4.1. Betweenness for points on a line in the Euclidean plane can be formulated via homeomorphisms between the line and $\mathbb{R}$ or intervals in $\mathbb{R}$, hence the same is true for $\mathbb{H}$-lines, if we use the subspace topology from $\mathbb{C}$. On $\mathbb{R}$ the easiest definition is:

$$a * b * c \iff a < b < c \quad \text{or} \quad a > b > c\,.$$

(Equivalently: $(a-b)(b-c) > 0$.) The simplest such homeomorphisms are projections to the imaginary axis from the vertical lines and to the real axis from the half–circles. It follows that betweenness for points on an $\mathbb{H}$–line ℓ can be characterized by

- $x * y * z \iff \operatorname{Im} x * \operatorname{Im} y * \operatorname{Im} z$ if ℓ is vertical,
- $x * y * z \iff \operatorname{Re} x * \operatorname{Re} y * \operatorname{Re} z$ otherwise.

Before we go on, we need a more precise notation for lines, rays etc. If z_1, z_2 are two points of $\mathbb{H}$, we write $\overleftrightarrow{z_1 z_2}$ for the uniquely determined hyperbolic line containing them, $\overrightarrow{z_1 z_2}$ for the ray from z_1 containing z_2 and $[z_1, z_2]$ for the segment between z_1 and z_2 — i. e. $[z_1, z_2] = \overrightarrow{z_1 z_2} \cap \overrightarrow{z_2 z_1}$. An $\mathbb{H}$–line l is uniquely determined by its endpoints p and q in $\overline{\mathbb{R}}$, and therefore we will also write $l = (p, q)$. With this notation, the identity $\overleftrightarrow{z_1 z_2} = (p, q)$ will tell us that the uniquely determined $\mathbb{H}$–line containing z_1 and z_2 has endpoints p and q. Similarly, we may also write $[z_1, q) = \overrightarrow{z_1 q} = \overrightarrow{z_1 z_2}$, expressing that q is the endpoint of the ray $\overrightarrow{z_1 z_2}$.

We say that z_1 is the *vertex* and q the *endpoint* of the ray $[z, q)$. An *angle* is then an unordered pair of rays with the same vertex, where the two rays do not lie on the same line. We use the notation $\angle uzv$ for the unordered pair $\{\overrightarrow{zu}, \overrightarrow{zv}\}$, where $z \in \mathbb{H}$ and u, v are either in $\mathbb{H}$ or in $\overline{\mathbb{R}}$.

Recall that we have defined the *two sides* of a line l by saying that two points z_1 and z_2 are on the same side of l if $[z_1, z_2] \cap l = \emptyset$. (Page 4 and exercise 1.1.4.) If r is the inversion in l, clearly $[z, r(z)] \cap l \neq \emptyset$. Hence r interchanges the two sides of l.

The congruence relation in $\mathbb{H}$ is now defined as follows:

Congruence of segments: $[z_1, z_2] \cong [w_1, w_2] \iff g([z_1, z_2]) = [w_1, w_2]$ for some $g \in \mathit{M\ddot{o}b}(\mathbb{H})$.

Congruence of angles: $\angle uzv \cong \angle u'z'v' \iff g(\overrightarrow{zu}) = \overrightarrow{z'u'}$ and $g(\overrightarrow{zv}) = \overrightarrow{z'v'}$ for some $g \in \mathit{M\ddot{o}b}(\mathbb{H})$. (Notation: $g(\angle uzv) = \angle u'z'v'$.)

The existence parts of the congruence statements say that there are enough Möbius transformations to move angles and segments freely around in $\mathbb{H}$, whereas the uniqueness means that there are not too many such transformations. The technical results we need are contained in the following Lemmas:

Lemma 2.4.2. *Suppose z_j lies on an $\mathbb{H}$–line l_j, with endpoints p_j and q_j, for $j = 1, 2$. Then there is a uniquely determined $f \in \mathit{M\ddot{o}b}^+(\mathbb{H})$ such that $f(p_1) = p_2$, $f(q_1) = q_2$ and $f(z_1) = z_2$ — hence also $f(l_1) = l_2$.*

In particular we have, for example, $f([z_1, q_1)) = [z_2, q_2)$. But since a ray determines the line containing it, we get

Corollary 2.4.3. *$\mathit{M\ddot{o}b}^+(\mathbb{H})$ acts transitively on the set of all rays: In fact, given two rays σ_1 and σ_2 with vertices z_1 and z_2, there is a unique $f \in \mathit{M\ddot{o}b}^+(\mathbb{H})$ such that $f(z_1) = z_2$ and $f(\sigma_1) = \sigma_2$.*

Lemma 2.4.4. *(i) An element in $\mathit{M\ddot{o}b}^+(\mathbb{H})$ is completely determined by its values at two points in $\mathbb{H}$.*

(ii) Suppose the segments $[z_1, z_2]$ and $[w_1, w_2]$ are congruent. Then there is a uniquely determined $f \in \mathit{M\ddot{o}b}^+(\mathbb{H})$ such that $f(z_1) = w_1$ and $f(z_2) = w_2$.

Lemma 2.4.5. *Given two rays σ_1 and σ_2 with a common vertex z_0. Then there is a unique inversion g such that $g(z_0) = z_0$, $g(\sigma_1) = \sigma_2$ and $g(\sigma_2) = \sigma_1$.*

Proof of Lemma 2.4.2. This is Exercise 2.2.7b, but, for completeness, here is a proof:

By Corollary 2.2.5 there exists a unique $f \in M\ddot{o}b^+(\mathbb{C})$ with the right properties, and all we have to prove is that it lies in $M\ddot{o}b^+(\mathbb{H})$. Let $\mathcal{C}_i, i = 1, 2$ be the $\overline{\mathbb{C}}$-circle containing l_i (and determined by p_i, q_i and z_i). Then we must have $f(\mathcal{C}_1) = \mathcal{C}_2$, and $f(\overline{\mathbb{R}})$ is a $\overline{\mathbb{C}}$-circle meeting $\mathcal{C}_2$ in p_2 and q_2 at right angles. Hence $f(\overline{\mathbb{R}}) = \overline{\mathbb{R}}$, and $f \in M\ddot{o}b(\mathbb{H})$. (Again by Corollary 2.2.5.) But since $f(z_1) = z_2$, we must have $f \in M\ddot{o}b^+(\mathbb{H})$. □

Proof of Lemma 2.4.4. (i) Two points in $\mathbb{H}$ determine a unique line l, and the endpoints of l must map to the endpoints of $f(l)$, in such a way that betweenness relations are preserved. Hence the values of f at *four* points are determined, and the uniqueness follows from uniqueness in Corollary 2.2.5.

(ii) Assume that $g([z_1, z_2]) = [w_1, w_2]$ for some $g \in M\ddot{o}b(\mathbb{H})$. If $g \in M\ddot{o}b^-(\mathbb{H})$, we replace g by $k \circ g$, where k is the inversion in the $\mathbb{H}$–line $\overleftrightarrow{w_1 w_2}$. Therefore we may assume that $g \in M\ddot{o}b^+(\mathbb{H})$.

The problem is that we might have $g(z_1) = w_2$ and $g(z_2) = w_1$. If so, choose an $h \in M\ddot{o}b^+(\mathbb{H})$ such that $h(\overleftrightarrow{w_1 w_2})$ is the imaginary axis, and write $h(w_1) = \omega_1 i$, $h(w_2) = \omega_2 i$. If we define $k(z) = -\omega_1\omega_2/z$, we see that k interchanges $\omega_1 i$ and $\omega_2 i$. Then $h^{-1}kh$ will interchange w_1 og w_2, and we let $f = h^{-1}khg$. □

Proof of Lemma 2.4.5. It follows easily from Lemma 2.4.2 that we can find an $h \in M\ddot{o}b^+(\mathbb{H})$ mapping σ_1 to σ_2. (h has z_0 as fixpoint and must necessarily be elliptic.) Let r be inversion in the line containing σ_1 and define $g = hr$. Then $g \in M\ddot{o}b^-(\mathbb{H})$ and has a fixpoint $z_0 \in \mathbb{H}$, hence is also an inversion (Remark at the end of 2.2.) Clearly $g(\sigma_1) = h(\sigma_1) = \sigma_2$, and then $g(\sigma_2) = g^2(\sigma_1) = \sigma_1$.

Uniqueness: Suppose g' is another inversion with the same properties. Then $g^{-1}g'$ is an element of $M\ddot{o}b^+(\mathbb{H})$ mapping both σ_1 and σ_2 to themselves. Therefore it has three fixpoints z_0, q_1 and q_2 (in $\overline{\mathbb{C}}$), hence it must be the identity map. Thus $g = g'$. □

We are now ready to prove that Hilbert's congruence axioms C1–6 are satisfied. The axioms are:

The axioms for congruence of segments:

C1: Given a segment $[z_1, z_2]$ and a ray σ with vertex w_1, there is a uniquely determined point w_2 on σ such that $[w_1, w_2] \cong [z_1, z_2]$.

C2: $\cong$ is an equivalence relation on the set of segments.

C3: If $z_1 * z_2 * z_3$ and $w_1 * w_2 * w_3$ and both $[z_1, z_2] \cong [w_1, w_2]$ and $[z_2, z_3] \cong [w_2, w_3]$, then also $[z_1, z_3] \cong [w_1, w_3]$.

The axioms for congruence of angles:

C4: Given a ray $[w, q)$ and an angle $\angle uzv$, there are unique angles $\angle p_1wq$ and $\angle p_2wq$ on opposite sides of $[w, q)$ such that $\angle p_1wq \cong \angle p_2wq \cong \angle uzv$.

C5: $\cong$ is an equivalence relation on the set of angles.

C6: (SAS) Given triangles $z_1z_2z_3$ og $w_1w_2w_3$. If $[z_1, z_2] \cong [w_1, w_2]$, $[z_1, z_3] \cong [w_1, w_3]$ and $\angle z_2z_1z_3 \cong \angle w_2w_1w_3$, then the two triangles are congruent — i. e. we also have $[z_2, z_3] \cong [w_2, w_3]$, $\angle z_1z_2z_3 \cong \angle w_1w_2w_3$ and $\angle z_2z_3z_1 \cong \angle w_2w_3w_1$.

C2 and **C5** follow immediately since we have defined congruence by a group action.

C1. The segment $[z_1, z_2]$ defines a ray $\overrightarrow{z_1z_2}$, and by Corollary 2.4.3 there exists an $f \in M\ddot{o}b^+(\mathbb{H})$ such that $f(z_1) = w_1$ and $f(\overrightarrow{z_1z_2}) = \sigma$. If we put $w_2 = f(z_2)$, we clearly get $[w_1, w_2] \cong [z_1, z_2]$.

Suppose that w_2' is another point on σ such that $[w_1, w_2'] \cong [z_1, z_2]$. By Lemma 2.4.4 we can find $h \in M\ddot{o}b^+(\mathbb{H})$ such that $h(z_1) = w_1$ and $h(z_2) = w_2'$. But then also $h(\overrightarrow{z_1z_2}) = \overrightarrow{w_1w_2'} = \sigma$, and by the uniqueness in Corollary 2.4.3 we must have $h = f$. Therefore $w_2 = f(z_2) = h(z_2) = w_2'$.

C3. Here we use Lemma 2.4.4(ii), saying that there exists a $g \in M\ddot{o}b^+(\mathbb{H})$ such that $g(z_1) = w_1$ and $g(z_2) = w_2$. Then $w_3' = g(z_3)$ is on the ray $\overrightarrow{w_2w_3}$, and g defines congruences $[z_2, z_3] \cong [w_2, w_3']$ and $[z_1, z_3] \cong [w_1, w_3']$. But by the uniqueness in C1 we must then have $w_3' = w_3$.

C4. We may assume that u and v are the endpoints of the rays in $\angle uzv$. To construct the angles is easy: let $g \in M\ddot{o}b^+(\mathbb{H})$ be such that $g([z, u)) = [w, q)$, and define $\angle p_1wq$ as $g(\angle uzv)$. Let $j \in M\ddot{o}b^-(\mathbb{H})$ be inversion in the $\mathbb{H}$–line l containing $[w, q)$, and let $\angle p_2wq = jg(\angle uzv) = j(\angle p_1wq)$. Since j interchanges the two sides of l, then $\angle p_1wq$ and $\angle p_2wq$ must lie on opposite sides of $[w, q)$.

It remains to prove uniqueness. Suppose that $h(\angle uzv) = \angle pwq$ and $h'(\angle uzv) = \angle p'wq$, with p and p' on the same side of $[w, q)$, and $h, h' \in$

Möb$(\mathbb{H})$. Then $h'h^{-1}(\angle pwq) = \angle p'wq$, and, using Lemma 2.4.5, we may assume $h'h^{-1}(w) = w$ and $h'h^{-1}(q) = q$. Hence $h'h^{-1}$ is either the identity or a reflection in the $\mathbb{H}$-line containing the ray $[w, q)$. But such a reflection must interchange the two sides of this $\mathbb{H}$-line, so we must have $h'h^{-1} = \mathrm{id}$, and therefore also $\angle pwq = \angle p'wq$.

C6. By assumption, we have $\angle w_2w_1w_3 = g(\angle z_2z_1z_3)$, for some $g \in$ *Möb*$(\mathbb{H})$, and after applying Lemma 2.4.5, if necessary, we may assume that $g(\overrightarrow{z_1z_2}) = \overrightarrow{w_1w_2}$ and $g(\overrightarrow{z_1z_3}) = \overrightarrow{w_1w_3}$. But uniqueness in C1 then means that $g(z_2) = w_2$ and $g(z_3) = w_3$. Hence it follows that also $g([z_2, z_3]) = [w_2, w_3]$, $g(\angle z_1z_2z_3) = \angle w_1w_2w_3$ and $g(\angle z_2z_3z_1) = \angle w_2w_3w_1$.

□

Remark 2.4.6. Except for **C6**, we could have defined congruence using the smaller group *Möb*$^+(\mathbb{H})$. The remaining axioms would still hold.

2.5 Distance in $\mathbb{H}$

Now that we have established the existence of a model for hyperbolic geometry based on the upper half–plane $\mathbb{H}$, it is time to start investigating the geometric structure itself. Classical geometry has a rich theory of triangles and circles, we can measure angles and the lengths of sides, and there is a theory of trigonometric functions theory relating them. Of great practical importance are also formulas for arc lengths and area of more general figures. To what extent can we do the same in hyperbolic geometry? Many classical geometric arguments do not use the parallel axiom, and these will automatically be valid in hyperbolic geometry, as well. So, which results carry over and which do not? And, what can we say when they do not?

Naturally, we can only scratch the surface here, but in the next sections we shall develop some of the basic theory. Enough, hopefully, to give a feeling for what the hyperbolic world looks like.

We start with the fundamental concept of *distance*, and we will show that there is a distance function in $\mathbb{H}$ which characterizes congruence, just as in Euclidean geometry. Thus, we want to define a *metric* on $\mathbb{H}$ — i. e. a function $d : \mathbb{H} \times \mathbb{H} \to \mathbb{R}$ such that

(d1) $d(z, w) \geq 0$, *and* $d(z, w) = 0$ *if and only if* $z = w$,

(d2) $d(z, w) = d(w, z)$ *for all* $z, w \in \mathbb{H}$,

(d3) $d(z, w) \leq d(z, u) + d(u, w)$ *for all* $z, u, w \in \mathbb{H}$.

The desired relation with geometry leads us to require that it should have the following additional properties:

(d4) *If* z, u, w *are distinct points in* $\mathbb{H}$, *then* $d(z, w) = d(z, u) + d(u, w)$ *if and only if* $u \in [z, w]$. *("Distance is measured along* $\mathbb{H}$*–lines".)*

(d5) $d(z, w) = d(z', w')$ *if and only if there exists a* $h \in \mathit{Möb}(\mathbb{H})$ *such that* $h(z) = z'$ *and* $h(w) = w'$.

("Two segments are congruent if and only if they have the same lengths".)

We first observe that (d4) and (d5) determine the metric almost completely, if it exists. Given z, w, there is a unique $g \in \mathit{Möb}^+(\mathbb{H})$ such that $g(z) = i$ and $g(w) \in [i, \infty)$, i. e. $g(w) = ti$, where $t \geq 1$. (This is Axiom C1 and Lemma 2.4.4.) Then, because of (d5), we must have $d(z, w) = d(i, ti)$, hence d is completely determined by the function $f : [1, \infty) \to [0, \infty)$ defined by $f(t) = d(i, ti)$.

We can say more about the function f. Let s and t be two numbers greater than or equal to 1. Then $si \in [i, sti]$ — hence from (d4) (and (d1), if s or t is 1) we have:

$$f(st) = d(i, sti) = d(i, si) + d(si, sti) = d(i, si) + d(i, ti) = f(s) + f(t)\,.$$

(The third inequality follows from (d5) applied to $h(z) = sz$ and $z = i$.) But then one can show that $f(t) = C\ln(t)$ for some real (positive) constant C.

(See Exercise 2.5.1 for the case when f is differentiable in one point. But it is easy to see that f must be *increasing,* and a famous Theorem due to Lebesgue says that an increasing function is differentiable almost everywhere.)

The choice of constant C just amounts to a scaling ('choice of unit'), and any positive number could be used in the following. We choose to set $C = 1$. (More on this in chapter 5, especially 5.8.) We now set

$$d(z, w) = \ln(|g(w)|)\,,$$

where $g \in \mathit{Möb}^+(\mathbb{H})$ maps z to i and w to a point ti, where $t \geqslant 1$. This is well defined, and we now want to show that this function satisfies (d1-d5) and hence defines the metric we want.

(d1) is obvious. To prove (d2), set $h(u) = \dfrac{-t}{g(u)}$ with the notation above. Then $h \in \textit{Möb}^+(\mathbb{H})$, $h(z) = -t/i = ti$, and $h(w) = -t/ti = i$. Hence $d(w,z) = \ln(t) = d(z,w)$, by definition.

Consider next (d5). If $d(z,w) = d(z',w') = \ln(t)$, it follows from the definition that there are $g, g' \in \textit{Möb}^+(\mathbb{H})$ sych that $g(z) = g'(z') = i$ and $g(w) = g'(w') = ti$. Then, setting $h = g'^{-1}g$, we have $h(z) = z'$ and $h(w) = w'$.

Conversely, if $h(z) = z'$ and $h(w) = w'$ for some $h \in \textit{Möb}(\mathbb{H})$, we can assume that $h \in \textit{Möb}^+(\mathbb{H})$, if necessary by composing with inversion in the $\mathbb{H}$-line through z and w. If g is the unique element in $\textit{Möb}^+(\mathbb{H})$ such that $g(w) = i$ and $g(w) \in [i, \infty)$, then $g' = gh^{-1}$ is the unique element in $\textit{Möb}^+(\mathbb{H})$ such that $g'(w') = i$ and $g'(w') \in [i, \infty)$.

Before we go on, we remark that if $g(w) = ti$, then $\dfrac{-1}{g(w)} = \dfrac{i}{t}$. Observing that $|\ln(t)| = |\ln(\frac{1}{t})|$, we then see that the condition $t \geqslant 1$ can be omitted from the definition of $d(z,w)$, provided we set

$$d(z,w) = |\ln|g(z)||.$$

We now prove the triangle inequality (d3), with the additional refinement (d4). For this, we need a more explicit expression for the transformation g. If p and q are the endpoints of the unique hyperbolic line through z and w, we can set

$$g(w) = [w, z, p, q]\, i.$$

Then

$$d(z,w) = |\ln|[w, z, p, q]||\,, \tag{2.5.1}$$

and we have the following lemma:

Lemma 2.5.1. *For every z, w in $\mathbb{H}$ we have $d(z,w) \geq |\ln(\operatorname{Im} w/\operatorname{Im} z)|$, with equality if and only if $\operatorname{Re} z = \operatorname{Re} w$.*

Proof. If $\operatorname{Re} z = \operatorname{Re} w$ we can choose $p = \operatorname{Re} z$ and $q = \infty$, and we have $[w, z, p, q] = (w-p)/(z-p) = \operatorname{Im} w/\operatorname{Im} z$.

Henceforth, assume $\operatorname{Re} z \neq \operatorname{Re} w$. The inequality is trivially satisfied if $\operatorname{Im} z = \operatorname{Im} w$. Hence we also assume $\operatorname{Im} z \neq \operatorname{Im} w$, and since $d(z,w) = d(w,z)$, we may choose the labeling such that $\operatorname{Im} w > \operatorname{Im} z$. Then

$$d(z,w) = d(w,z) = |\ln(|\frac{w-q}{w-p}\,\frac{z-p}{z-q}|)| = |\ln(\frac{|w-q|}{|w-p|}/\frac{|z-q|}{|z-p|})|\,,$$

where p and q are the endpoints of $\overleftrightarrow{zw}$. This expression does not change if we interchange p and q, hence we may assume that $p * w * z * q$. We now have the situation illustrated in figure 2.5.1 (or its mirror image).

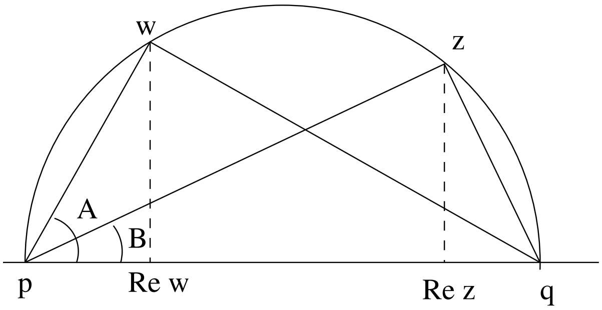


Fig. 2.5.1:

From the figure we see that $\dfrac{|w-q|}{|w-p|} = \tan A$ and $\dfrac{|z-q|}{|z-p|} = \tan B$. But we also have $\tan A = \operatorname{Im} w / |\operatorname{Re} w - p|$ and $\tan B = \operatorname{Im} z / |\operatorname{Re} z - p|$, hence we get

$$\frac{|w-q|}{|w-p|} \Big/ \frac{|z-q|}{|z-p|} = \frac{\tan A}{\tan B} = \frac{\operatorname{Im} w}{\operatorname{Im} z} \cdot \frac{|\operatorname{Re} z - p|}{|\operatorname{Re} w - p|} .$$

Because of the way we have chosen p, w, z and q, we have the inequalities $\dfrac{\operatorname{Im} w}{\operatorname{Im} z} > 1$ and $\dfrac{|\operatorname{Re} z - p|}{|\operatorname{Re} w - p|} > 1$. Therefore $\dfrac{|w-q|}{|w-p|} \Big/ \dfrac{|z-q|}{|z-p|} > \dfrac{\operatorname{Im} w}{\operatorname{Im} z} > 1$, and

$$d(z,w) = \ln\Big(\frac{|w-q|}{|w-p|} \Big/ \frac{|z-q|}{|z-p|}\Big) > \ln\Big(\frac{\operatorname{Im} w}{\operatorname{Im} z}\Big) .$$

Hence we have strict inequality when $\operatorname{Re} z \neq \operatorname{Re} w$ and equality when $\operatorname{Re} z = \operatorname{Re} w$. □

To prove the triangle inequality, it is enough to consider the case $z = i$, $w = ti$, where $t > 1$, since we can always move to this situation by (d5). If u is a third point, we get

$$\begin{aligned} d(z,u) + d(u,w) &\geq |\ln(\operatorname{Im} u/\operatorname{Im} z)| + |\ln(\operatorname{Im} w/\operatorname{Im} u)| \geq \\ & |\ln(\operatorname{Im} u/\operatorname{Im} z) + \ln(\operatorname{Im} w/\operatorname{Im} u)| = |\ln(\operatorname{Im} w/\operatorname{Im} z)| = d(z,w) . \end{aligned}$$

This is (d3). We have equality if and only if $\operatorname{Re} u = 0$ and $\ln(\operatorname{Im} u/\operatorname{Im} z)$ and $\ln(\operatorname{Im} w/\operatorname{Im} u)$ have the same sign — i. e. if and only if u also lies on the imaginary axis and $1 = \operatorname{Im} z \leq \operatorname{Im} u \leq \operatorname{Im} w$. But this means precisely that $u \in [z, w]$, and (d4) follows. □

Remark 2.5.2. We have now constructed a 'good' distance function on $\mathbb{H}$, but the formula (2.5.1) is not as explicit as we would have liked. A better formula will be given later: see formula (2.7.7).

Exercises for 2.5

1. Show that a function $f : (1, \infty) \to \mathbb{R}$ which is differentiable in one point and satisfies the equation $f(st) = f(s) + f(t)$ for all s and t must be equal to $C \ln(t)$ for some constant C. (Hint: Show that F is differentiable everywhere and compute its derivative.)

2. Find the midpoint of $[a + ib, a + ic]$, where $a, b, c \in \mathbb{R}$, $0 < b < c$.

3. The distance between two subsets U and V of a metric space is defined as $d(U, V) = \inf\{d(u, v) \,|\, u \in U,\ v \in V\}$. Show that this distance is preserved by Möbius transformations.

   Let l_1 and l_2 be two hyperbolic lines with a common endpoint. Show that $d(l_1, l_2) = 0$.

4. Compute $d(z, -\bar{z})$ for $z \in \mathbb{H}$. Show that the distance from z to the imaginary axis is $\dfrac{d(z, -\bar{z})}{2}$.

   What does the set of points in $\mathbb{H}$ having the same, fixed distance to the imaginary axis look like?

5. Assume that the function $f : \mathbb{H} \to \mathbb{H}$ preserves hyperbolic distance (i. e. $d(f(x), f(y)) = d(x, y)$ for all $x, y \in \mathbb{H}$). Prove that $f \in \mathit{M\ddot{o}b}(\mathbb{H})$.

2.6 Angle measure. $\mathbb{H}$ as a conformal model

Whereas the definition of distance required a fair amount of work, it turns out that angle measure is much simpler, due to the fact that all the congruence transformations are conformal as maps of $\mathbb{C}$ (Lemma 2.2.1ii).

A point z on a line divides the line into two rays with z as common vertex. A third ray from z determines two angles with these rays, and these angles are called *supplements* of each other. A given angle then has two supplementary angles, but they are easily seen to be congruent.

Definition. An angle is said to be a *right angle* if it is congruent to its supplements.

The usual (Euclidean) angular measure in $\mathbb{R}^2$ is a function which to any angle associates a number between 0 and $2R$, where R is the number we associate to a right angle (usually $\pi/2$ or 90 degrees), such that two angles are congruent (in the Euclidean sense) if and only if they are associated to the same number. The function is also *additive*, in the following precise sense:

Suppose A and D are on opposite sides of $\overleftrightarrow{BC}$, and suppose the angles $\angle ABC$ and $\angle CBD$ have angle measures U and V, respectively, If $U + V < 2R$, then the angle $\angle ABD$ has measure $U+V$. (The conditions mean that we are talking about a *sum*, rather than a *difference* of angles, and that we are only considering angles smaller than two right angles.) If we normalize the angle measure by requiring that right angles have the measure R and every number in the interval $(0, 2R)$ is realized by some angle, this determines the angle measure uniquely. Note that what determines the angle measure is then essentially the concept of *congruence.*

In $\mathbb{H}$ an angle measure should have exactly the same properties, except that two angles now should have the same measure if and only if they are congruent in the *hyperbolic* sense — i. e. there exists a Möbius transformation mapping one to the other. Now we define the *Euclidean angle* between two rays with common vertex to be the angle between their tangents at the vertex. Then Lemma 2.2.1 says that fractional linear transformations preserve the Euclidean angle measure, and complex conjugation trivially does the same. Therefore all Möbius transformations, and in particular those in $\textit{Möb}(\mathbb{H})$, also preserve Euclidean angle measure. In fact, the converse is also true, so we have:

Lemma 2.6.1. *Two angles A and B are congruent if and only if they have the same Euclidean angle measure.*

Proof. We only need to prove the *if* part. Let $A = \angle xyz$ and $B = \angle uvw$, and suppose they both have Euclidean angle measure θ. By Hilbert's axiom C4 we may reduce to the case $\overrightarrow{yx} = \overrightarrow{vu} = [i, 0)$. (Notation from section 4.) But there are exactly two rays from i making an angle θ with $[i, 0)$ — one on each side of the imaginary axis — and they are mapped to each other by the reflection $z \mapsto -\bar{z}$, which fixes $[i, 0)$. □

It follows that we can use the same measure of angles in $\mathbb{H}$ as in the Euclidean plane containing it.

We express this by saying that the Poincaré upper half–plane $\mathbb{H}$ is a *conformal* model for hyperbolic geometry. This is one of the properties that makes this model much more useful that the Beltrami–Klein model $\mathbb{K}$, which is not conformal.

Note that Lemma 2.6.1 then also says that two angles are congruent if and only if they have the same size.

Since stereographic projection preserves angles, it follows that the hemisphere model $\mathbb{B}$ also is conformal, and hence so is every other model obtained from it by stereographic projections. In particular, this is true for Poincaré's disk model $\mathbb{D}$, which we will investigate in the next section.

Exercises for 2.6

1. Show that if l is a hyperbolic line and z is point not on l, then there is a unique line l' which contains z and which meets l orthogonally.

2. Show that if l_1 and l_2 are two lines which do not have a common endpoint and which do not intersect, then there exists a line m which meets both orthogonally.

2.7 Poincaré's disk model $\mathbb{D}$

Because of its rotational symmetry, the *Poincaré disk model* $\mathbb{D}$ will in certain respects have great advantages over $\mathbb{H}$. Therefore this section is devoted to a closer study of this model. Having two different models to our disposal enables us in each situation to choose the one best suited. We will see examples of this in the last two sections.

We begin by transferring everything we have done with the upper half–plane model $\mathbb{H}$ to $\mathbb{D}$. We will do this using the bijection $G = \Phi \circ F \circ \Phi^{-1}$, where Φ is stereographic projection and $F : B_1 \approx B_2$ is an identification of the hemispheres $B_1 = \{(x,y,z)|y > 0\}$ and $B_2 = \{(x,y,z)|z < 0\}$. (Cf. Exercise 2.1.3.) If we choose $F(x,y,z) = (x,z,-y)$, the formulas (2.1.1) and (2.1.2) give:

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

or, if we write G in terms of complex numbers $z = u + iv$:

$$G(z) = \frac{z+\bar{z}+i(z\bar{z}-1)}{|z+i|^2} = \frac{(iz+1)(\bar{z}-i)}{(z+i)(\bar{z}-i)} = \frac{iz+1}{z+i}.$$

This is a fractional linear transformation which restricts to a bijection $G : \mathbb{H} \approx \mathbb{D}$. We see that $G(0) = -i$, $G(1) = 1$ and $G(-1) = -1$, and this determines G uniquely, by Corollary 2.2.5. (In fact, we could have used this to define G.) Observe also that $G(\infty) = i$ and $G(i) = 0$.

There are, of course, many other possible FLT's identifying $\mathbb{H}$ and $\mathbb{D}$, but G is particularly simple and will be our preferred choice.

G preserves angles and $\overline{\mathbb{C}}$–circles, hence it maps the circle through 0, 1 and -1, i. e. $\overline{\mathbb{R}}$, to the circle through $-i$, 1 and -1, i. e. $S^1 = \{z \in \mathbb{C} \,|\, |z| = 1\}$. Therefore the $\mathbb{H}$–lines are mapped to either circular arcs meeting S^1 orthogonally, or diameters. These curves are the hyperbolic lines in the disk model, or $\mathbb{D}$*–lines.*

Since G preserves angles, this model will also be *conformal* — i. e. the hyperbolic angle measure is the same as the Euclidean angle measure.

The group of real Möbius transformations $\textit{Möb}(\mathbb{H})$ corresponds to a group $\textit{Möb}(\mathbb{D})$ of transformations of $\mathbb{D}$ which preserve angles and hyperbolic lines, and which generally has the same properties with respect to $\mathbb{D}$ as $\textit{Möb}(\mathbb{H})$ to $\mathbb{H}$. $\textit{Möb}(\mathbb{D})$ is defined by

$$f \in \textit{Möb}(\mathbb{D}) \iff G^{-1}fG \in \textit{Möb}(\mathbb{H}),$$

In other words: every element $f \in \textit{Möb}(\mathbb{D})$ can be written as GgG^{-1} for some $g \in \textit{Möb}(\mathbb{H})$, and every transformation of this form is in $\textit{Möb}(\mathbb{D})$. It follows that $\textit{Möb}(\mathbb{D})$ and $\textit{Möb}(\mathbb{H})$ are conjugate subgroups of $\textit{Möb}(\mathbb{C})$, hence they are isomorphic as abstract groups. We will also use the notation $\textit{Möb}^+(\mathbb{D})$ and $\textit{Möb}^-(\mathbb{D})$ for the subgroup and coset corresponding to $\textit{Möb}^+(\mathbb{H})$ and $\textit{Möb}^-(\mathbb{H})$.

Congruence in $\mathbb{D}$ can now be defined as in $\mathbb{H}$, but using the group $\textit{Möb}(\mathbb{D})$ instead of $\textit{Möb}(\mathbb{H})$. The mappings G and G^{-1} will then preserve congruence.

To see what the transformations in $\textit{Möb}(\mathbb{D})$ look like, we use the matrix representation of Möbius transformations. G corresponds to the matrix

$\begin{bmatrix} i & 1 \\ 1 & i \end{bmatrix}$, and for G^{-1} we will use the formula $G^{-1}(z) = \dfrac{iz-1}{-z+i}$, corresponding to the matrix $\begin{bmatrix} i & -1 \\ -1 & i \end{bmatrix}$.

Consider first *Möb*$^+(\mathbb{D})$. If $g(z) = \dfrac{az+b}{cz+d}$, a, b, c, d real, GgG^{-1} will correspond to

$$\begin{aligned} \begin{bmatrix} i & 1 \\ 1 & i \end{bmatrix} \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} i & -1 \\ -1 & i \end{bmatrix} &= \begin{bmatrix} ia+c & ib+d \\ a+ic & b+id \end{bmatrix} \begin{bmatrix} i & -1 \\ -1 & i \end{bmatrix} \\ &= \begin{bmatrix} -a+ci-ib-d & -ia-c-b+id \\ ia-c-b-id & -a-ic+ib-d \end{bmatrix} \\ &= \begin{bmatrix} -(a+d)+(c-b)i & -(b+c)-(a-d)i \\ -(b+c)+(a-d)i & -(a+d)-(c-b)i \end{bmatrix}. \end{aligned}$$

The last matrix has the form $\begin{bmatrix} \alpha & \beta \\ \bar\beta & \bar\alpha \end{bmatrix}$, with

$$\alpha = -(a+d)+(c-b)i \textit{ and } \beta = -(b+c)-(a-d)i \tag{2.7.1}$$

This means that elements in *Möb*$^+(\mathbb{H})$ give rise to complex fractional linear transformations of the form

$$g(z) = \frac{\alpha z+\beta}{\bar\beta z+\bar\alpha}. \tag{2.7.2}$$

On the other hand, it is easy to see that every pair of complex numbers (α, β) can be written uniquely on the form (2.7.1), with a, b, c and d real. A simple calculation then gives $\alpha\bar\alpha - \beta\bar\beta = 4(ad-bc)$ (compare determinants), hence $\dfrac{\alpha z+\beta}{\bar\beta z+\bar\alpha}$ defines an element in *Möb*$^+(\mathbb{D})$ if and only if $\alpha\bar\alpha - \beta\bar\beta > 0$, and we can normalize α and β such that $\alpha\bar\alpha - \beta\bar\beta = 1$. Note that $(k\alpha, k\beta)$ defines the same function as (α, β) only if k is *real*. Hence we can only normalize by multiplication by real numbers. Any such normalization will preserve the sign of $\alpha\bar\alpha - \beta\bar\beta$.

Now we exploit the fact that we can think of *Möb*$^-(\mathbb{D})$ as the coset *Möb*$^+(\mathbb{D})\, g$ of any element $g \in$ *Möb*$^-(\mathbb{D})$. The simplest such element is complex conjugation $g(z) = \bar z$, which is easily verified to correspond to $z \mapsto 1/\bar z$ in *Möb*$^-(\mathbb{H})$. Thus we immediately conclude that *Möb*$^-(\mathbb{H})$ corresponds to maps of the form

$$f(z) = \frac{\alpha\bar z+\beta}{\bar\beta\bar z+\bar\alpha}, \tag{2.7.3}$$

with $\alpha\bar{\alpha} - \beta\bar{\beta} = 1$.

Examples. (1) Complex conjugation in $\mathbb{D}$ corresponds to inversion in the circle ($\mathbb{H}$–line) $|z| = 1$.

Similarly we see that inversion (reflection) in the imaginary axis $\mathbb{H}$ corresponds to reflection in the imaginary axis in $\mathbb{D}$.

(2) Let us determine the elements in $\mathit{M\ddot{o}b}(\mathbb{D})$ which have 0 as a fixpoint. For $\mathit{M\ddot{o}b}^+(\mathbb{D})$ this means $\dfrac{\alpha \cdot 0 + \beta}{\bar{\beta} \cdot 0 + \bar{\alpha}} = 0$ — i. e. $\beta = 0$. The condition $\alpha\bar{\alpha} - \beta\bar{\beta} = 1$ gives $|\alpha| = 1$, and we can write $\alpha = \mathrm{e}^{i\theta}$ for some $\theta \in \mathbb{R}$. Hence any element in $\mathit{M\ddot{o}b}^+(\mathbb{D})$ which has 0 as fixpoint can be written $f(z) = \alpha z/\bar{\alpha} = \alpha^2 z = e^{i2\theta}z$. But this is the formula for a rotation by the angle 2θ, written as complex multiplication. Conversely, any such rotation is an element in $\mathit{M\ddot{o}b}^+(\mathbb{D})$.

A short calculation shows that the transformation $G^{-1}fG$ of $\mathbb{H}$ that this rotation corresponds to is the elliptic transformation $g_\theta(z) = \dfrac{\cos\theta z + \sin\theta}{-\sin\theta z + \cos\theta}$ studied in section 3.

Similar considerations show that an element of $\mathit{M\ddot{o}b}^-(\mathbb{D})$ having 0 as fixpoint has the form $f(z) = e^{i\theta}\bar{z}$, which is a reflection in a line (diameter in D) forming an angle of $\theta/2$ with the x-axis. (For example, we can write $e^{i\theta}\bar{z} = e^{i\theta/2}\overline{e^{-i\theta/2}z}$, which means that we get $e^{i\theta}\bar{z}$ from z by first rotating by the angle $-\theta/2$, then reflecting in the x-axis, and finally rotating back by the angle $\theta/2$.) These mappings — rotations and reflections — form the group of orthogonal linear transformations in dimension 2, corresponding to the group $O(2)$ of orthogonal 2×2–matrices. Hence we have shown that the set of elements of $\mathit{M\ddot{o}b}(\mathbb{D})$ fixing 0 is precisely the group $O(2)$ acting on $\mathbb{R}^2$, restricted to the open unit disk $\mathbb{D}$.

Note that via the isomorphism with $\mathit{M\ddot{o}b}(\mathbb{H})$ (conjugation by G), the subgroup of rotations (isomorphic to ($SO(2)$)) corresponds to the *elliptic* transformations of $\mathbb{H}$ having i as fixpoint. This is the identification of these transformations with rotations that we promised in section 3.

More generally, also in $\mathit{M\ddot{o}b}(\mathbb{D})$ we have a classification of elements according to the behaviour of fixpoints, and we can define parabolic, hyperbolic and elliptic elements as before, corresponding to the elements of the same types in $\mathit{M\ddot{o}b}(\mathbb{H})$. Similarly, in $\mathit{M\ddot{o}b}^-(\mathbb{D})$ the inversions are the transformations with a $\mathbb{D}$–line of fixpoints, corresponding to the inversions in $\mathit{M\ddot{o}b}^-(\mathbb{H})$. Cf. exercises 2 and 3.

We can also transfer the metric d from $\mathbb{H}$ to $\mathbb{D}$. Let us write $d_{\mathbb{H}}$ for d as defined in section 5. Then the formula

$$d_{\mathbb{D}}(z_1, z_2) = d_{\mathbb{H}}(G^{-1}(z_1), G^{-1}(z_2))$$

will define a metric on $\mathbb{D}$ such that G and G^{-1} become inverse *isometries* (distance-preserving maps). In particular, the analogues of conditions (d1)-(d5) in section 5 will automatically hold. We will now derive a more explicit expression for $d_{\mathbb{D}}$.

Recall that $d_{\mathbb{H}}(G^{-1}(z_1), G^{-1}(z_2)) = |\ln(|[G^{-1}(z_1), G^{-1}(z_2), P, Q]|)|$, where P, Q are the endpoints of the $\mathbb{H}$–line through $G^{-1}(z_1)$ and $G^{-1}(z_2)$. But G^{-1} maps $\mathbb{D}$–lines to $\mathbb{H}$–lines, so $\{P, Q\} = \{G^{-1}(p), G^{-1}(q)\}$, where p and q are the endpoints of the $\mathbb{D}$–line through z_1 and z_2. Hence

$$d_{\mathbb{D}}(z_1, z_2) = |\ln(|[G^{-1}(z_1), G^{-1}(z_2), G^{-1}(p), G^{-1}(q) = |\ln(|[z_1, z_2, p, q]||\,, \tag{2.7.4}$$

by Proposition 2.2.10(iii). This is completely analogous to the formula for $d_{\mathbb{H}}$, but now all four points are in $\mathbb{C}$, so we can always write

$$d_{\mathbb{D}}(z_1, z_2) = |\ln(|\frac{z_1 - p}{z_1 - q}\,\frac{z_2 - q}{z_2 - p}|)|\,.$$

The extra symmetry in $\mathbb{D}$ can be used to study the metric in more detail, and in particular it will enable us to derive formulas for $d_{\mathbb{D}}(z_1, z_2)$ not involving the endpoints p and q.

First observe that since rotations around the origin are isometries, the distance from 0 to a point z must be equal to the distance from 0 to r, where $r = |z| \in [0, 1) \in \mathbb{D}$. But the endpoints of the $\mathbb{D}$–line through 0 and r are 1 and -1, so the distance formula gives

$$d_{\mathbb{D}}(0, z) = d_{\mathbb{D}}(0, r) = |\ln(|\frac{0 - (-1)}{0 - 1}\,\frac{r - 1}{r - (-1)}|)| = \ln(\frac{1 + r}{1 - r})\,.$$

This equation can also be solved for r, yielding

$$r = \tanh(\frac{d_{\mathbb{D}}(0, z)}{2})\,. \tag{2.7.5}$$

To find the distance between two arbitrary points z_1 and z_2, we now first move z_1 to 0 by an isometry $f \in Möb^+(\mathbb{D})$, and then use the formula above for the distance between 0 and $f(z_2)$.

If $f(z) = \dfrac{az + b}{\bar{b}z + \bar{a}}$ satisfies $f(z_1) = 0$, then $b = -az_1$, Therefore we can write

$$f(z) = \frac{a(z-z_1)}{-\overline{a z_1}z + \bar{a}} = \frac{a}{\bar{a}}\left(\frac{z-z_1}{-\bar{z}_1 z + 1}\right).$$

Introducing the notation $\rho = |f(z_2)| = \dfrac{|z_2 - z_1|}{|1-\bar{z}_1 z_2|}$, we now have

$$d_{\mathbb{D}}(z_1,z_2) = \ln(\frac{1+\rho}{1-\rho}), \quad \text{or} \quad \rho = \tanh(\frac{d_{\mathbb{D}}(z_1,z_2)}{2}).$$

Another expression is obtained from the second of these by using the formula of Exercise 8b:

$$\sinh^2(\frac{d_{\mathbb{D}}}{2}) = \frac{\tanh^2(d_{\mathbb{D}}/2)}{1-\tanh^2(d_{\mathbb{D}}/2)} = \frac{\rho^2}{1-\rho^2}.$$

Substituting for ρ, we have

$$\sinh^2(\frac{d_{\mathbb{D}}}{2}) = \frac{\dfrac{|z_2-z_1|^2}{|1-\bar{z}_1 z_2|^2}}{1-\dfrac{|z_2-z_1|^2}{|1-\bar{z}_1 z_2|^2}} = \frac{|z_2-z_1|^2}{|1-\bar{z}_1 z_2|^2 - |z_2-z_1|^2}.$$

The denominator here is

$$\begin{aligned}(1-\bar{z}_1 z_2)(1-z_1\bar{z}_2) - (z_1-z_2)(\bar{z}_1-\bar{z}_2) = \\ 1 - \bar{z}_1 z_2 - z_1\bar{z}_2 + |z_1|^2|z_2|^2 - (|z_1|^2 - z_1\bar{z}_2 - \bar{z}_1 z_2 + |z_2|^2) = \\ 1 - |z_1|^2 - |z_2|^2 + |z_1|^2|z_2|^2 = (1-|z_1|^2)(1-|z_2|^2).\end{aligned}$$

Hence we get

$$\sinh^2(\frac{d_{\mathbb{D}}}{2}) = \frac{|z_2-z_1|^2}{(1-|z_1|^2)(1-|z_2|^2)}.$$

Finally we make use of the identity $\cosh(2w) = 2\sinh^2(w) + 1$ with $w = d_{\mathbb{D}}/2$, to obtain the perhaps most convenient formula for the metric:

$$\cosh(d_{\mathbb{D}}(z_1,z_2)) = 1 + \frac{2|z_2-z_1|^2}{(1-|z_1|^2)(1-|z_2|^2)}. \tag{2.7.6}$$

We may now go back to $\mathbb{H}$, using the isometry G, to obtain a similar formula for $d_{\mathbb{H}}$:

$$d_{\mathbb{H}}(w_1,w_2) = d_{\mathbb{D}}(G(w_1),G(w_2)) = d_{\mathbb{D}}(\frac{iw_1+1}{w_1+i},\frac{iw_2+1}{w_2+i})\,.$$

A simple calculation gives

$$\frac{iw_2+1}{w_2+i} - \frac{iw_1+1}{w_1+i} = \cdots = \frac{2w_1-2w_2}{(w_2+i)(w_1+i)}\,, \quad \text{and}$$

$$1-|\frac{iw+1}{w+i}|^2 = \frac{(w+i)(\bar{w}-i)-(iw+1)(-i\bar{w}+1)}{|w+i|^2} =$$
$$\cdots = \frac{2\,i\bar{w}-2\,iw}{|w+i|^2} = \frac{4\operatorname{Im} w}{|w+i|^2}\,.$$

Substituting these expressions into the formula for $d_{\mathbb{D}}$ above, we get:

$$\cosh(d_{\mathbb{H}}(w_1,w_2)) = 1 + \frac{|w_2-w_1|^2}{2\,(\operatorname{Im} w_1)(\operatorname{Im} w_2)}\,. \qquad (2.7.7)$$

These equations (2.7.6 and 2.7.7) express the metrics in $\mathbb{D}$ and $\mathbb{H}$ as functions of the points only, without referring to the endpoints of lines. Note that the function $\cosh(t)$ is increasing for $t \geq 0$, hence the equations determine the metrics uniquely.

Exercises for 2.7

1. Show that the restriction $G^{-1}|\,S^1 : S^1 \to \overline{\mathbb{R}}$ of the fractional linear transformation G is the analogue of stereographic projection from $i \in S^1$.

2. Discuss a classification of the elements of $M\ddot{o}b^+(\mathbb{D})$ analogous to the classification of elements of $M\ddot{o}b^+(\mathbb{H})$ in section 3.

3. Show that an element of $M\ddot{o}b(\mathbb{D})$ has a $\mathbb{D}$–line of fixpoints if and only if it is the restriction of an inversion in a $\overline{\mathbb{C}}$–circle. (Hence the term 'inversion' is well–defined.)

   Show that $g(z) = \dfrac{\alpha\bar{z}+\beta}{\bar{\beta}\bar{z}+\bar{\alpha}}$ determines an inversion in $\mathbb{D}$ if and only if $|\alpha|^2-|\beta|^2>0$ and $\operatorname{Re}\beta = 0$.

4. (a) Show that *hyperbolic circles*, i. e. subsets of $\mathbb{D}$ of the form

   $\{z \in \mathbb{D} \,|\, d_{\mathbb{D}}(z, z_0) = r\}$, where z_0 is a fixed point and $r > 0$, also are Euclidean circles.

   (b) The same problem with $\mathbb{H}$ instead of $\mathbb{D}$.

5. Fix a point z_0 on a hyperbolic line l, and consider (hyperbolic) circles through z_0 with center on l. Show that as the center approaches an endpoint of l, the circle approaches a horocircles.

6. Show that $d_{\mathbb{D}}$ defines the subspace topology on $\mathbb{D} \subset \mathbb{C}$.

   (Hint: Show first (1) $d_{\mathbb{D}} : \mathbb{D} \times \mathbb{D} \to [0, \infty)$ is continuous in the Euclidean topology on $\mathbb{D}$, and (2) $d_{\mathbb{D}}(z_1, z_2) \geq |z_1 - z_2|$ for all z_1 and z_2 in D.)

7. Prove a converse to Exercise 2.5.3, i. e. show that if l_1 and l_2 are two lines which do not intersect and $d(l_1, l_2) = 0$, then they have a common endpoint. (You might need hint (2) in Exercise 6.)

   Same question for $\mathbb{D}$.

8. In this and the next two sections we use several relations between the hyperbolic functions. Verify the following formulas:

   (a) $\cosh^2 x - \sinh^2 x = 1$,

   (b) $\sinh^2 x = \dfrac{\tanh^2 x}{1 - \tanh^2 x}$,

   (c) $\sinh 2x = 2 \sinh x \cosh x$,

   (d) $\cosh 2x = \cosh^2 x + \sinh^2 x$,

   (e) $\dfrac{1 + \tanh^2 x}{1 - \tanh^2 x} = \cosh 2x$,

   (f) $\dfrac{2 \tanh x}{1 - \tanh^2 x} = \sinh 2x$.

2.8 Arc–length and area in the hyperbolic plane

Suppose that $\mathcal{C}$ is a curve in a metric space, given by a parametrization $z(t)$, $t \in [a, b]$, and assume for simplicity that z is injective. The *arc–length* of $\mathcal{C}$ is defined as

$$\sup_{a = t_0 < t_1 < \cdots < t_n = b} \sum_i d(z(t_i) z(t_{i+1}))\,, \tag{2.8.1}$$

provided this number is finite. (The supremum is taken over all partitions $a = t_0 < t_1 < \cdots < t_n = b$ of the interval $[a, b]$.) If so, we say that the curve is *rectifiable.* This is clearly the case if $z(t)$ satisfies a *Lipschitz* condition on $[a, b]$ — i. e. if there exists a constant K such that

$$d(z(t_1), z(t_2)) \leq K|t_1 - t_2|$$

for all t_1, t_2 in $[a, b]$.

If $\mathcal{C}$ is rectifiable, then the arc–length of the restriction of z to $[a, t]$ exists for every $t \in [a, b]$ and defines a continuous, non-decreasing function $s(t)$ on $[a, b]$. This is not the place to discuss the general theory, but one can show that if the limit

$$\sigma(t) = \lim_{h \to 0} \frac{d(z(t + h), z(t))}{|h|},$$

exists and is continuous at every t, then $s(t)$ is given as the integral

$$s(t) = \int_a^t \sigma(\tau)\, d\tau\,.$$

Hence

$$s'(t) = \sigma(t) = \lim_{h \to 0} \frac{d(z(t + h), z(t))}{|h|}\,. \tag{2.8.2}$$

In particular, this condition will be satisfied in $\mathbb{R}^2$, $\mathbb{H}$ or $\mathbb{D}$ whenever $z(t)$ is $\mathcal{C}^1$ as a curve in $\mathbb{R}^2$ — i. e. whenever both component functions are continuously differentiable.

The following is an important observation: Suppose $g : X \to Y$ is an isometry between (possibly different) metric spaces X and Y, and let $z(t),\ t \in [a, b]$ be a curve in X with image curve $gz(t)$ in Y. Then, if z has arc–length s, gz will also have arc–length s — i. e. *arc length is preserved by isometries.* This follows immediately from the definition (2.8.1), since we then always will have $d(gz(t_{i+1}), gz(t_i)) = d(z(t_{i+1})z(t_i))$.

Example 2.8.1. If $z(t)$ parametrizes a segment of a hyperbolic line, condition (d4) for a metric in in chapter 2.5 implies that the arc-length is equal to the hyperbolic distance between the endpoints. Moreover, an obvious generalization of the triangle inequality (d3) shows that $d(z(a), z(b)) \leq \Sigma_i d(z(t_i), d(z_{i+1}))$ for every partition $a = t_0 < t_1 < \cdots < t_n = b$ of $[a, b]$. Hence the hyperbolic line is the shortest possible curve between the two points.

Example 2.8.2. In $\mathbb{R}^2$ we have the formula

$$s'(t) = \sqrt{(\frac{dx}{dt})^2 + (\frac{dy}{dt})^2}\,, \tag{2.8.3}$$

where $z(t) = (x(t), y(t))$, provided $z(t)$ is continuously differentiable. We may also write this as $(\frac{ds}{dt})^2 = (\frac{dx}{dt})^2 + (\frac{dy}{dt})^2$. This formula is valid for any parametrization, and we express this by the relation

$$ds^2 = dx^2 + dy^2\,. \tag{2.8.4}$$

(This relation can be given a precise interpretation as an equation in an appropriate vector space, but here it suffices to read it as an generic notation for (2.8.3)).

We will now derive analogous expressions for the arc-length in the two models $\mathbb{H}$ and $\mathbb{D}$ for the hyperbolic plane.

We start with $\mathbb{H}$. The distance formula (2.7.7) gives:

$$\cosh(d_{\mathbb{H}}(z(t+h), z(t))) = 1 + \frac{|z(t+h) - z(t)|^2}{2\,(\operatorname{Im} z(t+h))(\operatorname{Im} z(t))}\,. \tag{2.8.5}$$

To simplify notation we now write $d(h) = d_{\mathbb{H}}(z(t+h), z(t))$. By Taylor's formula for cosh we can write $\cosh w = 1 + \frac{w^2}{2} + \eta(w)w^2$, where $\lim_{w\to 0}\eta(w) = 0$. (2.8.5) then yields

$$1 + \frac{(d(h))^2}{2} + \eta(d(h))(d(h))^2 = 1 + \frac{|z(t+h) - z(t)|^2}{2\,(\operatorname{Im} z(t+h))(\operatorname{Im} z(t))}\,,$$

and hence

$$(\frac{d(h)}{|h|})^2(1 + 2\,\eta(d(h))) = \left|\frac{z(t+h) - z(t)}{h}\right|^2 \frac{1}{(\operatorname{Im} z(t+h))(\operatorname{Im} z(t))}\,. \tag{2.8.6}$$

It follows that if $z(t)$ is $\mathcal{C}^1$, then $\lim_{h\to 0}\left(\frac{d(z(t+h), z(t))}{|h|}\right)^2$ exist and is equal to $\left((\frac{dx}{dt})^2 + (\frac{dy}{dt})^2\right)/(\operatorname{Im} z(t))^2$. Since the expressions involved are positive, we get

$$s'(t) = \lim_{h\to 0}\frac{d(z(t+h), z(t))}{|h|} = \frac{\sqrt{(\frac{dx}{dt})^2 + (\frac{dy}{dt})^2}}{\operatorname{Im} z(t)}\,.$$

Analogous to the example $\mathbb{R}^2$ above we can also write (remember that $y = \operatorname{Im} z$):

$$ds^2 = \frac{dx^2 + dy^2}{y^2}\,. \tag{2.8.7}$$

For the Poincaré disk we can make a similar analysis. The only difference is that the factor $\dfrac{1}{(\operatorname{Im} z(t+h))(\operatorname{Im} z(t))}$ on the right hand side of formula (2.8.6) is replaced by $\dfrac{4}{(1-|z(t+h)|^2)(1-|z(t)|^2)}$. Thus, in this case we obtain

$$ds^2 = 4\frac{dx^2 + dy^2}{(1-x^2-y^2)^2}\,. \tag{2.8.8}$$

As in the Euclidean case (2.8.4) we should think of these formulas as a way to describe how to get ds/dt from a parametrization $z(t) = (x(t), y(t))$ of the curve. Thus, (2.8.8) means that in $\mathbb{D}$ we have

$$(\frac{ds}{dt})^2 = 4\frac{(\frac{dx}{dt})^2 + (\frac{dy}{dt})^2}{(1-x(t)^2-y(t)^2)^2}\,.$$

Example 2.8.3. Let us apply this to compute the arc–length (circumference) of the hyperbolic circle $\mathcal{C}$ with hyperbolic radius ρ. (Cf. Exercise 2.7.5.) Since arc–length is preserved by isometries, we may assume that $\mathcal{C} \subset \mathbb{D}$ and with center in 0. $\mathcal{C}$ will then also be a *Euclidean* circle with center in 0, with Euclidean radius given by formula (2.7.5) — i. e. $r = \tanh(\rho/2)$. $\mathcal{C}$ may then be parametrized as $z(t) = r\mathrm{e}^{it} = (r\cos t, r\sin t)$, $t \in [0, 2\pi]$, and we get $dx/dt = -r\sin t$, $dy/dt = r\cos t$ and $x^2 + y^2 = r^2$. Thus

$$(\frac{ds}{dt})^2 = 4\frac{(-r\sin t)^2 + (r\cos t)^2}{(1-r^2)^2} = \frac{4r^2}{(1-r^2)^2}\,.$$

Hence the arc–length of $\mathcal{C}$ is

$$s(\mathcal{C}) = \int_0^{2\pi} \frac{2r}{1-r^2}\,dt = \frac{4\pi r}{1-r^2} = \pi\frac{4\,\tanh(\frac{\rho}{2})}{1-\tanh^2(\frac{\rho}{2})} = 2\pi\sinh(\rho)\,.$$

Now recall that

$$2\sinh(\rho) = \mathrm{e}^{\rho} - \mathrm{e}^{-\rho} = 2\rho + \frac{\rho^3}{3} + \frac{\rho^5}{60} + \cdots\,.$$

It follows that the circumference of a circle is greater and increases faster as a function of the radius in the hyperbolic than in the Euclidean plane.

More explicitly, we see that for small ρ the circumference is approximately equal to $2\pi\rho$ (i. e. the same formula as in the Euclidean Case), but when ρ is large, it increases approximately as πe^{ρ}.

Next we discuss *area* in the hyperbolic plane, and in particular we want to find the area of a triangle — i. e. the part of the plane bounded by the three segments between three points not on a line. Hence we do not need the most general concept of area possible, and we will limit our study to subsets of the plane bounded by a finite number of $\mathcal{C}^1$curves. ($\mathcal{C}^1$as curves in $\mathbb{R}^2$.) A reasonable area function A should be *additive* in the sense that for two such subsets U and V we should have $A(U \cup V) + A(U \cap V) = A(U) + A(V)$, and the area of points and smooth curves should be 0. Furthermore, congruent sets should have the same area — in other words: Möbius–transformations should preserve area.

It is not hard to see that such a function will be determined up to a constant scaling–factor (as was also the case for the distance function), so we might just write down the formulas below and show that they satisfy these properties. But hopefully the following informal discussion will help to explain the geometric reason for the formulas and why they are normalized as they are.

Again we first consider $\mathbb{H}$. Let $\Omega \subset \mathbb{H} \subset \mathbb{R}^2$ be a set as described above, and think of the identity map as a parametrization (i. e. as a map from Ω considered as a subset of $\mathbb{R}^2$ to Ω as a subset of $\mathbb{H}$). As usual we now cover *the parameter set* Ω with Euclidean rectangles with vertices (x_i, y_j), where $\{x_i\}_i$ and $\{y_j\}_j$ are increasing sequences of real numbers. Let $R(i,j)$ be the rectangle $[x_i, x_{i+1}] \times [y_j, y_{j+1}]$. Then $\sum_{R(i,j)\cap\Omega\neq\emptyset} A(R(i,j))$ will approximate $A(\Omega)$, and the approximation gets better as the rectangles get smaller. Therefore the area should be given by an integral of the form

$$A_{\mathbb{H}}(\Omega) = \iint_{\Omega} K(x,y)\, dx\, dy\,,$$

where

$$K(x,y) = \lim_{\Delta x, \Delta y \to 0} \frac{A_{\mathbb{H}}(R(\Delta x, \Delta y))}{\Delta x \Delta y}\,,$$

and $R(\Delta x, \Delta y)$ denotes the rectangle $[x, x+\Delta x] \times [y, y+\Delta y]$. The two edges $[x, x+\Delta x] \times \{y\}$ and $\{x\} \times [y, y+\Delta y]$ of $R(\Delta x, \Delta y)$ are curves meeting orthogonally in $\mathbb{H}$, hence $A_{\mathbb{D}}(R(\Delta x, \Delta y))$ ought to be approximated by the product of the hyperbolic lengths of these edges. It is therefore natural to normalize $A_{\mathbb{H}}$ by requiring

$$\lim_{\Delta x, \Delta y \to 0} \frac{A_{\mathbb{H}}(R(\Delta x, \Delta y))}{d_{\mathbb{H}}(x, x+\Delta x)\, d_{\mathbb{H}}(y, y+\Delta y)} = 1\,.$$

But from (2.8.7) we get

$$\lim_{\Delta x \to 0} \frac{d_{\mathbb{H}}(x, x+\Delta x)}{|\Delta x|} = \lim_{\Delta y \to 0} \frac{d_{\mathbb{H}}(y, y+\Delta y)}{|\Delta y|} = \frac{1}{y}\,.$$

Putting all this together we get $K(x,y) = \dfrac{1}{y^2}$, and hence

$$A_{\mathbb{H}}(\Omega) = \iint_{\Omega} \frac{dx\,dy}{y^2}\,. \tag{2.8.9}$$

This equation we now take to be our definition of the area function on $\mathbb{H}$, and the area is defined for every set for which the integral is defined.

This discussion can also be used to prove that the area is invariant under congruence — i. e. $A_{\mathbb{H}}(g\Omega) = A_{\mathbb{H}}(\Omega)$ if g is a Möbius transformation — but it may be instructive to see how this can also be verified from the formula.

Let $g(z) = \dfrac{az+b}{cz+d}$, with $a, b, c, d \in \mathbb{R}$ and $ad - bc = 1$, and assume that $\Omega' = g\Omega$. In order to distinguish between Ω and Ω' we use the notation $z = x + iy$ for points in Ω and $w = u + iv$ for points in Ω'.

The formula for change of variables in a double integral gives

$$A_{\mathbb{H}}(\Omega') = \iint_{\Omega'} \frac{du\,dv}{v^2} = \iint_{\Omega} |J(g)(z)|\, \frac{dx\,dy}{(\operatorname{Im} g(z))^2}\,. \tag{2.8.10}$$

where $|J(g)|$ is the determinant of the Jacobian of g considered as a mapping between subsets of $\mathbb{R}^2$. But the Cauchy–Riemann equations for the complex analytic mapping g imply that $|J(g)(z)| = |g'(z)|^2$. (Exercise 2.) In our case $g'(z) = (ad - bc)/(cz+d)^2$, hence $|J(g)(z)| = 1/|cz+d|^4$.

By formula (2.2.1) we have $\operatorname{Im}(g(z)) = \operatorname{Im} z/|cz+d|^2 = y/|cz+d|^2$. Substituting all this in (2.8.10), we get

$$A_{\mathbb{H}}(\Omega') = \iint_{\Omega} \frac{1}{|cz+d|^4} \frac{dx\,dy}{\left(\dfrac{y}{|cz+d|^2}\right)^2} = \iint_{\Omega} \frac{dx\,dy}{y^2} = A(\Omega)\,.$$

To show that *all* Möbius transformations preserve area it now suffices to observe that reflection in the imaginary axis, $\gamma(z) = -\bar{z}$, does, since $\textit{Möb}(\mathbb{H})$ is generated by γ and $M\ddot{o}b^{+}(\mathbb{H})$. But γ preserves the y–coordinate and has Jacobian equal to -1, so (2.8.10) again gives $A_{\mathbb{H}}(\Omega') = A_{\mathbb{H}}(\Omega)$.

Before we apply this to compute the area of a hyperbolic triangle, we need to remark that in addition to the ordinary triangles determined by three vertices in $\mathbb{H}$, we can also consider *asymptotic* triangles, with one or more "vertices" in $\overline{\mathbb{R}}$ — so–called *ideal* vertices. The two edges meeting at an ideal vertex are then $\mathbb{H}$–lines or rays with this vertex as common endpoint. We talk about *simply*, *doubly* or *triply* asymptotic triangles if there are one, two or three ideal vertices.

Example 2.8.4. Area of a triangle. Every finite triangle in $\mathbb{H}$ is congruent to a triangle with one side along the imaginary axis and where the third vertex has positive real part. Figure 2.8.1 shows such a triangle, with vertices A, B and C. m and r are the center and the radius of the circular arc (hyperbolic line) $\overline{AB}$, spanned by A and B. The other lower–case letters denote the sizes of the obvious angles, thus for example, b is the angular measure of the hyperbolic angle $\angle ABC$, which is the same as the Euclidean angle between the tangents of the two circular arcs meeting at B.

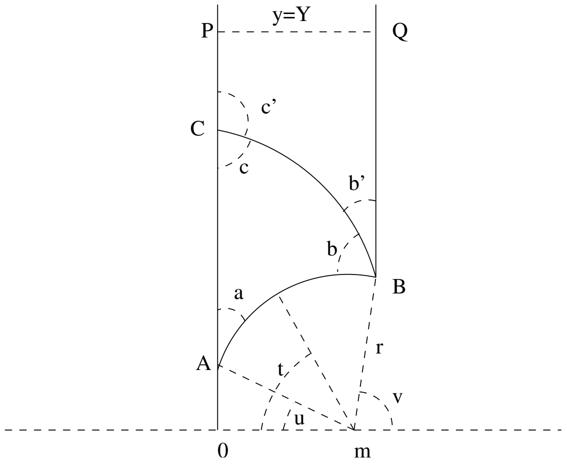


Fig. 2.8.1:

First we compute the area of the region $ABQP$ bounded by the horizontal Euclidean segment $y = Y$ and the hyperbolic segments $[A, P]$, $[A, B]$ and $[B, Q]$. From the figure we see that we can parametrize x as $x = m - r\cos t,\ t \in [u, \pi - v]$. Then $dx = r \sin t\, dt$, and for every t, y ranges from $r \sin t$ to Y. We get

$$\begin{aligned} \iint_{ABQP} \frac{dx\,dy}{y^2} &= \int_u^{\pi - v} \left[\int_{r \sin t}^{Y} \frac{dy}{y^2} \right] r \sin t\, dt \\ &= \int_u^{\pi - v} \left[\frac{1}{r \sin t} - \frac{1}{Y} \right] r \sin t\, dt \\ &= \pi - u - v - \frac{m + r \cos v}{Y}\,. \end{aligned}$$

Note that as Y goes to ∞, this expression approaches $\pi - u - v$. This means that the asymptotic triangle with vertices A, B and ∞ has *finite* area equal to $\pi - (u + v)$. Now observe that the Euclidean lines mO and mA meet the $\mathbb{H}$–segments $[AP]$ and $[AB]$ orthogonally, hence $u = a$. Similarly $v = b + b'$, hence $u + v$ equals the sum of the angles of the triangle, since the third angle is 0. Letting one or both of the vertices A and B approach the real axis, we see that this formula remains valid even for doubly or triply asymptotic triangles.

The area of the finite triangle ABC is equal to the difference between the areas of two such asymptotic triangles — one with area $\pi - (a + b + b')$ and the other with area $\pi - (b' + c')$. Hence the area of ABC is $\pi - a - b - b' - \pi + b' + c' = \pi - a - b - c$, since $c + c' = \pi$. (See figure 2.8.1.) Thus we have proved

Proposition 2.8.5. *The area of a triangle with angles a, b and c is given by*

$$\pi - (a + b + c)\,.$$

This formula is valid also for asymptotic triangles, i. e. if one or more of the angles are 0.

This is a striking result of fundamental importance. It says that the area only depends on the sum of the angles of the triangle, and since the area is always positive, the sum of the angles in a triangle is always less than π. We also see that the area of a triangle never exceeds π, and the maximal value π is only achieved by the triply asymptotic triangles, (which are all congruent, cf. Exercise 2.2.7).

We may also consider the area function for the disk model $\mathbb{D}$. A similar analysis then leads to the formula

$$A_{\mathbb{D}}(\Omega) = \iint_{\Omega} \frac{4\,dx\,dy}{(1 - x^2 - y^2)^2}\,. \tag{2.8.11}$$

Using the change of variables–formula 2.8.10 with g our standard isometry G^{-1} between $\mathbb{D}$ and $\mathbb{H}$ (section 7), we see that G^{-1} — hence also G — is area preserving, in the sense that $A_{\mathbb{H}}(\Omega)$ is defined if and only if $A_{\mathbb{D}}(G(\Omega))$ is defined, and

$$A_{\mathbb{H}}(\Omega) = A_{\mathbb{D}}(G(\Omega))\,.$$

This equation could, of course, also have been used to *define* $A_{\mathbb{D}}$, given $A_{\mathbb{H}}$.

Because of the rotational symmetry in $\mathbb{D}$ it is often convenient to use polar coordinates $x = r\cos\theta,\ y = r\sin\theta$. Then the formula becomes

$$A_{\mathbb{D}}(\Omega) = \iint_T \frac{4r\,dr\,d\theta}{(1-r^2)^2}\,, \tag{2.8.12}$$

where T is the appropriate parameter set in the (r,θ)–plane.

Example 2.8.6. Let us use this to compute the area of a hyperbolic circular disk $\mathcal{D}$ with hyperbolic radius ρ. As in example 2.8.3 we may assume that the circle is a Euclidean circle with center 0. The Euclidean radius is then $R = \tanh(\rho/2)$ (2.7.5), and we can parametrize $\mathcal{D}$ by polar coordinates $x = r\cos\theta,\ y = r\sin\theta$, where $r \in [0,R]$ and $\theta \in [0,2\pi]$. From formula (2.8.12) we get

$$\begin{aligned} A_{\mathbb{D}}(\mathcal{D}) = \int_0^R \left[\int_0^{2\pi} \frac{4r\,d\theta}{(1-r^2)^2}\right] dr =& 2\pi \int_0^R \frac{4r\,dr}{(1-r^2)^2} \\ =& 2\pi \left[\frac{2}{(1-r^2)}\right]_0^R = 4\pi\frac{R^2}{1-R^2}\,. \end{aligned}$$

But $R = \tanh(\rho/2)$, and the formula in Exercise 2.7.8 b gives

$$A_{\mathbb{D}}(\mathcal{D}) = 4\pi\frac{R^2}{1-R^2} = 4\pi\sinh^2(\frac{\rho}{2})\,.$$

Using more of Exercise 2.7.8 and Taylor expansion we get

$$A_{\mathbb{D}}(\mathcal{D}) = 2\pi(\cosh(\rho)-1) = \pi(\mathrm{e}^{\rho} + \mathrm{e}^{-\rho} - 2) = \pi(\rho^2 + \frac{\rho^4}{12} + \cdots)\,.$$

This means that the area of a circular disk is greater and increases faster with the radius in the hyperbolic plane than in the Euclidean plane. (Just as we observed for the circumference of a circle in example 2.8.3). In differential geometry this is expressed by saying that the hyperbolic plane has *negative curvature.*

It is also interesting to compare with geometry on a sphere of radius one. There the circumference of a circle of radius ρ is equal to $2\pi \sin \rho$, and the area is $2\pi \sin^2(\rho/2)$. Both are smaller and increase slower than in the Euclidean case. We say that the sphere has *positive curvature*, whereas the Euclidean plane has curvature 0.

Exercises for 2.8

1. Let $z_1 = a_1 + ib$ and $z_2 = a_2 + ib$ be two points in $\mathbb{H}$ with the same imaginary value. Let L be the hyperbolic arc–length of the *Euclidean* segment between z_1 and z_2. Compute L and show that $L > d_{\mathbb{H}}(z_1, z_2)$.

2. To show invariance of area under Möbius transformations we used that $|J(g)(z)| = |g'(z)|^2$ for a complex analytic function g. Verify this.

3. Show by calculation that the isometry $G : \mathbb{H} \to \mathbb{D}$ of section 7 is area preserving.

4. Find an expression for the area of a hyperbolic n–gon.

5. Let T_α be a doubly asymptotic triangle in $\mathbb{D}$ with one vertex in 0 and the angle there equal to α. Show that $\lim_{\alpha \to 0^+} A(T_\alpha) = \pi$, even though the triangles degenerate to a ray in the limit.

6. Let T be an asymptotic quadrilateral in $\mathbb{D}$ with one finite vertex with angle α and three ideal vertices.

   a) Find a formula for the area of T and show that the area only depends on α.

   b) Does α determine T up to congruence?

2.9 Trigonometry in the hyperbolic plane

In Euclidean geometry fundamental roles are played by the formulas known as the *Law of Sines* and the *Law of Cosines.* For instance, these formulas imply that certain combinations of three angles or sides in a triangle determine the triangle up to congruence (“congruence criteria”) — in fact, they even provide simple ways of computing the remaining angles and sides. In this section we will derive similar formulas for *hyperbolic* triangles. For this it will be convenient to use Poincaré's disk model $\mathbb{D}$ for the hyperbolic plane, but the resulting formulas will be independent of which model we use. In particular, they will also hold in $\mathbb{H}$.

We first consider *finite* triangles, i. e. triples of points in the plane (the *vertices*) which do not lie on a common hyperbolic line. Each pair of vertices spans a segment, a line and two rays. The segments are the *sides* of the triangle, and the *angle* at a vertex is the pair of rays having this vertex in common. An arbitrary such triangle is congruent to one which has one vertex in 0 and another on the positive real axis, and where the third vertex has positive imaginary part: if u, z and w are the vertices, we can move u to 0, rotate such that z lands on the positive real axis, and, if necessary, use complex conjugation to obtain $\operatorname{Im} w > 0$. We then say that the triangle is in "standard position". (But note that for any given triangle there are six possible ways of doing this.) The angles and sides of this new triangle will be congruent to the corresponding angles and sides of the original triangle, so if we want to study relations between their sizes, we may assume that the triangle is in standard position.

Admitting a slight lack of precision, we will simplify our terminology and use the word 'side' interchangeably for a segment and its length, measured in the hyperbolic metric. Similarly, an 'angle' can mean both the actual angle and its size, measured in radians. This is all in accordance with the usual (abuse of) language in Euclidean geometry.

Our general triangle will then have angles α, β and γ, and we denote the opposite sides by a, b and c, respectively. We may assume that α is the angle at 0 and β is the other angle on the real axis. The vertex on the real axis may be identified with a real number $r \in (0,1)$ and the third vertex can be written $w = s\mathrm{e}^{i\alpha}$, where $s \in (0,1)$ and $\alpha \in (0,\pi)$. Then the distance formula (2.7.6) yields:

$$\begin{aligned}\cosh a &= 1 + 2\frac{|r - s\mathrm{e}^{i\alpha}|^2}{(1-r^2)(1-s^2)} = 1 + 2\frac{r^2 + s^2 - 2\,rs\cos\alpha}{(1-r^2)(1-s^2)} \\ &= \frac{1 - r^2 - s^2 + r^2s^2 + 2r^2 + 2s^2 - 4rs\cos\alpha}{(1-r^2)(1-s^2)} \\ &= \frac{1+r^2}{1-r^2}\,\frac{1+s^2}{1-s^2} - \frac{2r}{1-r^2}\,\frac{2s}{1-s^2}\,\cos\alpha.\end{aligned}$$

But $r = \tanh(c/2)$ (2.7.5), and therefore

$$\frac{1+r^2}{1-r^2} = \frac{1+\tanh^2(c/2)}{1-\tanh^2(c/2)} = \cosh c\,, \text{ and}$$

$$\frac{2r}{1-r^2} = \frac{2\tanh(c/2)}{1-\tanh^2(c/2)} = \sinh c\,.$$

Similar formulas hold for s and b, and, substituting in the above expression for $\cosh a$, we have proved

Proposition 2.9.1. *(The first Law of Cosines.)*

$$\cosh a = \cosh b \cosh c - \sinh b \sinh c \cos \alpha.$$

(Obviously there are two more such relations, obtained by permuting the vertices.)

Corollary 2.9.2. *(The hyperbolic Pythagorean theorem)*
If $\alpha = \pi/2$, then $\cosh a = \cosh b \cosh c$.

To see the relationship with the classical Pythagorean Theorem, we substitute the Taylor series for cosh:

$$1 + \frac{a^2}{2} + \frac{a^4}{4!} + \cdots = (1 + \frac{b^2}{2} + \frac{b^4}{4!} + \cdots)(1 + \frac{c^2}{2} + \frac{c^4}{4!} + \cdots)\,.$$

Multiplying out the parentheses and solving with respect to the a^2–term, we see that $a^2 = b^2 + c^2 +$ {terms of order at least 4}. Hence, for small triangles this is approximately the Euclidean Pythagorean Theorem.

Now we put $A = \cosh a$, $B = \cosh b$ and $C = \cosh c$. Then $\sinh a = \sqrt{A^2 - 1}$, etc., and the first law of cosines may be written

$$\sqrt{B^2-1}\sqrt{C^2-1}\cos\alpha = BC - A,$$

or, squaring on both sides:

$$(B^2-1)(C^2-1)(1-\sin^2\alpha) = (BC-A)^2\,.$$

We can solve this equation for $\sin^2\alpha$:

$$\sin^2\alpha = \frac{(B^2-1)(C^2-1)-(BC-A)^2}{(B^2-1)(C^2-1)} =$$

$$\cdots$$

$$= \frac{2\,ABC - A^2 - B^2 - C^2 + 1}{(B^2-1)(C^2-1)}\,. \qquad (2.9.1)$$

Consequently, the quotient

$$\frac{\sin^2\alpha}{\sinh^2 a} = \frac{\sin^2\alpha}{A^2-1} = \frac{2\,ABC - A^2 - B^2 - C^2 + 1}{(A^2-1)(B^2-1)(C^2-1)}\,,$$

is completely symmetric in A, B, C. Hence we get exactly the same if we replace (α, a) by (β, b) or (γ, c), so we have shown:

$$\frac{\sin^2\alpha}{\sinh^2 a} = \frac{\sin^2\beta}{\sinh^2 b} = \frac{\sin^2\gamma}{\sinh^2 c}\,.$$

Since $\sin y$ is positive for $y \in (0, \pi)$ and $\sinh x$ is positive for all $x > 0$, we have proved

Proposition 2.9.3. *(The hyperbolic Law of Sines.)*

$$\frac{\sin\alpha}{\sinh a} = \frac{\sin\beta}{\sinh b} = \frac{\sin\gamma}{\sinh c}\,.$$

Since $\sinh x \approx x$ for small x, we see that for small triangles this is approximately the Euclidean sine relation.

For hyperbolic triangles there is also another cosine formula.

Proposition 2.9.4. *(The second Law of Cosines.)*

$$\cos\alpha = -\cos\beta\cos\gamma + \sin\beta\sin\gamma\cosh a.$$

(Again we obtain two additional formulas by permuting α, β and γ.)

Proof. From the first law of cosines we may write

$$\cos\alpha = \frac{BC - A}{\sqrt{(B^2-1)(C^2-1)}}\,,$$

and similarly for $\cos\beta$ and $\cos\gamma$. Substituting these three expressions, we get

$$\cos\alpha + \cos\beta\cos\gamma =$$
$$\frac{BC - A}{\sqrt{(B^2-1)(C^2-1)}} + \frac{(AC-B)(AB-C)}{(A^2-1)\sqrt{(B^2-1)(C^2-1)}} =$$
$$\cdots = A\frac{1 - A^2 - B^2 - C^2 + 2\,ABC}{(A^2-1)\sqrt{(B^2-1)(C^2-1)}}\,.$$

Using expressions analogous to 2.9.1 for $\sin\beta$ and $\sin\gamma$, we obtain

$$\sin\beta\sin\gamma = \frac{1 - A^2 - B^2 - C^2 + 2\,ABC}{(A^2-1)\sqrt{(B^2-1)(C^2-1)}}\,.$$

Hence we have

$$\cos\alpha + \cos\beta\cos\gamma = A\sin\beta\sin\gamma = \sin\beta\sin\gamma\cosh a\,.$$

□

Remark 2.9.5. This proof show that the second cosine law is a consequence of the first. However, the proof can be reversed, giving the first law as a consequence of the second. Thus the two laws are in fact equivalent.

The first law of cosines is analogous to the classical, Euclidean version, and like its Euclidean counterpart, it implies the SSS ('side–side–side') congruence criterion, stating that the lengths of all three sides determine all the angles; hence the whole triangle up to congruence. An important result that follows from this is:

Proposition 2.9.6. *Congruence of angles can be characterized in terms of congruence of segments.*

Proof. let r, s be rays with vertex A and r', s' rays with vertex A'. Choose vertices $B \in r$ and $C \in s$, both different from A. Using axiom $C1$ we can now find points $B' \in r'$ and $C' \in s'$ such that $A'B' \cong AB$ and $A'C' \cong AC$. Then it follows from the SSS–criterion that $\angle(r,s) \cong \angle(r',s')$ if and only if $B'C' \cong BC$. □

The second cosine relation, however, does not really have a direct counterpart in Euclidean geometry. For example, one striking consequence is that if we know all the angles of a triangle, the *sides* are also completely determined, and hence the whole triangle, up to congruence. Hence, in hyperbolic geometry similar triangles are congruent! This is the AAA congruence criterion, which also holds in *spherical geometry,* but definitely not in Euclidean geometry.

Note that this observation complements the area formula in Proposition 2.8.5, which says that the area is determined by the *sum* of the angles. In fact, if anything, the second law of cosines is another replacement for the result in Euclidean geometry saying that the sum of the angles in a triangle is π, which in terms of trigonometric functions is equivalent to

$$\cos\alpha = \cos(\pi - (\beta+\gamma)) = -\cos\alpha\cos\beta + \sin\alpha\sin\beta.$$

(For α and $\beta + \gamma$ in $[0, \pi]$.) Thus the second hyperbolic cosine law implies that the angle sum goes to 0 if and only if at least one of the sides goes to 0.

Because of the angle–sum formula in Euclidean geometry, two angles of a triangle determine the third. This is not true in hyperbolic geometry, but the second law of cosines says that two angles *and* the side between them determine the third angle. In both geometries the law of sines then determines the remaining two sides, and hence the whole triangle up to congruence. This is known as the *ASA* congruence criterion ('angle–side–angle').

It should be remarked that these congruence criteria, stating that certain combinations of three quantities determine the triangle up to congruence, can also be proved geometrically from the axioms. (Note that axiom C6 is the congruence criterion SAS.) In fact, except for the AAA–criterion, which is only valid in hyperbolic geometry, this can be done without using any parallel axiom. Hence the same geometric proofs are valid in both Euclidean and hyperbolic geometry. But the trigonometric formulas are needed in order to calculate the remaining quantities (angles and sides).

We conclude this section with some remarks on asymptotic triangles. These can be thought of as limiting positions of finite triangles as one or more vertices move to infinity. Recall that we call such vertices *ideal* vertices, and we call a triangle *simply, doubly* or *triply* asymptotic, depending on how many ideal vertices it has. The (size of the) *angle* at an ideal vertex is defined to be 0.

Let us consider more closely the three types of asymptotic triangles:

(i) Triply asymptotic. All three sides are then hyperbolic lines, and any two of these lines have a common endpoint in $\mathbb{R}$. By Exercise 2.2.7 (essentially Corollary 2.2.5), any triple of points can be mapped to any other triple by an element of $M\ddot{o}b(\mathbb{H})$. Hence any two triply asymptotic triangles are congruent.

(ii) Doubly asymptotic. Two of the sides are rays and the third is a hyperbolic line between their endpoints. Consider the triangle in the disk model $\mathbb{D}$. If we move the finite vertex to 0 by a Möbius transformation, we see that the two rays become radii in $\mathbb{D}$. Clearly the angle at 0 then determines the triangle up to congruence.

(iii) Simply asymptotic. Two of the sides are rays and the third is a finite segment of length c, say, between two finite vertices with angles α and β. The third angle is $\gamma = 0$.

Passing to the limit as $\gamma \to 0$, the second cosine relation will continue to hold, and we get

$$1 = \cos(0) = -\cos\alpha\cos\beta + \sin\alpha\sin\beta\cosh c\,.$$

This equation determines the third of the parameters α, β and c if the two others are given. For example, we have

$$\cosh c = \frac{1+\cos\alpha\cos\beta}{\sin\alpha\sin\beta}\,.$$

(See also Exercise 5.)

An important special case is when one of the angles, e. g. β, is $\pi/2$. (See Figure 2.9.1.) Then

$$\cosh c = \frac{1}{\sin\alpha}\,.$$

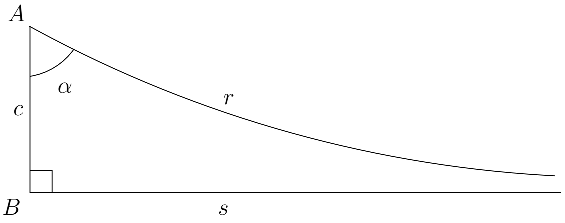


Fig. 2.9.1:

This relation is usually given a different form, obtained by solving the equation for e^{-c}:

$$e^{-c} = \frac{1-\cos\alpha}{\sin\alpha} = \tan\frac{\alpha}{2}\,.$$

These relations are possibly the simplest manifestations of the close relationship between units of measurement of angles and lengths in the hyperbolic plane. In fact, using Lemma 2.6.1 it can be considered as another version of Proposition 2.9.6.

Some classical terminology: let ℓ be the line containing the ray s. The ray r is one of two *limiting parallel rays* to ℓ through the point A. The lines containing the limiting parallel rays are called *asymptotic parallel lines* to ℓ, and the angle α is the *asymptotic angle*. Parallel lines that are not asymptotic are called *ultra-parallel.*

One final word: We have now developed enough of hyperbolic geometry to see that it differs dramatically from Euclidean geometry in the large.

However, several results show that in the small, i. e. locally around a point, the geometries become approximations of each other. Examples are the calculations of area and circumference of a circle, and the remarks on the trigonometric formulas above. This is analogous to the fact that although the Earth is a sphere, locally it looks flat, and the errors we make by using Euclidean geometry on small regions are usually very small.

In chapter 5 we shall see that these differences are consequences of the signs of the *curvatures* defined by the different metrics.

Exercises for 2.9

1 . In a triangle two of the angles are α and β, and the length of the side opposite to the vertex with angle β is b. Explain how to find the remaining angles and sides.

   Why does this establish the congruence criterion SAA?

2. Show that two of the angles of a finite triangle are equal if and only if the two sides opposite to these angles have the same length.

3. Use the hyperbolic sine relation to prove that in a hyperbolic triangle the greatest angle has the longest opposite side.

4. Show that the sides of a triangle all have the same length if and only if the three angles also are equal.

   If the sides have length a and the angles are α, prove the formulae

   $$\cos\alpha = \frac{\cosh a}{1 + cosha}$$
   $$\cosh a = \frac{\cos\alpha}{1 - \cos\alpha}$$
   $2\cosh(a/2)\sin(\alpha/2) = 1$. (Hint: cut the triangle into two pieces.)

5. Show that if a simply asymptotic triangle has finite angles α and β and finite side c, then β is determined by α and c.

6. Suppose given a *quadrilateral* with one ideal vertex and three right angles. Then two of the sides have finite lengths a and b. Show that

   $$\frac{1}{\cosh^2 a} + \frac{1}{\cosh^2 b} = 1\,.$$

7. Show that in a triangle with $\gamma = \pi/2$ the following formulas hold:

$$\begin{aligned} \sin\alpha &= \frac{\sinh a}{\sinh c} \\ \cos\alpha &= \frac{\tanh b}{\tanh c} \end{aligned}$$

8. Formulate and prove a 'converse' of Proposition 2.9.6.

9. Prove that a map $\mathbb{H} \to \mathbb{H}$ is a Möbius transformation if and only if it is distance preserving.

Appendix. Remarks on the Beltrami–Klein model

Since the Beltrami–Klein model played such an important part at the beginning of these notes, we should not end the discussion of hyperbolic geometry without some remarks on the missing bits of its geometry.

When we left it, we had all ingredients except congruence. Now we can define congruence as equivalence under transformations of the form $H^{-1}gH$, where H is our identification $\mathbb{K} \approx \mathbb{D}$ and $g \in Möb(\mathbb{D})$. This can be written out explicitly, but the formulas are ugly and not very enlightening. However, by Proposition 2.9.6 we now also know that congruence is completely determined by the distance measure, so we might instead ask what the distance formula looks like when transported back to $\mathbb{K}$. It turns out that this question does indeed have a nice and interesting answer.

Recall that the set of points of $\mathbb{K}$ is the interior of the unit disk in $\mathbb{R}^2$, and the 'lines' of the geometry are the chords in this disk.

Let Z_1, Z_2 be two points of $\mathbb{K}$. They span a unique chord which has endpoints P and Q on the boundary circle of $\mathbb{K}$ in $\mathbb{R}^2$. Let us denote the distance function on $\mathbb{K}$ by $d_{\mathbb{K}}$.

Proposition A.1. $d_{\mathbb{K}}(Z_1, Z_2) = \frac{1}{2}|\ln|[Z_1, Z_2, P, Q]||.$

Proof. The identification $H : \mathbb{K} \approx \mathbb{D}$ is the composition of the vertical projection from $\mathbb{K}$ to the upper hemisphere $\mathbb{B}$ with equator disk $\mathbb{K}$ and stereographic projection from $\mathbb{B}$ to $\mathbb{D}$. Note that $\mathbb{K}$ and $\mathbb{D}$ coincide as sets.

Let $W_1 = H(Z_1)$ and $W_2 = H(Z_2)$. The endpoints P and Q are left fixed, so we have by definition

$$d_{\mathbb{K}}(Z_1, Z_2) = d_{\mathbb{D}}(W_1, W_2) = |\ln|[W_1.W_2, P, Q]||.$$

Hence the result follows from

Claim 1: $[W_1.W_2, P, Q]^2 = [Z_1, Z_2, P, Q]$.

Now we simplify notation and denote by AB both the Euclidean segment from A to B and its Euclidean length $|A - B|$. Then the equation in Claim 1 reads

$$\left(\frac{W_1P}{W_1Q}\frac{W_2Q}{W_2P}\right)^2 = \frac{Z_1P}{Z_1Q}\frac{Z_2Q}{Z_2P}.$$

Hence Claim 1 is a consequence of

Claim 2: Let Z be a point on the chord with endpoints P and Q, and let $W = H(Z)$. Then $\left(\frac{WP}{WQ}\right)^2 = \frac{ZP}{ZQ}$.

We will prove this by a geometric argument entirely within Euclidean geometry, referring to Figure 2.9.2. Here O is the center of the unit disk, and we have drawn both the $\mathbb{K}$–line and the $\mathbb{D}$–line with endpoints P and Q. Note that the $\mathbb{D}$–line has the radii OP and OQ as tangents. By the definition of π and Φ the point W lies on the line between Z and O, so W must be the point of intersection between the $\mathbb{D}$–line and Euclidean segment OZ.

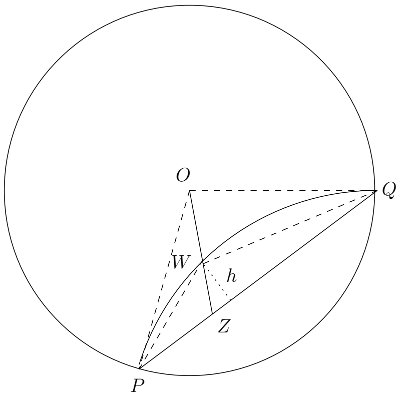


Fig. 2.9.2:

Consider now the triangles ZPO and ZQO. With ZP and ZQ as base segments the two triangles have the same heights, and we have

$$\frac{ZP}{ZQ} = \frac{A(ZPO)}{A(ZQO)},$$

where the right hand side is the ratio between the areas of the triangles.

Likewise, the triangles ZPW and ZQW have the same heights, so we deduce

$$\frac{ZP}{ZQ} = \frac{A(ZPW)}{A(ZQW)} = \frac{A(ZPO) - A(ZPW)}{A(ZQO) - A(ZQW)} = \frac{A(PWO)}{A(QWO)}.$$

But the areas in the last fraction can also be calculated with the two radii OP and OQ as bases, yielding

$$\frac{ZP}{ZQ} = \frac{\frac{1}{2}OP \cdot PW \sin(\angle WPO)}{\frac{1}{2}OQ \cdot QW \sin(\angle WQO)} = \frac{PW \sin(\angle WPO)}{QW \sin(\angle WQO)}.$$

The $\mathbb{D}$–line through P, W, Q is part of a Euclidean circle, hence we can use the theorem of peripheral angles to obtain

$$\angle WPO = \angle WQZ \quad \text{and} \quad \angle WQO = \angle WPZ.$$

But from the figure we see that

$$PW \sin(\angle WPZ) = h = QW \sin(\angle WQZ).$$

Combining these observations, we get

$$\frac{PW}{QW} = \frac{\sin(\angle WQZ)}{\sin(\angle WPZ)} = \frac{\sin(\angle WPO)}{\sin(\angle WQO)}.$$

Substituting this in the expression for $\dfrac{ZP}{ZQ}$ above proves Claim 2. □

Remark. Comparing the formula in Proposition A.1 with the formula for $d_{\mathbb{D}}$ in (2.7.4), it might look as if $d_{\mathbb{K}}$ is just a rescaling of $d_{\mathbb{D}}$. However, this is not the case. The reason is that the points P, Q (p, q) are not the same in the two formulas — in Proposition A.1 they are the endpoints of the chord through Z_1, Z_2, but in (2.7.4) they are endpoints of the $\mathbb{D}$–line through the same points. The only case where they coincide is when Z_1 and Z_2 lie on a common diameter. Then we have $d_{\mathbb{K}}(Z_1, Z_2) = \frac{1}{2}d_{\mathbb{D}}(Z_1, Z_2)$.

We can also derive a formula analogous to (2.7.6). First we observe that if we identify the underlying sets of $\mathbb{K}$ and $\mathbb{D}$ with the unit disk in the complex plane, the homeomorphism $H : \mathbb{K} \to \mathbb{D}$ is given by

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

(Exercise 5.) Using this we shall show that

Proposition A.2. $\cosh(d_{\mathbb{K}}(z_1, z_2) = \dfrac{1 - \langle z_1, z_2 \rangle}{\sqrt{1 - |z_1|^2}\sqrt{1 - |z_2|^2}}$, where $\langle z_1, z_2 \rangle$ is the Euclidean inner product of z_1 and z_2.

Proof. Write $w_i = H(z_i)$, i=1,2. A straightforward calculation starting with (2.7.6) for w_1, w_2 gives

$$\begin{aligned} &\cosh(d_{\mathbb{D}}(w_1,w_2)) \\ &\quad= \frac{1 - |w_1|^2 - |w_2|^2 + |w_1|^2|w_2|^2}{(1 - |w_1|^2)(1 - |w_2|^2)} + \frac{2(|w_1|^2 + |w_2|^2 - 2\langle w_1, w_2 \rangle)}{(1 - |w_1|^2)(1 - |w_2|^2)} \\ &\quad= \frac{(1 + |w_1|^2)(1 + |w_2|^2)}{(1 - |w_1|^2)(1 - |w_2|^2)} - \frac{4\langle w_1, w_2 \rangle}{(1 - |w_1|^2)(1 - |w_2|^2)} \end{aligned} \tag{2.9.2}$$

Now write $w = H(z) = \dfrac{z}{y}$, where $y = 1 + \sqrt{1 - |z|^2}$. Easy calculations then give $y^2 + |z|^2 = 2y$ and $y^2 - |z|^2 = 2y(y - 1)$; hence

$$1 + |w|^2 = 2/y \quad \text{and} \quad 1 - |w|^2 = 2(y - 1)/y.$$

Clearly we also have $\langle w_1, w_2 \rangle = \langle z_1, z_2 \rangle / y_1 y_2$. Substituting this in formula (2.9.2) proves the proposition. □

An application of this result to yet another model for the hyperbolic plane, the *hyperboloid model*, is given in exercise 6.

The Beltrami–Klein model is not conformal, and angle measure is much more complicated. However, one fact is worth noting: it is easy to draw all $\mathbb{K}$–lines which are *orthogonal* to a given line! We distinguish between two cases:

(1) If the given line is a diameter, the $\mathbb{K}$–lines orthogonal to it are the chords that are orthogonal to it in the Euclidean sense.

(2) Assume the $\mathbb{K}$–line ℓ is not a diameter, and let its endpoints be P and Q. The tangents to the circle at ∞ at P and Q intersect in a point W

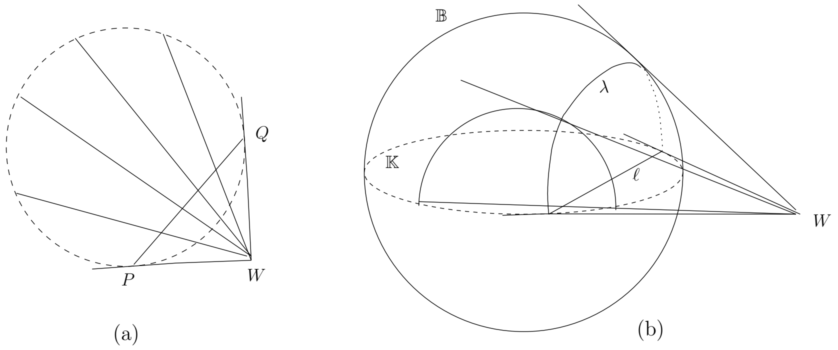


Fig. 2.9.3:

outside the unit disk. The $\mathbb{K}$ lines orthogonal to ℓ are the chords contained in Euclidean lines through W. (See Figure 2.9.3a.)

Note that case (1) follows from case (2) by a limiting procedure. Therefore it suffices to consider (2). We will prove this by a geometric argument using the hemisphere model $\mathbb{B}$ and a little three–dimensional geometry. Cf. Figure 2.9.3b. In this picture we think of $\mathbb{K}$ as lying in a standard $\mathbb{R}^2 \subset \mathbb{R}^3$, and $\mathbb{B} \subset S^2$ is the upper hemisphere.

Via vertical projection the $\mathbb{K}$–lines correspond to $\mathbb{B}$–lines, which are circular arcs (semi–circles) which lie in vertical planes and meet the equator circle at right (Euclidean) angles. In particular, ℓ correspond to one such semi–circle, which we call λ.

The reason why we pass to $\mathbb{B}$ is that the hemisphere model is conformal, since it is related to $\mathbb{D}$ by stereographic projection. Therefore we now only need to determine all $\mathbb{B}$–lines which meet λ orthogonally in the Euclidean sense.

A $\mathbb{B}$–line γ meets λ orthogonally if and only if its tangent line at the intersection point does. But the union of all the lines meeting the circle containing λ orthogonally is easily seen to be a circular cone, and all the lines meet at its vertex. Since the tangents at P and Q in Figure 2.9.3a are two such lines, we see that the cone vertex is precisely the vertex W.

It now only remains to observe that since γ lies in a vertical plane, its tangents also lie in this plane. Hence the plane contains W, and its projection to R^2 will be a line also containing W. But the $\mathbb{K}$–line corresponding to γ is contained in this line.

Exercises for Appendix

1. Which kind of curves are 'horocircles' in $\mathbb{K}$?

2. Verify that inversion in a diameter in $\mathbb{K}$ is ordinary reflection in that diameter.

3. Show that two hyperbolic lines have a common perpendicular if and only they are *ultra-parallel* (end of section 2.9), and show how it can be constructed in the Beltrami–Klein model.

4. Show that we can parametrize $\mathbb{K}$ by $(r, \theta) \mapsto (\tanh r \cos\theta, \tanh r \sin\theta)$, where r is the hyperbolic distance from the origin. ('Geodesic polar coordinates'.)

5. Show that if we identify the underlying sets of $\mathbb{K}$ and $\mathbb{D}$ with the unit disk in the complex plane, the homeomorphism $H : \mathbb{K} \to \mathbb{D}$ and its inverse are given by

$$H(z) = \frac{z}{1 + \sqrt{1 - |z|^2}} \quad \text{and} \quad H^{-1}(w) = \frac{2w}{1 + |w|^2}.$$

6. *The hyperboloid model of the hyperbolic plane.* The hyperboloid in $\mathbb{R}^3$ with equation $z^2 - x^2 - y^2 = 1$ has two sheets; let $\mathbb{L}$ be the one where $z > 0$. (Both sheets would work in the following discussion). Identify $\mathbb{K}$ with the disk where $x^2 + y^2 < 1$ in the plane $z = 1$. Then it is easy to see that every line through a point in $\mathbb{L}$ and the origin in $\mathbb{R}^3$ intersects $\mathbb{K}$ in exactly one point; in fact, this defines a homeomorphism $\Psi : \mathbb{L} \to \mathbb{K}$. (In coordinates: $\Psi(x, y, z) = (x/z, y/z)$.) Hence we get a model for that hyperbolic plane by mapping the geometry of $\mathbb{K}$ to $\mathbb{L}$ by Ψ^{-1}.

   Show that the metric in $\mathbb{L}$ is given by

$$\cosh(d_{\mathbb{L}}((x_1, y_1, z_1), (x_2, y_2, z_2))) = z_1 z_2 - x_1 x_2 - y_1 y_2.$$

   Remark. The expression

$$\ll (x_1, y_1, z_1), (x_2, y_2, z_2) \gg = z_1 z_2 - x_1 x_2 - y_1 y_2$$

   defines a *Minkowski inner product* on $\mathbb{R}^3$. It is not positive definite, but satisfies the other axioms for a non–degenerate inner product. Observe that $\mathbb{L}$ is defined as the set of "unit vectors" for this inner product, and the metric $d_{\mathbb{L}}$ can now be expressed by

$$\cosh(d_{\mathbb{L}}(A_1, A_2)) = \ll A_1, A_2 \gg .$$

Note the similarity with the distance formula on the unit sphere S^2 in Euclidean geometry:

$$\cos(d_{S^2}(A_1, A_2)) = \langle A_1, A_2 \rangle.$$

The model $\mathbb{L}$ is also called the *Minkowski model* or the *Lorentz model* of the hyperbolic plane. It generalizes easily to higher dimensions and leads to the identification of the group of hyperbolic isometries with the *Lorentz group* of linear transformations preserving the Minkowski inner product.

Chapter 3

Topological Classification of compact, closed surfaces

We now move on to study more general two–dimensional objects and their possible geometries. Note that our models for Euclidean or hyperbolic geometry are all homeomorphic to ordinary $\mathbb{R}^2$, and the geometries consist of some extra structure. Now we will consider objects (topological spaces) which are only *locally* homeomorphic to $\mathbb{R}^2$. Then the extra geometric structure can also be defined locally, and our goal is to study to which extent these local structures can be pieced together to give a structure on the whole space. These spaces are the objects we call "surfaces", and in this chapter we shall concentrate on the *topology* of such surfaces. The main result is a complete classification of those that are connected and compact as topological spaces.

To be more precise: A "surface" is here defined as a two–dimensional topological manifold, i. e. a Hausdorff topological space such that every point has a neighborhood which is homeomorphic to $\mathbb{R}^2$. (This is usually expressed by saying that it is *locally* homeomorphic to $\mathbb{R}^2$.)

Here are some examples (we shall see many more later):

(1) $\mathbb{R}^2$, S^2, $T^2 = S^1 \times S^1$,

(2) Open subsets of other surfaces, e. g. the models $\mathbb{H}$ and $\mathbb{D}$ of the hyperbolic plane,

(3) $\{(x, y, z) \in R^3 \mid F(x, y, z) = 0,\ \nabla F(x, y, z) \neq 0\}$, where F is a continuously differentiable function on an open subset of $\mathbb{R}^3$. Examples of

this type include graphs of functions of two variables.

(4) Analogous to (3) we can also consider the zero set of a complex analytic function $f(z,w)$ of two variables where at least one of the partial derivatives is different from 0. These are complex *curves*, but from our real point of view they have dimension 2 and will be considered as surfaces.

A surface is a disjoint union of its connected components, so it suffices to classify *connected* surfaces. Therefore, from now on we assume that all our surfaces are connected. (The word "closed" in the title means that we do not here consider surfaces with boundary.)

An important operation on surfaces (and manifolds in general) is *connected sum*: Given two surfaces M_1 and M_2, we can get another by removing an open disk from each of them and gluing the rest along their boundary circles. More precisely: Let D^2 be the closed 2–disk with boundary S^1, and let $h_i : D^2 \to M_i,\ i = 1, 2$ be two embeddings, — i. e. maps which are homeomorphism onto their images. Then the connected sum of M_1 and M_2, denoted $M_1 \# M_2$, is defined to be the surface

$$M_1 \# M_2 = (M_1 - h_1(\operatorname{int} D^2)) \cup_f (M_2 - h_2(\operatorname{int} D^2)),$$

where $f : h_1(S^1) \to h_2(S^1)$ is the homeomorphism $h_2 h_1^{-1}$.

Clearly, $M_1 \# M_2 \approx M_2 \# M_1$. (The symbol $\approx$ means "is homeomorphic to".)

Remark 3.1.7. This construction makes sense also in higher dimensions, but the result may depend on our choice of h_1, h_2. However, for surfaces all choices turn out to give homeomorphic results. This will be clear after we have proved the classification theorem.

In the following we will use the notation $M \approx N$ for "M is homeomorphic to N". Note, in particular, that $M \# S^2 \approx M$ for any surface M. It is also clear that we can iterate this operation and form the connected sum of several surfaces. The classifications theorem in dimension 2 states that a compact, connected surface is either homeomorphic to a sphere S^2, or it can be written as a connected sum of finitely many copies of only two distinct *irreducible* surfaces — i. e. surfaces which can not be decomposed further. These two surfaces are the *torus* T^2 and the *projective plane* P^2. We will use the following models for T^2 and P^2:

T^2 is defined as $S^1 \times S^1$, but since S^1 is homeomorphic to a closed interval (e. g. $[0, 1]$) where we have identified the endpoints, we can also think of T^2

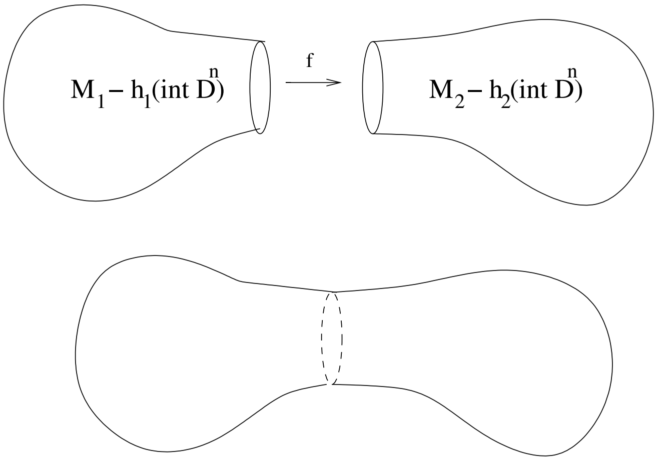


Fig. 3.1.1: Connected sum

as $[0,1] \times [0,1]$ where we identify opposite sides pairwise, as indicated on fig. 3.1.2.

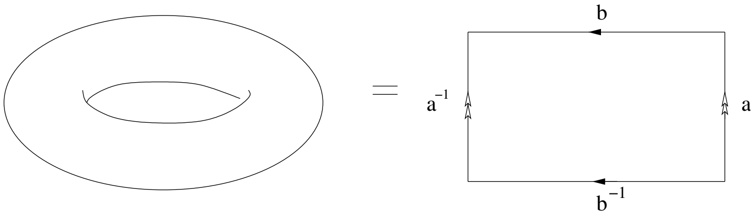


Fig. 3.1.2: Torus

The projective plane is obtained from the 2–sphere S^2 by identifying antipodal points. Then we identify every point in the upper hemisphere with a point in the lower hemisphere, hence we only need one of them to represent all of P^2. Since both hemispheres are homeomorphic to the disk D^2, we get P^2 by identifying antipodal points on the boundary circle S^1. This identification can be realized by cutting the boundary circle into two intervals and identifying as in fig. 3.1.3.

Note that we have described both the torus and the projective plane as identification spaces where we start with a $2n$-gon ($n = 2$ and 1 in these examples) and identify the sides pairwise according to some definite pattern.

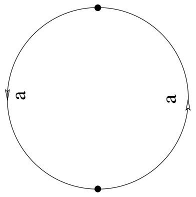


Fig. 3.1.3: P^2

The labels a, a^{-1} etc. on the boundary indicate that we identify the two sides in a way that preserves $(\dots a \dots a \dots)$ or reverses $(\dots a \dots a^{-1} \dots)$ the direction around the $2n$-gon. Thus, T^2 is represented by the string $aba^{-1}b^{-1}$ and P^2 by aa. Another example is the *Klein bottle* K^2, represented by $abab^{-1}$. This surface is not homeomorphic to a subset of $\mathbb{R}^3$, but fig. 3.1.4 is a picture with one circle of self–intersections. Using just the two surfaces T^2 and P^2 and connected sum, we will see that we obtain all other surfaces up to homeomorphism. $T^2 \# T^2$, for example, is a surfaces with “two holes”, as in fig. 3.1.5. In the same way, the connected sum of n copies of T^2 becomes a surface with “n holes”.

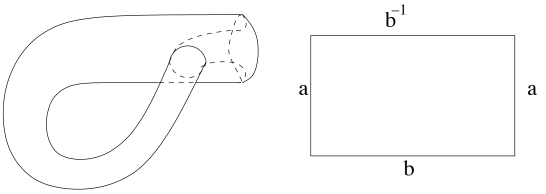


Fig. 3.1.4: Klein bottle

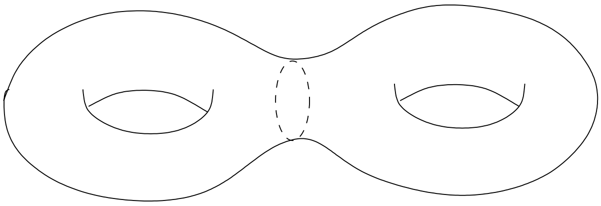


Fig. 3.1.5: $T^2 \# T^2$

How can we show that T^2 and P^2 are not homeomorphic? The simplest is probably to observe that T^2 is *orientable* and P^2 is not. An intuitive way of thinking about orientability of surfaces is as follows: Consider an embedded curve $\omega : [a,b] \to M$. Locally at any point we can distinguish between the two sides of the curve and e. g. name them "right" and "left". Moving along the curve, we can extend this and define the right and left hand side of the whole curve. However, if the curve is *closed*, i. e. $\omega(b) = \omega(a)$, the notions of right and left obtained when we come back to the starting point may not coincide with those we had at the beginning. An example of this is the *Möbius band,* shown in fig. 3.1.6. The arrows indicate a choice of side locally, but we see that no continuous choice is valid for the whole curve. We say that the curve is *one–sided.*

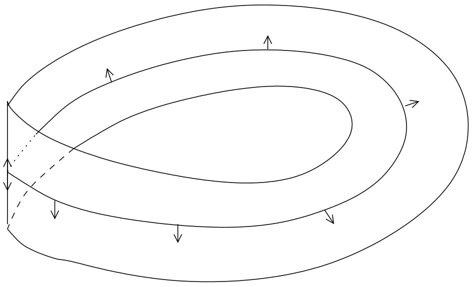


Fig. 3.1.6: Möbius band

If a surface has a one–sided curve, it is called *non–orientable*, and if all closed curves are *two–sided*, it is *orientable.* It is clear that homeomorphic surfaces are either both orientable or both non–orientable.

In fact, if M has a one–sided embedded, closed curve, a neighborhood of this curve looks like a Möbius band. Hence we have

Lemma 3.1.8. *A surface is non–orientable if and only if it contains a Möbius band.*

P^2 is non–orientable, since it contains a Möbius band. This can be seen by doing the identification in fig. 3.1.3 on a band going across the disk from the upper to the lower half–circle. To prove that T^2 or any other surface is orientable is harder, since we have to check a certain property for *all* closed curves. But here is at least a heuristic argument:

T^2 has a representation in $\mathbb{R}^3$ which is *two–sided* in the sense that it separates R^3 into two components. A choice of one of these makes it possible to define right and left hand sides of an embedded curve, “as seen from” the chosen component. One way to do this is to let the ordered triple of directions (“right, forward (moving along the curve) and up”) be chosen according to the right hand rule in $\mathbb{R}^3$. But this choice will not change when we move around the curve, so the curve is also two–sided on the surface. Hence T^2 is orientable.

Remark 3.1.9. This also proves that P^2 does not have a two–sided representation in $\mathbb{R}^3$. In fact, one can show that it has no embedding in $\mathbb{R}^3$ whatsoever.

We can now state the existence part of our classification theorem:

Theorem 3.1.10. *Any compact, connected surface is homeomorphic to a finite connected sum of tori and projective planes.*

Proof. To prove this we need a result by Radó from 1925, saying that every surface can be *triangulated.* This means that it is homeomorphic to a union of triangles, where every edge of every triangle is linearly identified with exactly one edge of another triangle. This is a deep result, but not very surprising. (However, the analogous statement in higher dimensions is false from dimension 4 on!)

Let $\Delta_1, \dots, \Delta_n$ be the triangles in a triangulation of a compact surfaces M. We may choose the ordering such that $\cup_{j<i}\Delta_j$ and Δ_i have at least one edge in common, and we choose such an edge E_i for every i. Let

$$F_k = \Delta_1 \cup_{E_1} \Delta_2 \cup \cdots \cup_{E_{k-1}} \Delta_k,$$

for $k = 1, 2, \dots, n$. Then $F_k = F_{k-1} \cup_{E_k} \Delta_k$, and an easy induction argument shows that each F_k is homeomorphic to a 2-disk D^2, which we may think of as a $(k+2)$–gon. In particular, F_n will be an $(n+2)$–gon, and M is obtained by identifying its edges pairwise. (It follows that $n+2$, hence also n, is even!)

We now give the edges names $a, b, \dots$, such that the edges which are identified get either the same $(\dots a \dots a \dots)$ or “inverse” $(\dots a \dots a^{-1} \dots)$ names, depending on whether they are identified in a way that preserves or reverses direction around the $(n+2)$-gon. (Cf. the descriptions of T^2, K^2 and P^2 above.)

If the word $W = a \dots a^{\pm 1} \dots$ lists the edges counterclockwise, we will write D^2/W for the result of the identifications. Thus the torus above can be written $D^2/aba^{-1}b^{-1}$ and P^2 as D^2/aa, whereas S^2 is homeomorphic to D^2/aa^{-1}. We let W^{-1} denote the same word read clockwise, e. g. $(abc^{-1})^{-1} = cb^{-1}a^{-1}$.

Let us call a word W defining a surface as above *admissible.* It is convenient to allow the empty word $\{\}$ and define $D/\{\} = S^2$. The proof of the theorem is based on the following lemma:

Lemma 3.1.11. (i) *If* $W = W_1W_2$*, then* $D^2/W_1W_2 = D^2/W_2W_1$*.*

(ii) *If* W_1 *and* W_2 *are admissible, then* W_1W_2 *is also admissible, and*

$$D^2/W_1W_2 \approx D^2/W_1 \# D^2/W_2$$

Proof. (i) is only the observation that when we go around the disk, it does not matter where we start reading the word.

(ii) We cut D^2 along a line c connecting the two endpoints on the boundary where W_1 and W_2 meet. (See fig. 3.1.7.) Since W_1 and W_2 are admissible, these two endpoints are identified, and c represents a closed curve in D^2/W_1W_2.

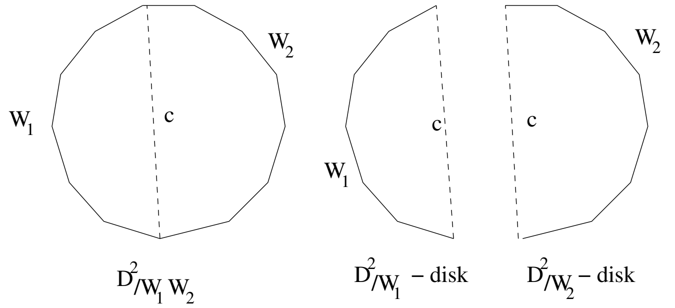


Fig. 3.1.7:

Thus we get two new disks where the edges are given by W_1 and W_2, respectively, plus an extra edge c in both. It is then easy to see that if we perform the identifications encoded by W_1 and W_2 in these two disks but do not identify the c's, we get spaces homeomorphic to $(D^2/W_1 - \text{disk})$ and $(D^2/W_2 - \text{disk})$.

We may construct D^2/W_1W_2 by first performing all the identifications in the two parts and then gluing along the dashed curves (which now have become circles). But this is clearly $D^2/W_1 \# D^2/W_2$.

□

In particular, we have

(1) $D^2/Waa^{-1} \approx D^2/W \# S^2 \approx D^2/W$

(2) $D^2/Waa \approx D^2/W \# P^2$

(3) $D^2/Waba^{-1}b^{-1} \approx D^2/W \# T^2$

Therefore we can eliminate all sequences having the form aa^{-1}, aa and $aba^{-1}b^{-1}$, in the sense that we can write $D^2/W \approx D^2/W' \# M'$, where M' is a connected sum of tori and projective planes, and W' is a word which is at most as long as W and which does not contain any such strings.

Suppose now that $W = W_1aW_2a$. Then we may, as in the proof of the lemma above, cut D^2 into two disks separating the words W_1a and W_2a. But this time we construct the surface by first gluing back along the edges labeled a. The result is a new disk with edge–identification given by the word $W_1W_2^{-1}cc$.

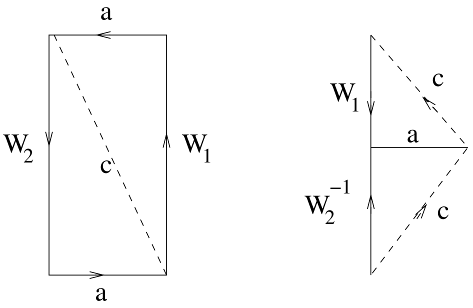


Fig. 3.1.8:

Hence we also have

(4) $D^2/W_1aW_2a \approx D^2/W_1W_2^{-1}cc \approx D^2/W_1W_2^{-1} \# P^2$

Example 3.1.12. The Klein bottle can be represented as $D^2/bab^{-1}a$, which by (4) is homeomorphic to $D^2/bb \# P^2 \approx P^2 \# P^2$.

By repeated use of (2) and (4) we now reduce to the case where all the edges occur in pairs (c, c^{-1}) etc., and by (1) we may eliminate all strings of type cc^{-1} (and $c^{-1}c$). Then we must either have $M \approx S^2$, or we can find (possibly after switching the labels c and c^{-1} and/or d and d^{-1}) edges c and d occurring in the order $c \dots d \dots c^{-1} \dots d^{-1} \dots$ in the word W. (Choose first a pair with minimal distance between them.)

We now cut and paste as in fig. 3.1.9.

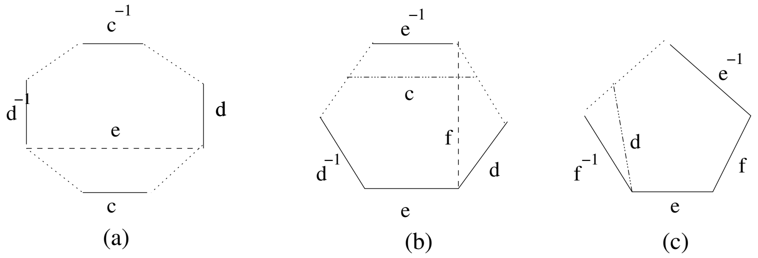


Fig. 3.1.9:

Figure 3.1.9(a) represents the original identifications given by the word W. We cut along a new line e as shown, and the two parts are glued together again along c and c^{-1}. Then we have the situation in (b), which we cut along a new edge f. The resulting parts are glued along d and d^{-1}, and we get (c) in the figure. This represents the same surfaces as (a), and if we put $g = f^{-1}$, we have

$$D^2/W \approx D^2/W'geg^{-1}e^{-1} \approx D^2/W' \# T^2$$

Here W' is a shorter word than W. Putting all this together we may split off summands P^2 and T^2 until we are left with a surface homeomorphic to S^2.

□

Thus every compact, connected surface is homeomorphic to a surface of the type

$$S(m,n) = \underbrace{T^2 \# \cdots \# T^2}_{m} \# \underbrace{P^2 \# \cdots \# P^2}_{n}$$

(We set $S(0,0) = S^2$.) Since T^2 is orientable and P^2 is not, we see that M is orientable if and only if $n = 0$ — i. e. M is homeomorphic to a surface of the type $S(m,0) = T^2 \# \cdots \# T^2$. Then we shall see that the number of summands is uniquely determined.

On the other hand, if M is *not* orientable, then m and n are *not* uniquely determined. In fact, we have:

Lemma 3.1.13. $P^2 \# P^2 \# P^2 \approx T^2 \# P^2$

Proof. We begin by recalling Example 1, from which it follows that

$$P^2 \# P^2 \# P^2 \approx K^2 \# P^2 \approx D^2/abab^{-1}cc$$

Successive applications of Lemma 3.1.11 and the homeomorphism (4) give

$$D^2/abab^{-1}cc \approx D^2/abacbc \approx D^2/acbcab \approx D^2/aca^{-1}c^{-1}bb \approx T^2 \# P^2$$

□

Now we are ready to complete the classification theorem:

Theorem 3.1.14. *Any compact, connected surface M is homeomorphic to a surfaces of the type $S(m, n)$, where $n = 0$, 1 or 2. m and n are then uniquely determined by M.*

Proof. The existence follows from Theorem 3.1.10 and Lemma 3.1.13, so it only remains to establish the uniqueness of n and m. We can not give all the details here, but at least we can explain the invariants we use and the ingredients of the proof.

The invariants are *orientability* and the so–called *Euler characteristic.*

We have already introduced orientability, and it is obvious that homeomorphic surfaces are either both orientable or both no–orientable.

The Euler characteristic is an integral invariant which can be associated to any triangulated surface M. If the triangulation has s triangles, e edges and v vertices, then the Euler characteristic is defined by

$$\chi(M) = v - e + s\,.$$

The fundamental fact is that this number is independent of which triangulation we use, such that if M and N are homeomorphic, then $\chi(M) = \chi(N)$. This is the only fact we cannot prove here, in full generality, but it is a nice exercise to show that the Euler characteristic does not change if we subdivide the triangles in a given triangulation by introducing new vertices and edges. (Exercise 9. The missing step to the general result is then to show that we can go from one triangulation to any other by a finite number of such operations or their inverses.)

Using suitable triangulations of the identification models D^2/W it is not difficult to verify that

$$\chi(S(m, n)) = 2 - 2m - n$$

This means that if M is orientable, then $\chi(M) = 2 - 2m$. Hence m, and therefore also M, is determined by $\chi(M)$.

If M is non–orientable, then $n = 1$ or 2, and $\chi(M) = 1 - 2m$ or $-2m$. In the first case $\chi(M)$ is odd ($n = 1$), and in the second case it is even ($n = 2$). In both cases $\chi(M)$ determines both m and n. □

Remark 3.1.15. Note that we also have proved that M is completely determined by $\chi(M)$ and whether M is orientable or not.

In the orientable case the number m is called the *genus* of the surface. It also follows that the genus determines an orientable surface up to homeomorphism.

Exercises for 3.1

1. Show that if B is a Möbius band with boundary circle S, then

   $P^2 \approx B \cup_{S \approx S^1} D^2$

   $K^2 \approx B \cup_S B$

2. Show that $S(m,n)$ can be represented as $D^2/[a_1,b_1]\cdots[a_m,b_m]c_1^2\cdots c_n^2$, where $[a,b] = aba^{-1}b^{-1}$ and $c^2 = cc$.

3. In the identification space D^2/W the edges of D are identified pairwise with each other to curves on the surface. Draw these curves on $S(2,0)$ in fig. 3.1.5 if we use the identifications in exercise 2.

4. Verify the formula $\chi(S(m,n)) = 2 - 2m - n$.

   Determine the Euler characteristics of S^2, T^2, P^2 and K^2.

   Show that these are the only compact, connected surfaces having Euler characteristic ≥ 0.

5. Determine which compact surface can be represented as $D^2/abc^{-1}bdacd^{-1}$. What is its Euler characteristic? Is it orientable?

6. Show that if M_1 and M_2 are compact surfaces such that $M_1 \# M_2$ is homeomorphic to M_1, then M_2 must be homeomorphic to S^2.

7. Show that $\chi(M \# N) = \chi(M) + \chi(N) - 2$ for all compact surfaces M and N.

   A surface is called *irreducible* if the only way it can be written as a connected sum is if one of the summands is S^2. Prove that M is irreducible if and only if M is homeomorphic to T^2, P^2 or S^2

8. Generalize the notion of triangulation to allow the two–dimensional pieces to be arbitrary n–gons for all n. Show that the formula

   $$\chi(M) = v - e + s$$

still holds, where now s is the number of polygons.

9. Show that $\chi(M)$ does not change under subdivision of the triangulation. (A triangulation τ_2 is a subdivision of a triangulation τ_1 if every simplex of τ_2 is contained in a simplex of τ_1.)

Chapter 4

Geometry on surfaces — basic concepts

4.1 Introduction, local structure

Our goal is to generalize our geometrical studies from plane geometries to geometries on surfaces. As we already have pointed out, this requires more structure, so we begin by discussing what it means to put structure on a surface. The word "structure" is here used somewhat loosely, but the examples below will, hopefully, make it clearer what we mean.

Recall that a surface M is locally homeomorphic to $\mathbb{R}^2$. This means that M has an open covering $\{U_j\}_{j\in J}$ such that there are homeomorphisms $x_j : U_j \approx x_j(U_j) \subset \mathbb{R}^2$, where $x_j(U_j)$ is open in $\mathbb{R}^2$. Such a pair (U_j, x_j) is called a (local) *chart* on M and $\{(U_j, x_j)\}_{j\in J}$ is called an *atlas.*

If we have some kind of structure on (subspaces of) $\mathbb{R}^2$ (e. g. complex, Euclidean, hyperbolic, . . .), a natural thing to try is to transfer this structure to M via the local charts. Of course, this ought to be independent of which chart we choose, so the problem, in general, will be compatibility in the overlaps between two or more charts.

Example. The fundamental example, and also the model for any further discussion, is that of a *differentiable structure.* It is clearly of great interest to extend the concept of derivatives and differentiable functions to surfaces (or more general manifolds). Using the philosophy above, a function $f : M \to \mathbb{R}$ should be differentiable at a point $p \in M$ if fx_j^{-1} is differentiable at $x_j(p)$, where (U_j, x_j) is a chart with $p \in U_j$. But if (U_i, x_i) is another chart with $p \in U_i$, this should be the same as requiring that fx_i^{-1} is differentiable at

$x_i(p)$. These two conditions are equivalent for every function f if and only if $x_i x_j^{-1}$ and $x_j x_i^{-1}$ are differentiable at $x_j(p)$ and $x_i(p)$, respectively. Hence we have a well defined concept of "differentiable function on M" if and only if M has an atlas $\{(U_j, x_j)\}_{j \in J}$ such that the *coordinate transformations* $x_j x_i^{-1}$ are differentiable wherever they are defined (i. e. in $x_i(U_i \cap U_j)$) for all i and j. Such an atlas is called a *differentiable atlas*, and it defines a *differentiable structure* on M.

However, if studying differentiable functions is the motivation for introducing this structure, we should identify structures coming from atlases defining the same set of such functions. We call two atlases *equivalent* if they do, and define a differentiable structure to be an equivalence class of differentiable atlases. It is easy to see that two atlases $\{(U_i, x_i)\}$ and $\{(V_j, y_j)\}$ are equivalent if and only if

- all compositions $y_j x_i^{-1}$ and $x_i y_j^{-1}$ are differentiable.

But this condition is clearly equivalent to

- $\{(U_i, x_i)\} \cup \{(V_j, y_j)\}$ is also a differentiable atlas.

The set of all differentiable atlases (as sets of charts) is partially ordered by inclusions, and it follows that the union of all atlases in an equivalence class is the unique maximal element in that class. Therefore we can also identify a differentiable structure with a *maximal* differentiable atlas.

A surface M with a choice of such structure is called a *differentiable* surface.

Remark. There is nothing in this discussion that confines it to dimension two. It applies to manifolds of arbitrary dimensions — just replace '$\mathbb{R}^2$' by '$\mathbb{R}^n$' everywhere — and defines a *differentiable manifold.*

We may also require more: if, e. g. , all the coordinate transformations are r times continuously differentiable, we have a *$\mathcal{C}^r$–structure.* Here r may even be ∞, and a surface (manifold) with a $\mathcal{C}^\infty$–structure is called *smooth.*

Thus, if M has a differentiable structure, we can talk about differentiable functions on M. But we can also define *differentiable mappings* between such manifolds. If M and N are two manifolds with $\mathcal{C}^r$-structures $\{(U_i, x_i)\}$ and $\{(V_j, y_j)\}$ and $f : M \to N$ is a mapping, we say that f also is $\mathcal{C}^r$ if all the mappings $y_j f x_i^{-1}$ are r times continuously differentiable wherever they are defined.

This way we can transfer concepts and results having to do with differentiation to manifolds, provided they are equipped with differentiable structures. We can also define *real analytic* manifolds and maps, and replacing $\mathbb{R}^n$ with $\mathbb{C}^n$ in the definitions, we arrive at the concept of *complex* *analytic* structures.

If $f : M \to N$ is differentiable and also has a differentiable inverse, we say that f is a *diffeomorphism* and that M and N are *diffeomorphic.* Then they are, of course, also homeomorphic. One can show that all surfaces have smooth structures that are unique up to diffeomorphism. Hence there is no loss of generality when we from now on tacitly assume that all surfaces are smooth.

4.2 Geometric structure on surfaces

Using these ideas we will now attempt to define a *geometric structure* on a surface M to be a geometry on (subsets of) $\mathbb{R}^2$ together with a (differentiable) atlas on M such that all the coordinate transformations preserve this geometry. This will become more precise as we proceed, but first we look at some examples.

Example 1. If we consider the integers $\mathbb{Z}$ as a subgroup of the additive group $\mathbb{R}$, we can define the torus T^2 as the quotient group $\mathbb{R}^2/\mathbb{Z}^2$ (with the quotient topology). The easiest way to see this is to observe that under addition of elements of $\mathbb{Z}^2$, every element of R^2 can be moved to an element in $[0, 1] \times [0, 1]$, and the only identifications here are of the form $(t, 0) \sim (t, 1)$ or $(0, s) \sim (1, s)$. This means that the quotient is the disk $D = [0, 1] \times [0, 1]$ with boundary identifications according to the pattern $aba^{-1}b^{-1}$.

The quotient map $p : \mathbb{R}^2 \to T^2$ is a local homeomorphism, since

$$p|(a, b) \times (c, d) : (a, b) \times (c, d) \to p((a, b) \times (c, d))$$

is a homeomorphism whenever $b-a < 1$ and $d-c < 1$. If we use the inverses of these homeomorphisms as charts, the coordinate transformations will look like $(s, t) \to (s + m, t + n)$ for integers m and n. Here (m, n) will depend continuously on (s, t), and are therefore locally constant.

These coordinate transformations are restrictions of Euclidean isometries (congruences), and they preserve the Euclidean geometry — i. e. they take lines to lines and preserve lengths and angles. Therefore this atlas defines what we may call a *Euclidean structure on* T^2.

The images of straight lines in R^2 define a system of curves on T^2 which we again may call “lines” in the Euclidean structure. This set of lines will, however, not satisfy all of Hilbert’s axioms. For example, lines in $\mathbb{R}^2$ with rational slope will map to closed curves on T^2 — hence *betweenness* will have no meaning — while lines with irrational slope will map injectively to a topologically dense subset of T^2. Two distinct curves may intersect in more than one point, even infinitely many, and there will be infinitely many such curves going through two given points.

This means that *segments* are not determined by their endpoints, but we have to also specify the lines containing them. Likewise, *triangles* are not determined by the three vertices only. The best way to define such geometric concepts is to lift them to $\mathbb{R}^2$ and use the geometry there.

Angular measure between two lines, however, will have a unique meaning, since we get the same value, independent of which chart we use. We may also measure Euclidean distance along lines, but this will only be uniquely determined in small neighborhoods, since lines may be closed. But this makes it possible to compare two arbitrary segments, and we may for instance, as in Hilbert's axiom C1, find a congruent copy of any given segment starting in any point and going in any direction.

This last property is closely related to a property called "completeness", saying that we may prolong any segment as far as we want. For instance, if M is the complement of a point in T^2 (or R^2), then M will obviously inherit a Euclidean structure in the sense that we can find an atlas where all the coordinate transformations are restrictions of Euclidean isometries, but this structure will not be complete, since segments on lines punctured by the missing point can not be prolonged across it. We will essentially only be interested in complete structures.

If we replace $\mathbb{Z}^2$ by another additive subgroup $\Lambda \subset \mathbb{R}^2$ of rank 2, then $\mathbb{R}^2/\Lambda$ will also have a Euclidean structure. $\mathbb{R}^2/\Lambda$ will be homeomorphic (even diffeomorphic) to T^2, but it may not be possible to find a homeomorphism which preserves the structure — i. e. the Euclidean structures may not be equivalent for different Λ.

To obtain these structures we have used the additive group structure on R^2, but if we look closer, we see that what we really have used is only that $\mathbb{Z}^2$ (or Λ) is a group acting via structure preserving diffeomorphisms of the surface R^2. From this point of view we could replace $\mathbb{R}^2$ by any surface X with a structure ("geometry") and $\mathbb{Z}^2$ by a group Γ of diffeomorphisms of X preserving this structure ("congruences"), provided that the quotient space X/Γ is a surface. The next examples are of this type.

Example 2. Again we let $X = \mathbb{R}^2$ with Euclidean geometry, but now we let Γ be the group of Euclidean isometries (congruences) generated by $\gamma(s,t) = (s+1,t)$ and $\tau(s,t) = (-s+1,t+1)$. As with the torus, we can show that the quotient map $\mathbb{R}^2 \to \mathbb{R}^2/\Gamma$ is a homeomorphism when restricted to open rectangles with sides less than 1, and the inverses of these give an atlas which defines a Euclidean structure. To see which surface $\mathbb{R}^2/\Gamma$ is, we first observe that γ identifies all the vertical bands $[n, n+1) \times \mathbb{R}$ via horizontal translation. τ identifies all the horizontal bands $\mathbb{R} \times [n, n+1)$, but in a way such that every other band is reflected in the line $\{1/2\} \times \mathbb{R}$. Hence we also obtain $\mathbb{R}^2/\Gamma$ from the square $[0,1] \times [0,1]$, but now we identify the sides according to the pattern '$abab^{-1}$'. Thus the quotient is a Klein bottle.

Example 3. In this example, we use a different surface than $\mathbb{R}^2$ to model a geometry. We say that a surface has *spherical structure* if it is locally homeomorphic to the sphere S^2 in such a way that the coordinate transformations are restrictions of spherical congruences, i. e. multiplication by matrices in the orthogonal group $O(3)$. It turns out that there is only one example of such a geometry except S^2 itself; namely the projective plane P^2. But P^2 is, in fact, of the type X/Γ. In fact, let $X = S^2$, and let $\Gamma = \mathbb{Z}/2 = \{\pm I\} \subset O(3)$, which acts on S^2 via the antipodal map. Then the quotient is precisely P^2, and the restrictions of the quotient map to open subsets contained in hemispheres are homeomorphisms. If we use these to define an atlas, then the coordinate transformations are either identities or restrictions of the antipodal map. These are both orthogonal, hence we have defined a spherical structure.

As remarked when we discussed Hilbert's axiom system for plane geometries, the spherical geometry is also characterized by a system of "lines", namely the great circles in S^2 or the images of those in P^2 (which also become circles). This defines an incidence geometry on P^2 where any two lines will intersect in exactly one point. This geometry is complete.

Two observations before we go on:

(i) In all the examples so far we have described the surfaces by a representation D^2/W of the type we used to classify surfaces. Notice that in the concrete models used for D^2, it appears as a polygon bounded by lines in the relevant geometry, and the identifications given by the word W are given by elements in a group Γ.

(ii) The concrete models for D^2 are not uniquely determined — not even up to congruence. (For the torus we could, for example, have used a parallelogram with vertices (0,0), (1,1), (1,2) and (0,1).) But all choices will have the same *area.* Therefore it is meaningful to say that the quotient surface also has this area.

Example 4. It should now be clear how we should define what it should mean for a surface M to have a *hyperbolic* structure. Motivated by the examples above, we should look for examples of the form $\mathbb{H}/\Gamma$ or $\mathbb{D}/\Gamma$, where Γ is a suitable subgroup of $Möb(\mathbb{H})$ or $Möb(\mathbb{D})$. Such subgroups are not quite so easy to find, but this is not because there are few examples:

Theorem 4.2.1. *Except for T^2, S^2, P^2 and K^2, every surface has at least one hyperbolic structure. (In fact, infinitely many!)*

The simplest examples are, of course, $\mathbb{H}$ and $\mathbb{D}$ themselves, and if $f \in Möb^+(\mathbb{H})$ is of hyperbolic or parabolic type and $\langle f \rangle$ is the cyclic subgroup of $Möb^+(\mathbb{H})$ generated by f, then $H/\langle f \rangle$ will be a hyperbolic structure on $S^1 \times \mathbb{R}$. If, for instance, f is the parabolic transformation $f(z) = z + 1$, we get $H/\langle f \rangle$ by identifying the two vertical (infinite) sides of the strip $0 \leq \operatorname{Re} z \leq 1$.

We can also get non–orientable surfaces: an example is the (open) Möbius band, which can be realized as $\mathbb{H}/\langle g \rangle$, with $g(z) = -k\bar{z}$, with k a positive real number different from 1.

Much more interesting are complete hyperbolic surfaces having *finite area* in the sense discussed before.

The simplest such example is perhaps the following:

Consider the asymptotic quadrilateral in $\mathbb{D}$ with vertices in $\mathrm{e}^{\pi i/4}$, $\mathrm{e}^{3\pi i/4}$, $\mathrm{e}^{-3\pi i/4}$ and $\mathrm{e}^{-\pi i/4}$. (See fig. 4.2.1.)

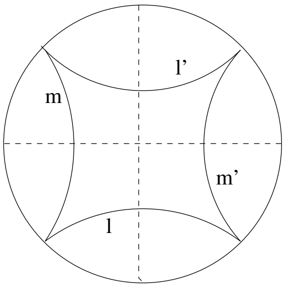


Fig. 4.2.1:

Let γ be the hyperbolic transformation having the real axis as axis and mapping m to m', and let τ be the hyperbolic transformation having the imaginary axis as axis and mapping l to l'. Let $\Gamma \subset Möb^+(\mathbb{D})$ be the subgroup generated by γ and τ. (Explicitly: $\gamma(z) = \dfrac{\sqrt{2}z + 1}{1z + \sqrt{2}}$ and $\tau(z) = \dfrac{\sqrt{2}z + i}{-iz + \sqrt{2}}$.) Then one can show that we can get $\mathbb{D}/\Gamma$ by identifying opposite sides in the asymptotic quadrilateral via γ and τ. The identification space is then clearly a torus minus one point. ("The point at infinity".)

This means that if p is a point on the torus T^2, then $T^2 - \{p\}$ has a hyperbolic structure of the form $\mathbb{D}/\Gamma$, with finite area equal to 2π. (The asymptotic quadrilateral is the union of two triply asymptotic triangles intersecting in one side.)

We can also use fig. 4.2.1 to describe another example: Let γ be a parabolic transformation which identifies l and m and fixes $e^{-3\pi i/4}$, and let τ be a parabolic transformation which identifies l' and m' and fixes $e^{\pi i/4}$. Then we get $\mathbb{D}/ < \gamma, \tau >$ by identifying l with m, and l' with m'. This surface is homeomorphic to a sphere were we have removed *three* points. (Alternatively $\mathbb{R}^2$ minus two points.) Hence this surface also has a hyperbolic structure of the form $\mathbb{D}/\Gamma$, with finite area equal to 2π.

Both of these examples are complete. In both examples the group Γ is a free group on two generators. (Isomorphic to the so–called *fundamental group* of the surface). But the surfaces are not homeomorphic, so the abstract group does not determine the surface up to homeomorphism.

The hyperbolic surfaces constructed so far have all been non–compact. The simplest compact surfaces admitting a hyperbolic structure are of genus two (i. e. they are homeomorphic to $T^2\#T^2$), and we now sketch how one can construct such examples.

From the proof of the classification theorem, we know that a surface of genus two can be obtained from an *octagon* where the edges are identified according to the pattern $a_1b_1a_1^{-1}b_1^{-1}a_2b_2a_2^{-1}b_2^{-1}$. We now need this to be a *hyperbolic* octagon, where the identifications are made via hyperbolic isometries. In particular, we need the edges that are identified (e. g. the edges labeled a_1 and a_1^{-1}) to be segments of the same length. Moreover, all the vertices are identified to one point and all angles are preserved, so the sum of the interior angles at all vertices must be 2π. But in the hyperbolic plane we may achieve this by a *regular* octagon. In the disk model this can be constructed as follows: Draw the radii (the rays) from $0 \in \mathbb{D}$ to the eight points $\{\mathrm{e}^{k\pi i/4}\}_{k=0,\ldots,7}$ in S^1. Every circle in $\mathbb{D}$ with center at 0 will intersect all these radii, and the eight points of intersection will be the vertices of a regular, hyperbolic octagon. When the (hyperbolic) radius of the circle increases from 0 to ∞, the interior angle at the vertices will decrease from $3\pi/4$ (as in the Euclidean case) to 0 (with all the vertices at ∞). Somewhere in between the angle will be $\pi/4$, which is what we need. (We can, of course, also use the hyperbolic trigonometric relations to calculate exactly for which radius this will happen.)

We now use this octagon as our disk D^2. Then there are unique elements $\gamma_i \in M\ddot{o}b^+(\mathbb{H})$ mapping a_i to a_i^{-1}, $i = 1, 2$, and $\tau_i \in M\ddot{o}b^+(\mathbb{H})$ mapping b_i

to b_i^{-1}, $i = 1, 2$, with the right orientations. If we identify the edges of D^2 by these congruences, the identification space becomes a surface of genus two with a hyperbolic structure. This surface also has the form $\mathbb{D}/\Gamma$, where now Γ is generated by the elements $\gamma_1,\ \gamma_2,\ \tau_1$ and τ_2.

The area of this surface is equal to the area of the regular octagon D^2, i. e. 4π.

This hyperbolic structure is also complete, and the "lines" are images of $\mathbb{D}$-lines under the quotient map $\mathbb{D} \to \mathbb{D}/\Gamma$.

Remark 4.2.2. From the formula for the area of hyperbolic triangles it is easy to see that an n–gon with peripheral angle sum 2π has area $(n-4)\pi$. It follows that the area of a compact hyperbolic surface of genus g constructed as above from a $4g$–gon will have area $4\pi(g-1)$, independent of lengths of the sides. This will later be a consequence of a much more general result — the *Gauss–Bonnet theorem.*

After all these examples, we are now ready for a more formal definition: Let (X, G) be one of the pairs

- $(\mathbb{H}, \textit{Möb}(\mathbb{H}))$ (or $(\mathbb{D}, \textit{Möb}(\mathbb{D}))$,
- $(\mathbb{R}^2,\ E(2))$, where $E(2)$ is the group of Euclidean isometries,
- $(S^2,\ O(3))$, where $O(3)$ is the group of spherical isometries (orthogonal matrices).

We say that a surface has a *geometric structure modeled on* (X, G) (or shorter: a hyperbolic, Euclidean or spherical structure, respectively) if it has an atlas $\{(U_i,\ x_i : U_i \to x_i(U_i) \subset X)\}_i$ such that all the coordinate transformations are restrictions of elements of G.

One can show that a complete, connected surface with such a structure must have the form X/Γ for some subgroup $\Gamma \subset G$. The classification of geometric structures on surfaces then becomes the classification of subgroups occurring this way. The richest and most interesting such theory is in the hyperbolic case, were it leads to the theory of *Fuchsian groups.*

The idea of characterizing geometries using group actions goes back to Felix Klein and Sophus Lie, and it was presented by Klein in the famous *Erlangen program* in 1872. The groups G are examples of *Lie groups.* Note that in all the three cases above G acts *transitively* on X, i. e. for any two points $x, y \in X$ we can find a $g \in G$ such that $y = gx$. Moreover, the *stabilizer subgroup*

$$G_x = \{g \in G | gx = x\}$$

is isomorphic to the orthogonal group $O(2)$. Then, for any $x \in X$, the map

$$g \mapsto gx : G \to X$$

induces a bijection (even a homeomorphism) between X and the left cosets G/G_x, with G–action corresponding to left multiplication. This way we obtain identifications

$$\begin{aligned}
&\mathbb{R}^2 \approx E(2)/O(2),\\
&S^2 \approx O(3)/O(2),\\
&\mathbb{H} \approx SL_2(\mathbb{R})/SO(2).
\end{aligned}$$

Chapter 5

Differential geometry on surfaces

We are now going to broaden our perspective and generalize dramatically the idea of geometric structures on surfaces. Until now we have restricted ourselves to the classical geometries or structures built from them. But the geometries of most surfaces in space are not of this nature — still they are very natural objects for geometric studies, and we would like to develop a theory which also includes these. This will require completely new ideas, exploiting methods from calculus and differential equations — whence the term *differential geometry.* The structures built from the classical geometries are very special cases, and one of our main results will be a characterization of these geometries inside differential geometry.

The geometric structures we defined earlier were *homogeneous*; they look "the same" at every point in the sense that neighborhoods of any two points can be mapped to each other by local congruences, as, for instance, on a perfectly round sphere in space. However, most surfaces are not like this. On a two–dimensional graph a maximum point and a saddle point look very different, and whatever we may mean by "geometric structure" should capture this difference.

In this chapter we will introduce a notion of geometric structure which is allowed to vary continuously over the surface. This notion will not be based on lines and incidence, even locally, but rather on infinitesimal versions of distance and angles, encoded in a *Riemannian metric.* Maps preserving this metric — *local isometries* — are then analogues of local congruences in the classical geometries, but we do not require such maps to exist between neighborhoods of arbitrary points.

Although not part of the initial structure, analogues of “lines” in the classical geometries will appear eventually, but now they will be constructed as curves satisfying local properties that can be used to characterize lines in the classical geometries (*geodesic curves*).

The history of differential geometry on surfaces is long and fascinating. The literature is enormous, and we can only begin to explore it in these notes. It started with studies of special surfaces in Euclidean 3–space and only gradually developed into a general theory. Central characters in the story were Euler, Monge and, above all, Gauss. Gauss was the first to introduce the *intrinsic* point of view, realizing that many aspects of geometry could be understood from metric properties of the surfaces itself, without reference to the surrounding space. This paved the way for Riemann, who generalized the theory to abstract manifolds of any dimension. In honour of him the abstract theory is also called *Riemannian geometry.*

Our treatment will be based on the abstract point of view, but using geometry on surfaces in space as motivation. After covering the most fundamental ideas, we will concentrate on two major results: characterization of the classical geometries and the Gauss–Bonnet theorem.

But before we can venture into this new terrain, we need some more background material on differentiable surfaces and maps. Everything will be formulated for *smooth* (i. e. $\mathcal{C}^\infty$) surfaces. We will need to differentiate functions several times, and then it is convenient to agree once and for all that all surfaces and maps are smooth.[1]

The fundamentals on tangent planes and derivatives are covered in section 1, followed by a discussion of orientability in section 2. Then we are ready for the definition of the fundamental concepts of Riemannian metrics and local isometries, which are found in sections 3 and 4. Section 5 treats the most important tool we have for measuring the variation of a metric over a surface: *Gaussian curvature.* The promised reappearance of distinguished lines, *geodesics,* follows in section 6. They are used to construct particularly well behaved local coordinates (section 7), which, in turn, are used in section 8 to characterize the classical geometries as those with constant curvature. The last section deals with the famous *Gauss–Bonnet theorem*, which is a powerful formula bringing together topological information (Euler characteristic) and curvature.

A remark on terminology: we have defined differentiable structures and maps using local charts $x : \mathcal{V} \to \mathbb{R}^2$, where the $\mathcal{V}$'s are open subsets of the

[1] A careful analysis will reveal that as long as we just study surfaces in $\mathbb{R}^3$, $\mathcal{C}^2$ will do. The extension to abstract surfaces via Gauss' *Theorema egregium* requires $\mathcal{C}^3$.

surface S. It will here be convenient to replace charts with their inverses $x^{-1} : x(\mathcal{V}) \to S$, which we will call (local) *parametrizations.* Thus, a surface can also be defined as a space S such that for every $p \in S$ there is a map $x : \mathcal{U} \to S$ from an open set $\mathcal{U} \subset \mathbb{R}^2$, such that $p \in x(\mathcal{U})$ and $x : \mathcal{U} \to x(\mathcal{U})$ is a homeomorphism. $(x, \mathcal{U})$ is also called a *coordinate patch.* The *coordinate transformation* are now the maps $y^{-1}x : U \cap x^{-1}y(V) \to V \cap y^{-1}x(U)$, defined for coordinate patches $(x, \mathcal{U})$ and $(y, \mathcal{V})$.

We will here also use the word *atlas* for a collection of coordinate patches $\{(x_i, \mathcal{U}_i)\}$ such that $\cup_i \mathcal{U}_i = S$. A differentiable structure on S is then a maximal atlas such the all coordinate transformations are smooth.

We emphasize that this is only a change of language; one can easily pass between charts and parametrizations by taking inverses.

5.1 Tangent planes and derivatives of maps

We have already defined what it means for a map between smooth surfaces to be smooth, but we also need to define what we mean by the *derivatives* of such maps. The idea is that the derivative of a map f should be a linear approximation to f, so we first have to define vector spaces which approximate the surface: the *tangent planes.* This can be done in several different ways, but here our point of view is that tangent vectors should be tangent vectors of *curves* on the surface. To illustrate the idea, let q be a point in the plane $\mathbb{R}^2$. Every smooth curve through q has a tangent vector there, and the set of all possible tangent vectors of curves through q can naturally be identified with the vector space $\mathbb{R}^2$ itself. But each vector can be realized as the tangent vector of infinitely many curves through q. Hence, if we define an equivalence relation on the set of smooth curves through q by declaring two curves to be equivalent if they have the same tangent vector at this point, we can identify the vector space $\mathbb{R}^2$ with the set of equivalence classes. This is a formulation that is easy to generalize to an arbitrary smooth surface.

Let S be a smooth surface, and let p be a point on S. By a *curve at p* we shall mean a smooth map $\omega : J \to S$, where J is an open interval containing 0, such that $\omega(0) = p$. Let $\Omega_p = \Omega_p(S)$ be the set of all such curves at p.

We now choose a local parametrization $x : \mathcal{U} \to S$ around p. Then $x^{-1}\omega$ is a curve at $x^{-1}(p) \in \mathbb{R}^2$, and $(x^{-1}\omega)'(0)$ is a vector in $\mathbb{R}^2$. Define the equivalence relation $\sim$ on Ω_p by

$$\omega \sim \tau \iff (x^{-1} \circ \omega)'(0) = (x^{-1} \circ \tau)'(0).$$

(Cf. Fig. 5.1.1. Note that ω and τ need not be defined on the same interval J for this to make sense. It is also not necessary that the image of the whole curve is contained in $x(\mathcal{U})$.

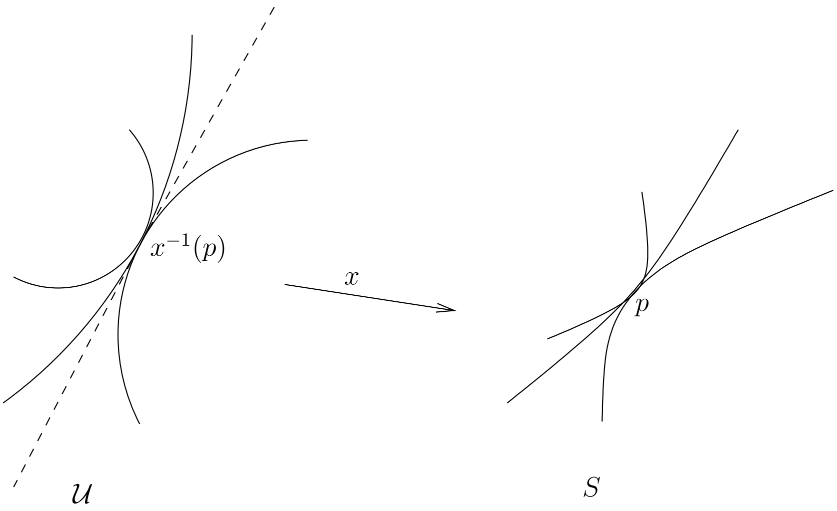


Fig. 5.1.1: Curves with the same tangent.

Lemma 5.1.1. *(1) The equivalence relation $\sim$ does not depend on the choice of local parametrization x.*
Define T_pS to be the set of equivalence classes.

(2) $\omega \mapsto (x^{-1}\omega)'(0)$ induces a bijection $T_pS \approx \mathbb{R}^2$.

(3) The bijection in (2) defines a vector space structure on T_pS which is independent of x.

Proof. (1): If $y : \mathcal{V} \to S$ is another parametrization, then

$$(y^{-1}\omega)'(0) = (y^{-1}xx^{-1}\omega)'(0) = J(y^{-1}x)_{x^{-1}(p)}(x\omega)'(0)\,,$$

where $J(y^{-1}x)_{x^{-1}(p)}$ is the Jacobian matrix of $y^{-1}x$ at the point $x^{-1}(p)$. But $J(y^{-1}x)_{x^{-1}(p)}$ is invertible, since $y^{-1}x$ is a diffeomorphism near $x^{-1}(p)$.

(2) Injectivity follows immediately from the definition of $\sim$.

Surjectivity: If $v \in \mathbb{R}^2$, let $\omega(t) = x(x^{-1}(p) + tv)$. Then $\omega(t)$ is defined for t near 0, and $(x^{-1}\omega)'(0) = v$.

(3): Let $\alpha_x : T_pS \to \mathbb{R}^2$ be the bijection in (2), and let α_y be the analogous bijection obtained using the parametrization y. Then

$$\alpha_y \circ \alpha_x^{-1}(v) = y^{-1}(x(x^{-1}(p) + tv))'(0) = J(y^{-1}x)_{x^{-1}(p)}\, v.$$

But multiplication by $J(y^{-1}x)_{x^{-1}(p)}$ is an isomorphism of vector spaces.

□

Definition 5.1.2. *T_pS with the vector space structure given by Lemma 5.1.1 is called the* tangent plane of S at p.

This definition is quite abstract, and valid for an arbitrary surface. However, for surfaces in $\mathbb{R}^3$, we can identify T_pS with something which is easier to visualize.

Definition 5.1.3. *Let S be a subset of $\mathbb{R}^3$ which is also a smooth surface, and let $\iota : S \subset \mathbb{R}^3$ be the inclusion map. S is called a* regular surface *if for every local parametrization x of S the Jacobian of the composition $\iota \circ x$ has rank 2 at every point.*

From now on, when we write $S \subset \mathbb{R}^3$, we will always assume that S is regular.

One can show that S is regular if and only if it locally (around every point) has the form $\{(x, y, z) \in \mathbb{R}^3 : F(x, y, z) = 0\}$ for some smooth function F with $\nabla F \neq 0$ on S. Spheres are examples of surfaces defined this way. Other important examples include *graphs* of smooth functions $f(u, v)$ defined on open subsets $\mathcal{O} \subset R^2$, parametrized by $x(u, v) = (u, v, f(u, v))$. In fact, we shall see below that locally every regular surface can be identified with a graph, possibly rotated and translated in space.

If S is regular, we can identify T_pS with the plane in $\mathbb{R}^3$ which contains p and all the tangent lines to curves on S through p. This plane can be further identified with a linear subspace of $\mathbb{R}^3$ via the translation taking p to $0 \in \mathbb{R}^3$. Thus we may naturally identify every T_pS with such a subspace, and we will always assume that we have done so. The vector space structure given by Lemma 5.1.1 makes this identification an isomorphism. Note that if S is given by an equation $F(x, y, z) = 0$, the tangent space is determined by its *normal vector* $\nabla F(x, y, z)$.

Notation. Let $\omega'(0)$ denote the equivalence class of ω in T_pS. This is the *tangent vector* of ω in p. More generally, we let $\omega'(t)$ denote $\omega'_t(0)$, where $\omega_t(s) = \omega(t+s)$ (where this is defined). This means that we take the tangent vector of ω in $\omega(t)$ instead of $\omega(0)$. Note that if $S \subset \mathbb{R}^3$, we may think of ω as a curve in $\mathbb{R}^3$, and $\omega'(t)$ is then the tangent vector of this curve in $\mathbb{R}^3$.

A parametrization $x : \mathcal{U} \to S$ gives rise to a natural basis for T_pS for every $p \in x(\mathcal{U})$. Let (u, v) be the coordinates in $\mathcal{U} \subset \mathbb{R}^2$. If we fix v and vary u, we get a curve $u \mapsto \beta(u) = x(u, v)$. The tangent vector $\beta'(u)$ of this curve we denote $x_u(u, v)$. Fixing u and varying v we similarly get tangent vectors $x_v(u, v)$. Via the identifications $T_pS \approx \mathbb{R}^2$, these vectors correspond to the standard basis vectors $(1, 0)$ and $(0, 1)$ in $\mathbb{R}^2$, and therefore they form a basis for $T_{x(u,v)}S$.

The curves $u \mapsto x(u, v)$ and $v \mapsto x(u, v)$ are called *coordinate curves.* x_u and x_v are the tangent vectors of these curves, and they may be thought of as the partial derivatives of x with respect to u and v. If $S \subset \mathbb{R}^3$, we may also think of them as the partial derivatives of x considered as a map $x : \mathcal{U} \to \mathbb{R}^3$. In fact, we see that S is regular if and only if x_u and x_v are always linearly independent as vectors in $\mathbb{R}^3$, or if the cross product $x_u \times x_v$ is non–zero everywhere. The cross product is then a normal vector to the tangent plane.

Remark 5.1.4. Another common notation for x_u and x_v is $\frac{\partial}{\partial u}$ and $\frac{\partial}{\partial v}$. This notation refers to a different interpretation of tangent vectors and will not be used here.

Derivatives of smooth maps. Let $f : S \to S'$ be smooth, let p be a point of S, and $q = f(p)$.

If $\omega \in \Omega_p(S)$, then $f \circ \omega \in \Omega_q(S')$.

Lemma 5.1.5. *$(f\omega)'(0)$ depends only on $\omega'(0)$.*

Proof. Let $x : \mathcal{U} \to S$ and $y : \mathcal{V} \to S'$ be local parametrizations around p and q, respectively.

Suppose that $\omega'(0) = \tau'(0)$ — i. e. $(x^{-1}\omega)'(0) = (x^{-1}\tau)'(0)$. Then

$$\begin{aligned}(y^{-1}f\omega)'(0) &= ((y^{-1}fx)(x^{-1}\omega))'(0) = J(y^{-1}fx)_{x^{-1}(p)}(x^{-1}\omega)'(0)\\ &= J(y^{-1}fx)_{x^{-1}(p)}(x^{-1}\tau)'(0) = (y^{-1}f\tau)'(0).\end{aligned}$$

□

Definition 5.1.6. *The* derivative of f at p *is the map* $df_p : T_pS \to T_qS'$ *defined by* $df_p(\omega'(0)) = (f\omega)'(0)$. *This is well–defined by Lemma 5.1.5.*

Examples 5.1.7. If U is an open subset of $\mathbb{R}^2$, we have a natural identification between $\mathbb{R}^2$ and T_pU for every $p \in U$. This identification will often be understood in the following. (See also exercise 1.) Using this, the bijection $T_pS \approx \mathbb{R}^2$ of Lemma 5.1.1 is nothing but $d(x^{-1})_p = (dx_{x^{-1}(p)})^{-1}$.

If U' is another such subset and $f : U \to U'$ is smooth, then df_p is multiplication by the Jacobi matrix of f at p, which is a linear transformation $\mathbb{R}^2 \to \mathbb{R}^2$.

More generally, we have:

Lemma 5.1.8. *Let $f : S \to S'$ be a smooth map between surfaces. Then*

(1) df_p is a linear transformation for every $p \in S$.

(2) (Chain rule) If $g : S' \to S''$ is another smooth map and $q = f(p)$, Then $d(g \circ f)_p = dg_q \circ df_p$.

Proof. We first prove (2):
$d(gf)_p(\omega'(0)) = (gf\omega)'(0) = dg_q((f\omega)'(0)) = dg_q(df_p(\omega'(0))$.

(1): Using (2) we can write $df = d(y(y^{-1}fx)x^{-1}) = dy \circ d(y^{-1}fx) \circ d(x^{-1})$, which is a composition of three linear transformations by Example 5.1.7. □

Writing vectors in $\mathbb{R}^2$ as (a, b), we see by Example 5.1.7 that $x_u = dx(1, 0)$ and $x_v = dx(0, 1)$. We can use this to express a tangent vector $\omega'(t)$ of a curve ω in the basis $\{x_u, x_v\}$. Suppose ω has image in $x(\mathcal{U})$, such that we can write $\omega(t) = x(u(t), v(t))$, for uniquely determined functions $u(t)$ and $v(t)$. Then $\tau(t) = (u(t), v(t)) = x^{-1}\omega(t)$ defines a curve in $\mathcal{U}$, and $\tau'(t) = (u'(t), v'(t)) = u'(t)(1, 0) + v'(t)(0, 1)$. Therefore we get

$$\omega'(t) = dx(\tau'(t)) = u'(t)\, dx(1, 0) + v'(t)\, dx(0, 1) = u'(t)\, x_u + v'(t)\, x_v\,.$$

It is also easy to see that if $x : \mathcal{U} \to S$ and $y : \mathcal{V} \to S'$ are local parametrizations and $f : S \to S'$ is smooth, then the Jacobian matrix $J(y{-}1fx)$ is the matrix of df with respect to the bases (x_u, x_v) and (y_u, y_v). (Exercise 3.)

The inverse function theorem easily generalizes to say that if df_p is nonsingular (has rank 2), then f is a diffeomorphism of a neighborhood of p onto its image. As an application, let us show that locally every regular surface can be thought of as the graph of a function.

Let $H \subset \mathbb{R}^3$ be a plane with a unit normal vector N and let p be a point in H. If $f : \mathcal{V} \to \mathbb{R}$ is a function defined on an open subset of H, we define the *graph* of f to be the subset

$$S_f = \{q + f(q)N \mid q \in \mathcal{V}\} \subset \mathbb{R}^3.$$

This is an obvious generalization of the standard definition of a graph, and S_f is clearly a regular surface if and only if f is smooth.

Proposition 5.1.9. *A regular surface $S \subset \mathbb{R}^3$ coincides with a graph of a smooth function in a neighborhood of every point.*

Proof. Suppose first that the regular surface S contains the origin $0 \in \mathbb{R}^3$ and consider S near 0. Let $\pi : S \to T_0S$ be the restriction to S of the orthogonal projection to T_0S. Then it is easy to verify that $d\pi_0$ is the identity map. (Strictly speaking we here identify the tangent plane of T_0S at 0 with T_0S itself. See exercise 1.) Therefore π has a local inverse $\gamma : \mathcal{W} \to \gamma(\mathcal{W}) \subset S$, where $\mathcal{W}$ is a neighborhood of 0 in T_0S, and γ can be written $\gamma(w) = w + h(w)N$, where N is a normal vector to T_0 and $h : \mathcal{W} \to \mathbb{R}$ is a smooth function. But this means that $\gamma(\mathcal{W})$ is identified with the graph of the function h.

If $p \in S$ is an arbitrary point, we can clearly reduce to the case $p = 0$ by a translation in space. □

Remark 5.1.10. For later use we observe the following two easy facts about the function h constructed in the preceding proof:

(i) Let $L(p)$ be the line spanned by the surface normal $N(p)$ at p. Then $h = \pi_L \circ \gamma$, where $\pi_L : S \to L(p)$ is the restriction to S of the orthogonal projection to $L(p)$.

(ii) The point $0 \in W$ is a critical point for h and (equivalently) p is a critical point for π_L.

Exercises for 5.1

1. Let V be a vector space of dimension n. Verify that V has a smooth structure such that taking coordinates with respect to any basis defines a diffeomorphism between V and $\mathbb{R}^n$. We will always assume that vector spaces have this structure.

   Show that there is a natural isomorphism of vector spaces $T_vV \cong V$ for every $v \in V$.

   Using this isomorphism, show that if $L : V \to V$ is linear, then $dL_v = L$ for every $v \in V$.

   (From now we shall usually make these identifications without further comments.)

2. Consider the set of complex numbers $\mathbb{C}$ with its natural smooth structure (Exercise 1.) Show that the tangent planes $T_z\mathbb{C}$ have natural structures as complex (one–dimensional) vector spaces.

If $f(z)$ is an analytic function on an open subset of $\mathbb{C}$, show that f is smooth. Finally, show that, via the identifications $T_z\mathbb{C} \cong \mathbb{C}$ in exercise 1, df_z is (complex) multiplication by $f'(z)$.

3. Let $x : \mathcal{U} \to S$ and $y : \mathcal{V} \to S'$ be local parametrizations and $f : S \to S'$ a smooth map. Verify that the Jacobian matrix $J(y^{-1}fx)$ is the matrix of df with respect to the bases (x_u, x_v) and (y_u, y_v).

4. Let $\alpha(u) = (f(u), g(u))$, $u \in [a, b]$ be an embedded curve in the xz–plane such that $f(u) > 0$ for all u. Find a parametrization of the *surface of rotation* obtained by rotating α around the z–axis and find conditions for this to be a regular surface.

   Discuss what happens if we remove the condition $f(u) > 0$.

5. Let α and δ be curves in R^3 defined on the same interval. Define $x(u, v) = \alpha(u) + v\delta(u)$. If this is a parametrization of a surface S, we call S a *ruled* surface — S is then a union of lines going through points in $\operatorname{im}\alpha$ and with directions given by δ. Discuss general conditions for S to be a regular surface, and apply this to the cases

   a) α constant ("generalized cones").

   b) δ constant ("generalized cylinders").

   c) $\delta(t) = \alpha'(t)$ (The "tangent developable" of α.)

6. Find a parametrization of a Möbius band and a torus realized as regular surfaces in $\mathbb{R}^3$.

7. Consider the proof given that any regular surface locally is a graph of a smooth function. Discuss what happens when we replace the tangent plane with an arbitrary plane.

8. Prove the assertions in Remark 5.1.10.

   More generally, let $\pi_L :\to L$ be orthogonal projection to a line L spanned by a vector V and write $\pi_L(z) = g(z)V$. As a converse to (ii), prove that at a critical point q for g, the surface normal also spans L.

   The next two exercises discuss an extension of the definitions of smooth maps and derivatives needed in Section 9.

9. Let $g : I \to S$ be a map from a not necessarily open interval to a surface S. We say that g is smooth if we can find a smooth map $G : J \to S$ on an open interval containing I, such that $g = G|I$. Show that if $a \in I$ is an endpoint of I, then $G'(a)$ is independent of G.

This justifies the notation $g'(a) = G'(a)$.

10. Let $\triangle$ be a triangle in $\mathbb{R}^2$. In analogy with exercise 9.we say that a map $f : \triangle \to S$ is smooth if it can be extended to a smooth map G defined on an open neighborhood of $\triangle$ in $\mathbb{R}^2$.

    Show that if p is a point on the boundary of $\triangle$, the derivative dF_p only depends on f.

    For this reason we will write $df_p = dF_p$.

5.2 Orientation

When we discussed the topological classification of compact surfaces, we needed to distinguish between *orientable* and *non–orientable* surfaces. We will now show how these concepts can be defined more precisely for smooth surfaces, in terms of properties of smooth atlases.

An orientation of a *vector space* is determined by an ordered basis, and two ordered bases define the same orientation if and only if the transition matrix between them has positive determinant. This divides the set of ordered bases into two equivalence classes, and an orientation is the same as a choice of one of them. A basis in the chosen equivalence class will be called "positively oriented".

In dimension 2, for example, we see that if (e_1, e_2) is an ordered basis representing one orientation, then $(e_1, -e_2)$ represents the other ("opposite") orientation. Hence, if we think of e_1 as pointing "ahead" and e_2 "left", then $-e_2$ points "right", and changing orientation amounts to interchanging the notions of left and right. Similarly, we can change the orientation by replacing (e_1, e_2) by (e_2, e_1).

Another way of thinking about orientation in dimension 2 is as a preferred choice of one of the two possible senses of rotation around a point. ("Rotation in the direction from e_1 to e_2.)

We will now concentrate on dimension two, though suitably generalized most of the following will also apply in higher dimensions.

If S is a surface, each tangent plane T_pS has two possible orientations. If $x : \mathcal{U} \to S$ is a local parametrization, then the ordered bases (x_u, x_v) determine one of these orientations for every $p \in x(\mathcal{U})$. This is what we will mean by a *continuous* choice of orientations on $x(U)$. If $y : \mathcal{V} \to S$ is another local parametrization, (y_u, y_v) will determine the same orientations on $x(\mathcal{U}) \cap y(\mathcal{V})$ if and only if the Jacobian $J(y^{-1}x)$ has positive determinant (where it is defined). Hence a precise definition of orientability will be:

Definition 5.2.1. *The surface S is* orientable *if it admits a differentiable atlas such that the Jacobians of all the coordinate transformations have positive determinant everywhere. An* orientation of S *is a choice of a maximal such atlas.*

Note that we obtain an atlas defining the opposite orientation by reversing the order of the parameters (u, v) everywhere.

For regular surfaces we have the following convenient characterization of orientability:

Proposition 5.2.2. *If $S \subset \mathbb{R}^3$ is a regular surface, S is orientable if and only if it has a smooth normal vector field, i. e. if and only if we can choose a unit normal vector $N_p \in \mathbb{R}^3$ to every tangent plane T_pS such that the resulting map $N : S \to S^2$ is smooth.*

Proof. Given such a normal vector field, we can use the “right–hand rule” to give an ordering (v_1, v_2) of any basis for T_pS by requiring that (v_1, v_2, N_p) is a right–hand system of vectors in $\mathbb{R}^3$, that is: if the (3×3) matrix $[v_1^t, v_2^t, N_p^t]$ has positive determinant.

If $x : \mathcal{U} \to S$ is a parametrization and (x_u, x_v) is the “wrong” ordering, we can correct that by interchanging u and v. Doing this with (the inverses of) all charts in an arbitrary atlas produces an orientable atlas. (We will refer to this orientation as *the orientation determined by the normal N*).

Conversely, given an orientable, smooth atlas, we can use the vector product in $\mathbb{R}^3$ and define the normal vector field by

$$N = \frac{x_u \times x_v}{\|x_u \times x_v\|} .$$

□

Here is a sketch of how to relate this definition of orientability to the previous one, which was based on the notions of one–sided or two–sided curves:

Let $\alpha : [0, 1] \to S$ be a smooth, closed curve (i. e. $\alpha(1) = \alpha(0)$, and assume that $\alpha'(t) \neq 0$ for all t — i. e. α is *regular*. Locally, for small variations of the parameter t, we can distinguish between ”the two sides” of the curve, and the issue is whether or not it is possible to choose one side continuously along the whole curve. More precisely: can we find a continuous family of tangent vectors $v(t) \in T_{\alpha(t)}S$ such that $\alpha'(t)$ and $v(t)$ are linearly independent for every t and $(\alpha'(0), v(0))$ and $(\alpha'(1), v(1))$ define the same orientation of $T_{\alpha(0)}S = T_{\alpha(1)}S$?

First we reduce to an infinitesimal situation, using the observation that a vector $v \in T_{\alpha(s)}S$ which is not a multiple of $\alpha'(s)$, points to exactly one of the two sides of the curve (in the local picture), and the sides can be distinguished by the orientation classes of the ordered bases $(\alpha'(t), v)$ for $T_{\alpha(s)}S$. Thus the problem of a continuous choice of sides along the whole curve, becomes the problem of a continuous choice of orientations of $T_{\alpha(s)}S$ for all s. If this is possible, we call α *orientation preserving.* This definition can easily be extended to closed curves which are only *piecewise regular*, i. e. regular except at a finite number of points.

The result is:

Proposition 5.2.3. *A surface S is orientable if and only if every closed, piecewise regular curve is orientation preserving.*

Proof. One way is easy: if S is orientable, we can just take the restriction of any orientation on S to the curve.

To prove the converse, assume that every closed, regular curve is orientation preserving. We want to construct an orientable, smooth atlas on S. It is clearly enough to treat each connected component separately, so we assume that S is connected.

Start by choosing some local parametrization $x_0 : \mathcal{U}_0 \to S$ with $\mathcal{U}_0$ connected, and a basepoint $p_0 \in x_0(\mathcal{U}_0)$. Let p be another point in S. Since S is connected, we can find a regular path β such that $\beta(0) = p_0$ and $\beta(1) = p$. Every point $\beta(t)$ lies in the image of a local parametrization, and, using compactness of the interval $[0.1]$, we can find a partition $t_0 = 0 < t_1 < \cdots < t_n = 1$ and local parametrizations $x_i : \mathcal{U}_i \to S,\ i = 1, \ldots, n$, with all $\mathcal{U}_i$ connected, such that $\beta([t_{i-1}, t_i]) \subset x_i(\mathcal{U}_i)$ for all i. By inductively interchanging the parameters of x_i, if needed, we may assume that $\det d(x_i^{-1} x_{i-1})_u > 0$ for $i = 1, \ldots, n$ and all $u \in \mathcal{U}_{i-1}$ such that $x_i^{-1} x_{i-1}(u)$ is defined. In particular, this means that we have a continuous choice of orientations along β, extending the orientation defined by x_0 at p_0. Doing this for every point $p \in S$, we clearly obtain an atlas for the smooth structure on S. The claim is that this atlas is orientable.

If not, we would have $\det d(y^{-1}x)_u < 0$ for some local parametrizations $x : \mathcal{U} \to S$ and $y : \mathcal{V} \to S$ obtained as above, and $u \in \mathcal{U} \cap x^{-1}y(\mathcal{V})$. But, by construction, $x(\mathcal{U})$ contains a point p connected to p_0 by a curve β with a continuous choice of orientations extending a fixed orientation at p_0, and $y(\mathcal{V})$ contains a point q connected to p_0 by a curve γ with the same property.. But p and q can be connected to $x(u)$ by curves β_1 and γ_1 contained in $x(\mathcal{U})$ and $y(\mathcal{V})$, respectively.

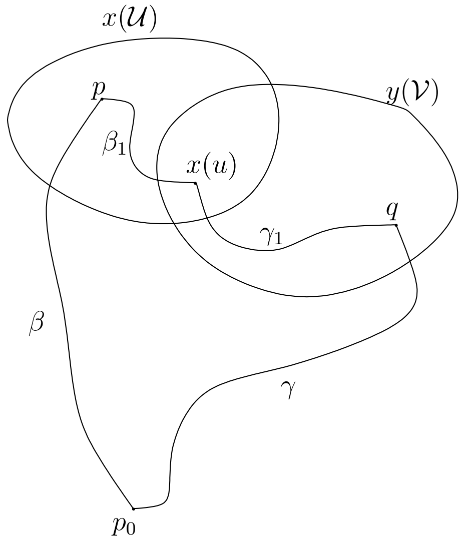


Fig. 5.2.1:

The composition of these four curves is a piecewise regular closed curve. But it is not orientation preserving, since prolonging an orientation from $x(u)$ around the curve leads to the opposite orientation when we come back to the same point.

□

Let $f : S \to S'$ be a local diffeomorphism between oriented surfaces. We say that f is *orientation preserving* if df_p maps positively oriented bases to positively oriented bases. Equivalently: the Jacobian matrix $J(y^{-1}fx)$ has positive determinant wherever it is defined, for all local parametrizations x and y in orientable atlases for S and S', respectively. In particular, the parametrizations (or charts) in orientable atlases are themselves orientation preserving. Clearly, compositions and inverses of orientation preserving diffeomorphisms are again orientation preserving.

Exercises for 5.2

1. Show that on a connected, orientable surface an orientation is completely determined by a choice of orientation of the tangent space at one point. Hence the surface has exactly two orientations.

2. Let $f : S \to S$ be a diffeomorphism of an orientable surface. Show that the question of whether f is orientation preserving or not (with respect to the same orientation on both copies of S), does not depend on which orientation we choose for S.

3. Show that the orientation preserving hyperbolic congruences in $\mathbb{H}$ are precisely the elements in $\textit{Möb}^+(\mathbb{H})$. What are the analogous results for the other classical geometries?

4. Show that P^2 and the Möbius band are non–orientable.

5.3 Riemannian surfaces

We are now ready to introduce the most general notion of a *geometric structure* on a smooth surface, or rather an infinitesimal version of it. In our earlier treatment of geometry (e. g. Euclidean and hyperbolic geometry), geometric structure was defined in terms of the system of subsets we called "lines". Next we tried to extend this to surfaces by first defining the geometric structure locally, and then patching pieces together in such a way that a global system of lines could be defined by prolongations of local lines. The system quickly becomes very complicated, but it is important to observe that the global system is completely determined by the local pictures. Now we will take this local point of view to the extreme and start with geometry on the *infinitesimal* approximations of the surface, i. e. the *tangent planes.* This may look unnatural at first, but in the remaining sections we shall see that all we need in order to develop the geometry is contained in this infinitesimal information.

We have seen that as we pass to smaller and smaller neighborhoods of a point in the hyperbolic plane the geometry resembles more and more Euclidean geometry, and the same is true of spherical geometry. Hence we should expect that the infinitesimal approximation of a general geometry should, indeed, be Euclidean. But essentially all of Euclidean geometry is encoded in the *inner product.* This leads up to the following definition:

Definition 5.3.1. *A* Riemannian metric on S *is a choice of inner product on every tangent space T_pS, varying smoothly with p.*

We use the notation $\langle w_1, w_2\rangle_p$ for the inner product on $T_p(S)$, or just $\langle w_1, w_2\rangle$ if the point p is understood.

The smooth dependence on p means the following: if $x : \mathcal{U} \to S$ is a local parametrization, the inner product is completely determined by its values on the basis $\{x_u, x_v\}$ — i. e. by the functions

$$E = \langle x_u, x_u\rangle,\ G = \langle x_v, x_v\rangle \text{ and } F = \langle x_u, x_v\rangle = \langle x_v, x_u\rangle\,. \tag{5.3.1}$$

The inner product is *smooth* if these functions are smooth for all local parametrizations.

A surface with a given Riemannian metric we be called a *Riemannian surface,* and we say that the surface has a *Riemannian structure.*

Caution: This is not the same as *Riemann surface,* which in the literature means a surface with a *complex analytic* structure, i. e. a complex manifold of (complex) dimension one. However, the definitions of Riemannian metric/structure etc. above have obvious generalizations to manifolds of any dimension n, which are then called *Riemannian n–manifolds.* We choose to use the terminology “Riemannian surface” rather than “Riemannian 2-manifold”.

Given a Riemannian metric, we define the *norm* $\|w\|_p$ of a vector $w \in T_pS$ by $\|w\|_p^2 = \langle w, w\rangle_p$. If $\omega(t) = x(u(t), v(t))$, $t \in [a, b]$ is a $\mathcal{C}^1$–curve and $p = \omega(t)$, we get

$$\|\omega'(t)\|_p^2 = E(p)(u'(t))^2 + 2F(p)u'(t)v'(t) + G(p)(v'(t))^2\,.$$

Motivated by the theory of curves in space, we define the *arc–length* of ω between the parameter values a and t by

$$s(t) = \int_a^t \|\omega'(t)\|_{\omega(t)}\,dt\,.$$

This formula can also be written

$$(\frac{ds}{dt})^2 = E(\frac{du}{dt})^2 + 2F\frac{du}{dt}\frac{dv}{dt} + G(\frac{dv}{dt})^2\,.$$

This should hold for any parametrization of the same curve, so we write more concisely

$$ds^2 = Edu^2 + 2F du\,dv + Gdv^2. \tag{5.3.2}$$

(Note the similarity with the arc–length formula in Euclidean and hyperbolic geometry.)

Since the metric is determined over $x(\mathcal{U})$ by E, F and G, we also use (5.3.2) as a notation for the Riemannian metric. The relation with arc length is also a good reason for using the word "metric": given two points p and q, let $\Omega_{p,q}$ be the set of all $\mathcal{C}^1$–curves on S from p to q. Then we define

$$d(p,q) = \inf\{\ell(\gamma) | \gamma \in \Omega_{p,q}\}\,, \tag{5.3.3}$$

where $\ell(\gamma)$ is the arc length of γ. One can show that this defines a metric on S in the sense of topology (exercise 6), and that the topology defined by this metric is the underlying topology on S (exercise 7). Hence we can also think of the Riemannian metric as an *infinitesimal* version of a topological metric defining the topology on S.

I addition to giving rise to a distance function along curves, the metric also makes it possible to define the *angle* between two curves at a point of intersection as the angle between the tangent vectors at that point. Thus, if $\alpha(t)$ and $\beta(t)$ are two smooth curves such that $\alpha(0) = \beta(0)$, they meet at the angle $\phi \in [0, \pi]$ determined uniquely by

$$\langle \alpha'(0), \beta'(0) \rangle = \cos\phi \, \|\alpha'(0)\| \, \|\beta'(0)\|.$$

Hence we already see that the Riemannian metric captures features that are clearly *geometric* in nature. *Differential geometry* deals with how and to what extent the Riemannian metric determines geometric properties of the surface.

Examples 5.3.2. (1) Parametrizing $\mathbb{R}^2$ by the identity map, we get $ds^2 = dx^2 + dy^2$. This is the Euclidean plane as a Riemannian surface. Using *polar coordinates*, we parametrize $\mathbb{R}^2 - \{0\}$ by $z(r,\theta) = (r\cos\theta, r\sin\theta)$. Then $z_r = (\cos\theta, \sin\theta)$ and $z_\theta = (-r\sin\theta, r\cos\theta)$. Thus $E(r,\theta) = 1,\ F(r,\theta) = 0$ and $G(r,\theta) = r^2$, hence $ds^2 = dr^2 + r^2 d\theta^2$.

(2) If $S \subset R^3$ is a regular surface, the tangent planes inherit inner products from the standard inner product $\eta \cdot \xi = \eta\xi^t$ on R^3. This defines a Riemannian structure on S, and we will always assume that a regular surface comes with this Riemannian structure. The inner product, or the local expression $Edu^2 + 2F du\, dv + G dv^2$, is then called *the first fundamental form* of S.

For example, let S be the graph of the smooth function $f(u,v)$. The natural choice of parametrization is $x(u,v) = (u, v, f(u,v))$, with $x_u = (1, 0, f_u)$, $x_v = (0, 1, f_v)$. Hence

$$E = \|x_u\|^2 = 1 + f_u^2,\ F = x_u \cdot x_v = f_u f_v,\ G = \|x_v\|^2 = 1 + f_v^2.$$

(f_u and f_v are the partial derivatives of f with respect to u and v).

(3) Let $S = \mathbb{H}$, i. e. Poincaré's upper half plane model for the hyperbolic plane. We have seen that arc length can be described infinitesimally by

$$ds^2 = \frac{dx^2 + dy^2}{y^2}\,. \tag{5.3.4}$$

This defines a Riemannian metric on $\mathbb{H}$ via the local parametrization given by the identity map. (Or, more precisely, the inverse of the inclusion $\mathbb{H} \subset \mathbb{R}^2$.) Thus $E(x, y) = G(x, y) = 1/y^2$, and $F = 0$. The identity $F = 0$ expresses that the two systems of coordinate lines $x =$constant and $y =$constant are orthogonal to each other, as in the Euclidean plane. In fact, since the metric at every point is just a multiple of the Euclidean metric, all angles will be the same in the two geometries. We say that the two metrics are *conformally equivalent.* (Cf. exc. 4.7.)

This is an example of a Riemannian structure on a surface which is not given as a regular surface in $\mathbb{R}^3$, and it is not a priori clear that (5.3.4) can be realized as a fundamental form. However, this will follow from Exercise 5.4.5.

On Poincaré's disk model $\mathbb{D}$, we similarly get a Riemannian metric with $E = G = 4/(1 - x^2 - y^2)^2$ and $F = 0$.

Area. Let $S \subset \mathbb{R}^3$ be a regular surface, and let $x : \mathcal{U} \to S$ be a local parametrization. If $\Omega \subset \mathcal{U}$ is a region bounded by, say, smooth curves, we know from calculus that the area of $R = x(\Omega)$ is

$$A(R) = \iint_\Omega \left\| \frac{\partial x}{\partial u} \times \frac{\partial x}{\partial v} \right\| dudv = \iint_\Omega \|x_u \times x_v\|\, dudv. \tag{5.3.5}$$

The area can also be expressed in terms of the functions E, F and G, using the identity

$$\|x_u \times x_v\|^2 = \|x_u\|^2 \|x_v\|^2 - \langle x_u, x_v\rangle^2 = EG - F^2. \tag{5.3.6}$$

Substituting this in (5.3.5) we obtain a formula which makes sense for general Riemannian surfaces:

$$A(R) = \iint_\Omega \sqrt{EG - F^2}\, dudv\,. \tag{5.3.7}$$

Note that if $R = R_2 \cup R_2$ with $R_1 \cap R_2$ a union of smooth curves, then $A(R) = A(R_1) + A(R_2)$. It follows that if R, more generally, is a region which is a union along such curves of a finite number of regions R_i of the type considered above, then we can define the area of R by

$$A(R) = \Sigma_i A(R_i)\,,$$

and this will be independent of choice of such subdivision of R. In particular, one can prove that any *compact* Riemannian surface has such subdivision. (This will be more or less obvious for all examples we will ever deal with.) Hence a compact Riemannian surface S has a well–defined area $A(S)$.

Remark 5.3.3. Since the tangent vectors x_u and x_v are linearly independent, it follows from the Schwarz inequality that $EG - F^2$ is always strictly positive on a Riemannian surface. This fact will be important several times in what follows.

Example 5.3.4. In $\mathbb{H}$ we have seen that $E = G = 1/y^2$ and $G = 0$. Hence, if we use the identity map as parametrization, we get the area formula

$$A(R) = \iint_R \frac{dxdy}{y^2}.$$

as before. By the same argument we recover the area formula for the Poincaré disk $\mathbb{D}$.

Exercises for 5.3

1. Compute the first fundamental form of the parametrization of a surface of rotation that you found in exercise 5.1.4.

2. Compute the first fundamental form of the sphere of radius r minus two antipodal points, parametrized using spherical coordinates.

3. Let $E\,du^2 + 2F du\,dv + G\,dv^2 = E'ds^2 + 2F'ds\,dt + G'dt^2$ be the metric in a coordinate neighborhood expressed in two parametrizations $x(u, v)$ and $y(s, t)$. Prove that

$$\begin{pmatrix} E' & F' \\ F' & G' \end{pmatrix} = J(x^{-1}y)^t \begin{pmatrix} E & F \\ F & G \end{pmatrix} J(x^{-1}y)\,,$$

where $J(x^{-1}y)$ is the Jacobian of the coordinate transformation $x^{-1}y$.

4. Use exercise 3 to prove the following:

(i) For a metric on S to be smooth it suffices that the functions E, F and G are smooth for all parametrizations in some differentiable atlas for S.

(ii) $E'G' - F'^2 = (EG - F^2)|J(x^{-1}y)|^2$.

(iii) The area given by formula (5.3.7) is independent of the parametrization.

5. Referring to Exercise 5.1.10, let $f : \triangle \to S$ be a smooth map from a triangle to a Riemannian surface S. Let $p \in \triangle$ be a vertex and assume that df_p is nonsingular. Show that the angle between the images of the two sides meeting at p is not 0 or π.

6. Show that formula (5.3.3) defines a topological metric on S.

7. Show that the metric in exercise 6 defines the original topology on S. (Hint: observe first that it suffices to do this for coodinate patches $x(U)$. Then reduce to comparing the metric on U defined by the fundamental form with the Euclidean metric on U.)

5.4 Isometries

While *homeomorphisms* are bijections preserving the topological structure and *diffeomorphisms* are homeomorphisms preserving the differentiable structure, we define *isometries* to be diffeomorphisms preserving the *Riemannian* structure. In precise terms, we say that a map $f : S \to S'$ between Riemannian surfaces is an isometry5 if it is a diffeomorphism and

$$\langle df_p(v), df_p(w)\rangle_{f(p)} = \langle v, w\rangle_p \tag{5.4.1}$$

for every $p \in S$ and $v, w \in T_pS$. If we omit the condition that f should be a diffeomorphism, we call it a *local isometry.* Condition (5.4.1) implies that df_p is an isomorphism (apply it to $v = w$), so f will automatically be a local diffeomorphism by the inverse function theorem.

The polarization formula $2\langle v, w\rangle = \|v + w\|^2 - \|v\|^2 - \|v\|^2$ shows that an inner product is determined by the norm it defines. Hence it suffices to check that $\|df_p(v)\|_{f(p)} = \|v\|_p$ for every $v \in T_pS$.

The identity map is trivially an isometry, and, using the chain rule, we see that compositions and inverses of isometries are again isometries. Hence the set of isometries of a Riemannian surface to itself is a group under composition — the *isometry group* of S.

Examples 5.4.1. (1) The Euclidean isometries $g \in E(2)$, i. e. the distance preserving maps $g(x) = Ax + b$, where A is an orthogonal matrix and $b \in \mathbb{R}^2$, are also the isometries of the Riemannian surface $\mathbb{R}^2$.

(2) If S and S' are surfaces in R^3 and $f : S \to S'$ is the restriction of a Euclidean isometry of $\mathbb{R}^3$, then f is clearly an isometry between S and S'. For example, any spherical isometry, i. e. multiplication by a matrix in $O(3)$, is an isometry of the Riemannian structure on every sphere with center at the origin.

(3) Every $f \in \textit{Möb}(\mathbb{H})$ is an isometry of the Riemannian structure on $\mathbb{H}$. This follows from the fact that f preserves arc length, since the infinitesimal expression for arc length defines the Riemannian metric. (Cf. exercise 2.) But it can also be instructive to check directly that df_z preserves the norm on $T_z\mathbb{H}$ for every $z \in \mathbb{H}$:

Let $f(z) = \dfrac{az+b}{cz+d}$, with a, b, c, d real and $ad - bc = 1$. Then $f'(z) = 1/(cz+d)^2$, so formula (2.2.1) gives $\operatorname{Im} f(z) = \operatorname{Im} z \, |f'(z)|$.

Identifying $T_z\mathbb{H}$ with $\mathbb{C}$, we have $df_z(w) = f'(z) \cdot w$. (Complex multiplication, see exercise 5.1.2.) If $|\ |$ denotes Euclidean norm in $T_z\mathbb{H}$ (i. e. $|x+iy|^2 = x^2 + y^2$), then the hyperbolic norm is given by $\|w\|_z = \dfrac{|w|}{\operatorname{Im} z}$. Hence

$$\|df_z(w)\|_{f(z)} = \frac{|f'(z)|\,|w|}{\operatorname{Im} f(z)} = \frac{|w|}{\operatorname{Im} z} = \|w\|_z \,.$$

Thus it follows that every $f \in \textit{Möb}^+(\mathbb{H})$ is an isometry of the Riemannian structure. On the other hand, the inversion $r(z) = -\bar{z}$ (reflection about the imaginary axis) preserves both the imaginary part and Euclidean norm, hence r is also an isometry. Since $\textit{Möb}(\mathbb{H})$ is generated by $\textit{Möb}^+(\mathbb{H})$ and r, every element of $\textit{Möb}(\mathbb{H})$ is an isometry.

A similar analysis for $\mathbb{D}$ shows that $\textit{Möb}(\mathbb{D})$ also consists of isometries of $\mathbb{D}$ as a Riemannian surface. Moreover, any hyperbolic isometry between $\mathbb{H}$ and $\mathbb{D}$ is also an isometry between the Riemannian structures. Hence the two models are isometric representations of the hyperbolic plane as a Riemannian surface.

These examples show that in the three classical geometries, geometric isometries are also Riemannian isometries. One can show that the opposite is also the case, such that the word “isometry” is uniquely defined in these three geometries.

A consequence is that surfaces with a geometric structure built on one of the models (X, G) in "Geometry on surfaces…", page 6, also have Riemannian structures locally isometric to the structure on X. For every chart $x : \mathcal{U} \to x(\mathcal{U}) \subset X$, we can define a Riemannian structure on $\mathcal{U}$ such that x becomes an isometry. (Define $\langle v, w\rangle$ as $\langle dx(v), dx(w)\rangle$.) Since the coordinate transformations preserve inner products, this will be independent of x.

Definition 5.4.2. *Aspects of the geometry which are preserved under all isometries are called* intrinsic*, whereas those which depend on a particular representation of the surface — e. g. as a regular surfaces in $\mathbb{R}^3$ — are called* extrinsic.

Arc length is an example of an intrinsically defined quantity, since it can be expressed only in terms of the norms in the tangent planes, and the norms are preserved under isometries. More explicitly: if $\omega : [a, b] \to S$ is a curve, then its arc length is defined as $l(\omega) = \int_a^b \|\omega'(t)\|_{\omega(t)} dt$. If $f : S \to S'$ is a smooth map, then $l(f\omega) = \int_a^b \|(f\omega)'(t)\|_{\omega(t)} dt = \int_a^b \|df_{\omega(t)}(\omega'(t))\|_{f(\omega(t))} dt$. But if f is an isometry, this is equal to $l(\omega)$. (A converse to this observation is given as Exercise 2.)

Similar remarks apply to the area function, proving that area is also an intrinsic quantity. (Exercise 4.)

The following proposition records some simple observations which will be very useful later. We use the notation E, G, F for the functions defining the metric using a local parametrization x, and similarly E', G', F' when we use a local parametrization x'.

Proposition 5.4.3. *a) Assume $p \in S$ has a neighborhood which is isometric to a neighborhood of q in S' by a local isometry f. If $x : \mathcal{U} \to S$ is a local parametrization around p, then $x' = f \circ x : \mathcal{U} \to S'$ is a local parametrization around q such that $E = E'$, $F = F'$ and $G = G'$.*

b) Conversely, suppose $x : \mathcal{U} \to S$ and $x' : \mathcal{U} \to S'$ are local parametrization of two Riemannian surfaces such that $E = E'$, $F = F'$ and $G = G'$. Then the composition $x' \circ x^{-1}$ is an isometry between $x(\mathcal{U})$ and $x'(\mathcal{U})$.

□

Exercises for 5.4

1. Let $f : S \to S'$ be a map of surfaces which is a local diffeomorphism. Show that if S' is Riemannian, then S has a unique Riemannian structure such that f is a local isometry.

2. Suppose that $f : S \to S'$ is a diffeomorphism between two Riemannian surfaces. Show that f is an isometry if and only if f preserves arc–length — i. e. such that $l_S(\omega) = l_{S'}(f \circ \omega)$ for all curves ω in S, where l_S $(l_{S'})$ is arc–length in S (S').

3. Show that a generalized cylinder (cfr. exercise 5.1.5b) is locally isometric to the standard Euclidean plane. (Hint: Reduce, by a suitable projection, to the case where the curve α lies in a plane orthogonal to δ. Then parametrize by arc–length.)

4. Show that isometric compact surfaces have the same area.

5. (a) Let $\mathbb{H}$ be the upper half–plane model for the hyperbolic plane, and let M be the subset $\{z \in \mathbb{H} \,|\, \operatorname{Im} z > 1\}$. Let $\mathcal{U} = \mathbb{R} \times (0, \infty)$ and define $x : \mathcal{U} \to M$ by $x(u, v) = u + \cosh v\, i$.

   Show that x is a parametrization of M and compute the metric $ds^2 = E\,du^2 + 2F du\,dv + G\,du^2$ in these coordinates.

   b) Define $y : \mathcal{U} \to \mathbb{R}^3$ by

   $$y(u, v) = (\frac{\cos u}{\cosh v}, \frac{\sin u}{\cosh u}, v - \tanh v)\,.$$

   Show that y parametrizes a regular surface of rotation Σ in $\mathbb{R}^3$ and compute its metric in these coordinates.

   c) Show that there is an element γ generating an infinite cyclic subgroup $\Gamma \subset \textit{Möb}(\mathbb{H})$ such that M/Γ and Σ are isometric.

   Explain why this shows that Σ is a hyperbolic surface.

   > This is a famous example of a hyperbolic surface realized as a regular surface — the "pseudosphere". In fact, this was the first example of a geometry ever noticed to provide a local model for hyperbolic geometry, by Beltrami in 1868. A few months later he published his models for the hyperbolic plane.
   >
   > Note that the pseudosphere is not compact. We will see later that there are no compact, hyperbolic regular surfaces. (Proposition 5.5.6.)

   d) Let $z_0 \in \mathbb{H}$ be an arbitrary point. Explain how we can use the results in this exercise to show that the standard metric on $\mathbb{H}$ in a neighborhood of z_0 (given by formula (5.3.4)) can be realized as a fundamental form.

6. Let S be a Riemannian surface with metric $\langle , \rangle$. Show that if λ is a positive real number, we can define a new metric $\langle , \rangle^\lambda$ by setting

   $$\langle v, w\rangle_p^\lambda = \lambda\langle v, w\rangle_p$$

   for every $p \in S$ and $v, w \in T_pS$. We shall refer to this as a *rescaling* of the metric.

   Let S_λ be S with this new Riemannian structure. Show that if S is compact, then S_λ and $S_{\lambda'}$ are not isometric if $\lambda \neq \lambda'$, but that all $\mathbb{R}^2_\lambda$ are isometric.

7. We say that the smooth map $g : S \to S'$ is *conformal* if there is a smooth function $k > 0$ on S such that

   $$\langle dg_p(v), dg_p(w)\rangle_p = k(p)\langle v, w\rangle_{g(p)}$$

   for every $p \in S$ and $v, w \in T_pS$. Show that g must be an angle preserving local diffeomorphism and that g^{-1} is also conformal if g is a diffeomorphism. (If so, g is a *conformal equivalence.*)

   Show that given the Riemannian surface S, all S_λ of exercise 6 are conformally equivalent.

8. Show that stereographic projection (section 2.1) is conformal. (This reproves Lemma 2.1.2.)

   Remark. One can prove that all Riemannian metrics on surfaces are *locally* conformally equivalent to the Euclidean metric $du^2 + dv^2$. (*Isothermal* coordinates.)

5.5 Curvature

Curvature is probably the most important concept in differential geometry. It comes in many variations, but they all try to measure different aspects of how the surface curves and bends as we move around on it. We shall here concentrate on what is called *Gaussian curvature,* which is motivated by the study of regular surfaces in $\mathbb{R}^3$. One reason why Gaussian curvature is so useful is that it is intrinsic. This is not at all obvious from the way we define it. Indeed, when the great mathematician C. F. Gauss discovered that it is, he named the result "Theorema egregium" — remarkable theorem.

Let first S be a regular surface in $\mathbb{R}^3$. To analyze how S varies near a point p is essentially the same as to analyze how the tangent planes of S turn in space as we move around p. The tangent plane is determined by the direction of a *normal vector*, so we may as well study how the unit normal vector $N(q)$ varies with q. In fact, a normal vector to a tangent plane can be thought of as a normal vector to the surface itself.

If $x : \mathcal{U} \to S \subset \mathbb{R}^3$ is a local parametrization, the vector cross product $x_u \times x_v$ is a normal vector. (It is non–zero since S is regular.) Therefore the unit normal is

$$N = \frac{x_u \times x_v}{\|x_u \times x_v\|}. \tag{5.5.1}$$

N is uniquely determined up to sign, and we get $-N$ by, for instance, changing the order of u and v (or x_u and x_v). Formula (5.5.1) clearly defines $N(q)$ as a smooth function of q with values in $\mathbb{R}^3$, and since $\|N\| = 1$, we may think of N as a smooth map $N : x(\mathcal{U}) \to S^2$. This is the *Gauss map.*

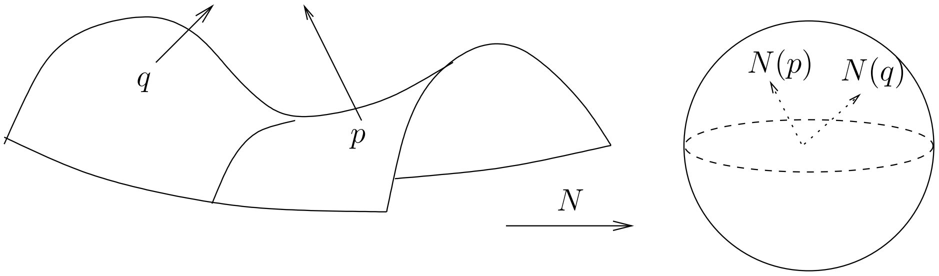


Fig. 5.5.1: The Gauss map

The derivative of the Gauss map at the point p is a linear transformation $dN_p : T_pS \to T_{N(p)}S^2$. But considered as subspaces of $\mathbb{R}^3$, T_pS and $T_{N(p)}S^2$ have the same normal vector (namely $N(p)$), hence they must coincide. Thus dN_p can naturally be thought of as a linear transformation $T_pS \to T_pS$. We define the *Gaussian curvature* (or just *curvature*) of S at the point p as the *determinant* of this linear transformation:

$$K(p) = \det(dN_p). \tag{5.5.2}$$

Observe that if we had used the opposite normal $(-N)$ instead, formula (5.5.2) would have given $K(p) = \det(d(-N_p)) = \det(-dN_p)$, which equals

$\det(dN_p)$ since T_pS has dimension 2. Hence it does not matter which unit normal vector we use.

The linear transformation $-dN_p : T_pS \to T_pS$ is also called the *shape operator* or *Weingarten map* of S at p.

The determinant of a matrix A measures the rate of distortion of area by the linear transformation $v \mapsto Av$, so $K(p)$ measures the infinitesimal distortion of area under the Gauss map. In fact, we have the following formula:

Proposition 5.5.1. *Let $p \in S$. Then*

$$|K(p)| = \lim_{R \to p} \frac{A_{S^2}(N(R))}{A_S(R)} .$$

where A_{S^2} and A_S are the area functions on S^2 and S, and the limit is taken over all regions R containing p.

If $K(p) \neq 0$, the Gauss map is a local diffeomorphism near p, and $K(p)$ is positive if N is orientation preserving, negative if it is orientation reversing.

To better understand the definition of curvature, it may be illuminating to consider the more familiar case of curves in the plane, from a similar point of view. Let $\alpha(t)$ be a regular parametrization of a curve $\mathcal{C}$ (i. e. $\alpha'(t) \neq 0$ for all t). The curvature κ of $\mathcal{C}$ is usually defined as the rate of change of the direction of the tangent as we move along the curve. Thus, if $T(\alpha(t)) = \alpha'(t)/|\alpha'(t)|$ is the unit tangent vector, then κ is defined by $|(T\alpha)'(t)| = \kappa(\alpha(t))|\alpha'(t)|$. But in the plane we might as well have considered the unit normal vector, e. g. defined by $N = \rho T$, where ρ is rotation by $\pi/2$. Then we also have $\kappa(\alpha(t))|\alpha'(t)| = |(N\alpha)'(t)|$. But N can be considered as a map $N : \mathcal{C} \to S^1$, which is the one–dimensional analogue of the Gauss–map. Since the tangent line of $\mathcal{C}$ at $\alpha(t)$ is generated by $\alpha'(t)$, the derivative dN of N is defined by $dN(\alpha'(t)) = (N\alpha)'(t)$. Hence this curvature measures the infinitesimal rate of distortion of *length* under the analogue of the Gauss map for curves in the plane.

Examples 5.5.2. (1) If S is a plane, then N is constant, and $dN = 0$. Hence the curvature is constant, equal to 0.

(2) If S is a cylinder (on an arbitrary curve), the normal will always lie in the plane normal to the axis of the cylinder. Hence the Gauss map will have image contained in a great circle in S^2. Therefore dN will have rank ≤ 1 and determinant 0, so the curvature will again be 0 everywhere.

(3) Let $S \subset \mathbb{R}^3$ be a sphere of radius r and center at the origin. Then the Gauss map is given as $N(p) = \dfrac{1}{r}\,p$, and $dN_p = \dfrac{1}{r}I_p$, where I_p is the identity map on T_pS. Hence the curvature is $1/r^2$ everywhere. ("A small sphere curves more than a big sphere.")

We will now derive an important formula for the $K(p)$ in terms of local coordinates. The formula is the basis for most direct calculations, and also for the proof of Gauss' *Theorema egregium* below.

Let $S \subset R^3$ be a regular surface and let $x : \mathcal{U} \to S$ be a local parametrization around p. Then the basis vectors $\{x_u, x_v\}$ can be thought of as functions $\mathcal{U} \to \mathbb{R}^3$, and we can take further partial derivatives and define, e. g., x_{uu}, x_{vv} and $x_{uv} = x_{vu}$. (The last equality holds because x is at least three times differentiable.)

The composition $N \circ x$ is a smooth map $\mathcal{U} \to \mathbb{R}^3$. The partial derivatives we denote by N_u, N_v, i. e.

$$N_u = \frac{d}{dt} N(x(u+t, v))|_{t=0} = dN(x_u),$$

and similarly for N_v.

Differentiating the identities $N \cdot x_u = N \cdot x_v = 0$ with respect to u and v, we get $N_u \cdot x_u + N \cdot x_{uu} = 0$, $N_v \cdot x_u + N \cdot x_{uv} = 0$, $N_u \cdot x_v + N \cdot x_{vu} = 0$ and $N_v \cdot x_v + N \cdot x_{vv} = 0$.

Now define functions e, f, g by

$$\begin{aligned} e &= N \cdot x_{uu} = -N_u \cdot x_u, \\ g &= N \cdot x_{vv} = -N_v \cdot x_v, \\ f &= N \cdot x_{uv} = N \cdot x_{vu} = -N_u \cdot x_v = -N_v \cdot x_u. \end{aligned} \tag{5.5.3}$$

Then we have

Proposition 5.5.3.

$$K = \frac{eg - f^2}{EG - F^2}\,. \tag{5.5.4}$$

Proof. Let $\begin{bmatrix} \alpha & \gamma \\ \beta & \delta \end{bmatrix}$ be the matrix of dN relative to the basis $\{x_u, x_v\}$. Thus

$$\begin{aligned} N_u &= dN(x_u) = \alpha\, x_u + \beta\, x_v, \\ N_v &= dN(x_v) = \gamma\, x_u + \delta\, x_v. \end{aligned} \tag{5.5.5}$$

Taking inner product of both sides of these equations with x_u and x_v we get, using (5.5.3):

$$\begin{aligned} -e &= N_u \cdot x_u = \alpha\, E + \beta\, F, \\ -f &= N_u \cdot x_v = \alpha\, F + \beta\, G, \\ -f &= N_v \cdot x_u = \gamma\, E + \delta\, F, \\ -g &= N_u \cdot x_v = \gamma\, F + \delta\, G. \end{aligned}$$

These equations can be written on matrix form as

$$-\begin{bmatrix} e & f \\ f & g \end{bmatrix} = \begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix} \begin{bmatrix} E & F \\ F & G \end{bmatrix} . \tag{5.5.6}$$

But curvature is defined as the determinant of the matrix $\begin{bmatrix} \alpha & \gamma \\ \beta & \delta \end{bmatrix}$, hence

$$eg - f^2 = K(EG - F^2). \tag{5.5.7}$$

□

Remark 5.5.4. The functions e, f and g define the *second differential form.* This is definitely not intrinsic, but composed of the normal components of the second derivatives of the parametrization $x(u, v)$, it gives important information on how S lies in $\mathbb{R}^3$.

Because of the normalization of N, it is usually best to avoid differentiating N. Hence, actual calculations are often simpler if we use the formulas $e = N \cdot x_{uu}$, $g = N \cdot x_{vv}$ and $f = N \cdot x_{uv}$, rather than the other variants.

Example 5.5.5. Let $S \subset \mathbb{R}^3$ be the graph of a smooth function $h(x, y)$ defined on an open subset $\mathcal{O} \subset \mathbb{R}^2$. The parametrization $z(x, y) = (x, y, h(x, y))$ gives, as we have seen, $z_x = (1, 0, h_x)$, $z_y = (0, 1, h_y)$ and $E = 1 + h_x^2$, $G = 1 + h_y^2$, $F = h_x h_y$. The unit surface normal is

$$N = \frac{z_x \times z_y}{\|z_x \times z_y\|} = \frac{(-h_x, -h_y, 1)}{\sqrt{1 + h_x^2 + h_y^2}} .$$

Differentiating z once more, we get $z_{xx} = (0, 0, h_{xx})$, $z_{yy} = (0, 0, h_{yy})$ and $z_{xy} = (0, 0, h_{xy})$, hence

$$e = \frac{h_{xx}}{\sqrt{1 + h_x^2 + h_y^2}} , \quad g = \frac{h_{yy}}{\sqrt{1 + h_x^2 + h_y^2}} \quad \text{and} \quad f = \frac{h_{xy}}{\sqrt{1 + h_x^2 + h_y^2}} .$$

Consequently,

$$K = \frac{h_{xx} h_{yy} - h_{xy}^2}{(1 + h_x^2 + h_y^2)^2} = \frac{\det H(h)}{(1 + h_x^2 + h_y^2)^2} ,$$

where $H(h)$ is the *Hessian* of h. It follows, for example, that at a non–degenerate critical point of h (i. e. a critical point where $\det H(h) \neq 0$), the curvature is positive if the point is a maximum or minimum point for f and negative if it is a saddle point. In fact, at a critical point the curvature is *equal to* the determinant of the Hessian.

Recalling from Proposition 5.1.9 that any regular surface is locally a graph, we now have a geometric interpretation of the *sign* of the curvature at a point p, if $K(p) \neq 0$. In particular, from property (ii) of Remark 5.1.10 and its converse in Exercise 5.1.8, we see that if the curvature is negative at every point, a function defined by projection to a line cannot have a maximum point. It follows that a *compact* regular surface cannot have everywhere negative curvature.

We will discuss further the problem of realizing compact geometric surfaces later, but first we will prove one of the most important results of differential geometry:

Theorema egregium. *Gaussian curvature is intrinsic.*

Proof. We want to prove that if we choose local coordinates $x(u, v)$, the curvature can be expressed entirely in terms of the functions E, F and G and their derivatives. As in Proposition 5.4.3b, on an isometric surface we can find local coordinates defined for the same values of (u, v) and giving rise to the same functions E, F, G. But then the curvatures also have to be the same.

We start with formula (5.5.4), written as

$$K = \frac{(x_{uu} \cdot N)(x_{vv} \cdot N) - (x_{uv} \cdot N)^2}{EG - F^2}.$$

Now we substitute the expression $N = \dfrac{x_u \times x_v}{\|x_u \times x_v\|} = \dfrac{x_u \times x_v}{\sqrt{EG - F^2}}$ for N and use the formula $((\mathbf{a} \times \mathbf{b}) \cdot \mathbf{c}) = \det \begin{pmatrix} \mathbf{a} \\ \mathbf{b} \\ \mathbf{c} \end{pmatrix} = \det \begin{pmatrix} \mathbf{a} \\ \mathbf{b} \\ \mathbf{c} \end{pmatrix}^t$ for the triple vector product. Then

$$\begin{aligned} K(EG - F^2)^2 &= (x_{uu} \cdot (x_u \times x_v))(x_{vv} \cdot (x_u \times x_v)) - (x_{uv} \cdot (x_u \times x_v))^2 \\ &= \det \begin{pmatrix} x_{uu} \\ x_u \\ x_v \end{pmatrix} \det \begin{pmatrix} x_{vv} \\ x_u \\ x_v \end{pmatrix}^t - \det \begin{pmatrix} x_{uv} \\ x_u \\ x_v \end{pmatrix} \det \begin{pmatrix} x_{uv} \\ x_u \\ x_v \end{pmatrix}^t \\ &= \det \left[\begin{pmatrix} x_{uu} \\ x_u \\ x_v \end{pmatrix} \begin{pmatrix} x_{vv} \\ x_u \\ x_v \end{pmatrix}^t \right] - \det \left[\begin{pmatrix} x_{uv} \\ x_u \\ x_v \end{pmatrix} \begin{pmatrix} x_{uv} \\ x_u \\ x_v \end{pmatrix}^t \right] \end{aligned}$$

$$= \det\begin{pmatrix} x_{uu}\cdot x_{vv} & x_{uu}\cdot x_u & x_{uu}\cdot x_v \\ x_u\cdot x_{vv} & E & F \\ x_v\cdot x_{vv} & F & G \end{pmatrix} - \det\begin{pmatrix} x_{uv}\cdot x_{uv} & x_{uv}\cdot x_u & x_{uv}\cdot x_v \\ x_u\cdot x_{uv} & E & F \\ x_v\cdot x_{uv} & F & G \end{pmatrix}.$$

It remains to express the inner products in the last two matrices in terms of E, F and G.

Taking derivatives with respect to u and v yields

$$\begin{aligned} E_u &= 2\,x_{uu}\cdot x_u, & E_v &= 2\,x_{uv}\cdot x_u, \\ F_u &= x_{uu}\cdot x_v + x_u\cdot x_{uv}, & F_v &= x_{uv}\cdot x_v + x_u\cdot x_{vv}, \\ G_u &= 2\,x_{uv}\cdot x_v, & G_v &= 2\,x_{vv}\cdot x_v. \end{aligned}$$

From these equations we easily get expressions for the six inner products $x_a\cdot x_{bc}$, $a,b,c\in\{u,v\}$ in terms of the derivatives E, F and G. For example, $x_u\cdot x_{uv} = \frac{1}{2}E_v$ and $x_u\cdot x_{vv} = F_v - \frac{1}{2}G_u$. Substituting these six expressions, we get

$$\begin{aligned} &K(EG-F^2)^2 = \\ &= \det\begin{pmatrix} x_{uu}\cdot x_{vv} & \frac{1}{2}E_u & F_u - \frac{1}{2}E_v \\ F_v - \frac{1}{2}G_u & E & F \\ \frac{1}{2}G_v & F & G \end{pmatrix} - \det\begin{pmatrix} x_{uv}\cdot x_{uv} & \frac{1}{2}E_v & \frac{1}{2}G_u \\ \frac{1}{2}E_v & E & F \\ \frac{1}{2}G_u & F & G \end{pmatrix} \\ &= \det\begin{pmatrix} x_{uu}\cdot x_{vv} - x_{uv}\cdot x_{uv} & \frac{1}{2}E_u & F_u - \frac{1}{2}E_v \\ F_v - \frac{1}{2}G_u & E & F \\ \frac{1}{2}G_v & F & G \end{pmatrix} - \det\begin{pmatrix} 0 & \frac{1}{2}E_v & \frac{1}{2}G_u \\ \frac{1}{2}E_v & E & F \\ \frac{1}{2}G_u & F & G \end{pmatrix}. \end{aligned}$$

(The last identity can be verified, for instance, by expansion of the determinants along the first columns.) To finish the proof, observe that

$$x_{uu}\cdot x_{vv} - x_{uv}\cdot x_{uv} = (x_u\cdot x_{vv})_u - (x_u\cdot x_{uv})_v = (F_v - \frac{1}{2}(G_{uu}+E_{vv}))\,,$$

which is an expression in the derivatives of E, F and G. Dividing by the factor $(EG-F^2)^2$, we now have an expression for K which only depends on the metric on S (*Brioschi's formula*). □

For the purpose of calculations, the actual formula yields little insight and is quite meaningless to memorize. A somewhat simpler version is given in exercise 6 when $F = 0$, and we will see later (section 5.7) that it is always possible to find parameters where this is the case.

But the fact that Gaussian curvature is preserved under (local) isometries has many important consequences. For example, it implies that $\mathbb{R}^2$ and S^2 are not locally isometric, since $\mathbb{R}^2$ has curvature 0 everywhere, and S^2 has positive curvature everywhere. This means that it is impossible to draw a map (on flat, "Euclidean" paper) of any part of the earth which scales all distances by the same amount. Hence any map of the earth must be distorted some way or another.

Another application is that any surface with a *transitive* group of isometries (like $\mathbb{R}^2$ and S^2) must have constant curvature, since the transitivity means that for every pair of points p and q there is an isometry mapping p to q.

On a more fundamental level, we can use the result to extend the definition of curvature to surfaces that are not given as regular surfaces in $\mathbb{R}^3$. It suffices that we know that they are *locally isometric* to such surfaces. More specifically, if $p \in S$ is a point in a Riemannian surface and $f : \mathcal{U} \to S'$ is an isometry of a neighborhood of p to a regular surface, we can define the curvature of S at p to be equal to the curvature of S' at $f(p)$. *Theorema egregium* then says that a different choice of local isometry would give the same result. Moreover, it follows from Proposition 5.4.3b that we can use Brioschi's formula to compute the curvature using any local coordinates on S'.

An important example is the hyperbolic plane. By a famous result of Hilbert, it is impossible to realize the hyperbolic plane as a regular surface in $\mathbb{R}^3$, but if you did Exercise 5.4.5, you have proved that the upper half–plane model $\mathbb{H}$ is *locally* isometric to the 'pseudosphere'. Therefore $\mathbb{H}$ has a well–defined curvature function. In fact, since $\mathbb{H}$ is an example of a Riemannian surface with a transitive group of isometries, the curvature must be constant. Using the standard metric on $\mathbb{H}$, given by formula (5.3.4), we have $E = G = 1/v^2$ and $F = 0$, and Brioschi's formula gives $K = -1$. Hence the hyperbolic plane is not even *locally* isometric to the Euclidean plane or to any sphere.

The following natural question now arises: can we also use the formula to *define* curvature, even for Riemannian surfaces which are not known to be locally isometric to regular surfaces?

The answer is indeed affirmative, but unfortunately we can not conclude this from our proof of *Theorema egregium*, since it uses very explicitly the Euclidean geometry in $\mathbb{R}^3$ and the assumption that E, F and G have the form given in formulas (5.3.1). One could, of course, try to prove that Brioschi's formula is invariant under *all* local coordinate changes by a brute force calculation. This is certainly theoretically possible, but the amount of work involved would be formidable, and probably not to be recommended. A better proof goes by verifying that the formula coincides with a more general definition which is manifestly invariant, using *Riemannian connections*. However, this would lead us too far astray, so we will accept this fact without proof. Thus, for abstract surfaces we will use Brioschi's formula as the definition of curvature.

Earlier we have seen that one can construct many *compact* hyperbolic surfaces, hence also compact Riemannian surfaces with constant curvature -1. But a consequence of the preceding discussion and the computation in Example 5.5.5 is that such surfaces cannot be realized as regular surfaces in $\mathbb{R}^3$. This means that to obtain a general geometric theory it is essential that we consider general abstract surfaces, and not just those that lie in $\mathbb{R}^3$.

What about compact *Euclidean* surfaces, like tori? The following result shows that neither can they be realized as regular surfaces:

Proposition 5.5.6. *A compact surface $S \subset \mathbb{R}^3$ must have a point where the curvature is strictly positive.*

Proof. Let $S \subset R^3$ be compact. Previously we have seen that a maximum point for a function gotten by projection to a line will be a point with curvature greater than or equal to zero. To obtain the sharper result we instead use the function

$$f(q) = |q|^2 = q \cdot q\,.$$

Let p be a maximum point for f. This exists since S is compact. Obviously $p \neq 0$. Let $\omega \in \Omega_p S$ represent a tangent vector $v = \omega'(0) \in T_pS$ and set $g(t) = f(\omega(t))$. Then 0 is a maximum point for $g(t)$, and we have

$$\text{(i)} : g'(0) = 0 \quad \text{and} \quad \text{(ii)} : g''(0) \leqslant 0\,. \tag{5.5.8}$$

Note that $g'(t) = 2\omega(t) \cdot \omega'(t)$ and $g''(t) = 2\omega(t) \cdot \omega''(t) + 2|\omega'(t)|^2$. From (5.5.8 i) we get $0 = g'(0) = 2p \cdot v$, and since this is true for every $v \in T_pS$, it follows that the vector $p \in \mathbb{R}^3$ is orthogonal to T_pS. Hence we can choose the surface normal such that

$$N(p) = \frac{1}{|p|} p\,. \tag{5.5.9}$$

The second condition (5.5.8 ii) yields the inequality $|\omega'(0)|^2 + \omega(0) \cdot \omega''(0) \leqslant 0$, or

$$p \cdot \omega''(0) \leqslant -|\omega'(0)|^2 = -|v|^2\,. \tag{5.5.10}$$

In order to prove that $K(p) = \det dN_p > 0$ we will use the following observation:

Algebraic lemma. *Let A be a real 2×2 matrix which is self adjoint with respect to an inner product $\langle\, ,\rangle$ on $\mathbb{R}^2$. Then $\det A > 0$ if and only if $\langle A(v), v\rangle \neq 0$ for all $v \neq 0$ in $\mathbb{R}^2$.*

The easy proof of this lemma is left as an exercise (8), as is the verification that it applies to dN_p (5.7).

It now remains to verify that $dN_p(v) \cdot v \neq 0$ for every nonzero $v \in T_pS$. Let v be represented by $\omega \in \Omega_p S$. Then $dN_{\omega(t)}(\omega'(t)) = (N\omega)'(t)$, hence $dN_p(v) \cdot v = (N\omega)'(0) \cdot \omega'(0)$. But since $(N\omega)(t) \cdot \omega'(t) = 0$ for all t, we have

$$0 = ((N\omega) \cdot \omega')'(t) = (N\omega)'(t) \cdot \omega'(t) + (N\omega)(t) \cdot \omega''(t).$$

For $t = 0$ this gives

$$dN_p(v) \cdot v = -(N\omega)(0) \cdot \omega''(0) = -\frac{1}{|p|} p \cdot \omega''(0) \geqslant \frac{|v|^2}{|p|} > 0$$

since $v \neq 0$, by (5.5.9) and (5.5.10). □

Since compact Euclidean surfaces, e. g. the geometric surfaces R^2/Λ where $\Lambda \subset R^2$ is a discrete rank two subgroup, are locally isometric to $\mathbb{R}^2$, they must have constant curvature 0. Hence they can not occur as regular surfaces. The only remaining compact geometric surfaces are then those with *spherical* geometry. But there are only two such surfaces: S^2 and P^2, and one can show that P^2 cannot even be *topologically* realized in $\mathbb{R}^3$. (In fact, only *orientable* compact surfaces can be realized in R^3.) Consequently, S^2 is the only compact surface with geometric structure built on the classical geometries which can be realized as a regular surface in $\mathbb{R}^3$.

Note that the compactness condition here is essential. In addition to the pseudosphere of curvature -1, the Euclidean plane $\mathbb{R}^2 \subset \mathbb{R}^3$ or cylinders are obvious examples of regular surfaces of curvature 0.

Exercises for 5.5

1. Let $r(u, v) = (g(u) \cos v, g(u) \sin v, h(u))$ be a parametrization of a regular surface of rotation. Show that the Gaussian curvature is given by
$$K(r(u, v)) = \frac{h'(g'h'' - g''h')}{g(g'^2 + h'^2)^2}\,.$$

Which such surfaces have constant curvature 0?

2. Assume $0 < a < b$. Compute the curvature of the torus in $\mathbb{R}^3$ given by the parametrization

$$r(u, v) = ((a \cos u + b) \cos v, (a \cos u + b) \sin v, a \sin u).$$

Where is the curvature positive and where is it negative?

3. Consider again the surfaces of rotation of exercise 1, and assume that $g'(u)^2 + h'(u)^2 = 1$. (I. e. the generating curve is parametrized by arc length — see section 5.6.) Show that the formula simplifies to

$$K = -\frac{g''}{g}\,.$$

Discuss how we can find all surfaces of rotation of constant curvature and show that the only compact such surfaces are the standard spheres.

4. a) Let S_1 be the surface obtained by rotating the curve

$$u \mapsto (u, \ln u),\ \ u > 0.$$

in the xz–plane around the z–axis. Compute its curvature both using exercise 1 and by the method of Example 5.5.5.

b) Let S_2 be the "helicoid" given by the parametrization

$$y(u, v) = (u \cos v, u \sin v, v),\ \ u > 0,\ \ v \text{ arbitrary}.$$

Compute its curvature.

c) Show that there is a smooth map $f : S_2 \to S_1$ such that $K(f(p)) = K(p)$ for every p, but that no such map can be a local isometry. This proves that there is no "converse" to Theorema egregium: the curvature function does not determine the surface up to local isometry.

(Hint: The u–coordinates must correspond. Now use exercise 5.3.3.)

5. What happens to the Gaussian curvature when we scale the metric as in exercise 5.4.6?

6. Show that if $F = 0$, then

$$K = -\frac{1}{2\sqrt{EG}}\left(\frac{\partial}{\partial v}\left(\frac{E_v}{\sqrt{EG}}\right) + \frac{\partial}{\partial u}\left(\frac{G_u}{\sqrt{EG}}\right)\right)\,.$$

As an application, show that the hyperbolic plane has curvature -1 at every point.

7. Show that the derivative dN_p of the Gauss map for a regular surface in R^3 is self adjoint with respect to the first fundamental form. Conclude that it has real eigenvalues and an orthogonal basis of eigenvectors.

   (The negatives k_1 and k_2 of the eigenvalues are called the *principal curvatures* at the point p, and the directions of the eigenvectors are called the *principal directions* at p.

   Note that $K(p) = k_1 k_2$. This is sometimes taken as the *definition* of curvature.)

8. Prove the algebraic lemma in the proof of Proposition 5.5.6.

9. If $S \subset R^3$ is a regular surface, define the *mean curvature* of S at a point p to be

   $$H(p) = -\frac{1}{2} tr(dN_p) = \frac{1}{2}(k_1 + k_2).$$

   Prove that $H^2 \geq K$. When do we have equality here? Is H intrinsic?

10. The definition of mean curvature in the previous exercise depends (up to sign) on the choice of normal direction, but the condition that $H(p) = 0$ everywhere does not. If this condition is satisfied, S is called a *minimal* surface. Why can a compact surface never be minimal?

11. Show that $N_u \times N_v = K(x_u \times x_v)$.

    Show that if $K \neq 0$ everywhere on a region $R \subset x(U) \subset S$, then

    $$A(N(R)) = \iint_{x^{-1}} K|x_u \times x_v|\, du\, dv.$$

    Use this to prove Proposition 5.5.1 when $K(p) \neq 0$.

12. Suppose that the first fundamental form of a surface $S \subset \mathbb{R}^3$ has the form $h(u, v)(du^2 + dv^2)$, with respect to some parametrization $x(u, v) = (f_1(u.v), f_2(u, v), f_3(u, v))$. Show that the image of x is minimal if and only if the three coordinate functions f_i are *harmonic*, i. e.

    $$\frac{\partial^2 f_i}{\partial u^2} + \frac{\partial^2 f_i}{\partial v^2} = 0 \quad \text{for } i = 1, 2, 3\,.$$

5.6 Geodesics

We will now introduce the “lines” of the geometry on Riemannian surfaces. They will be defined as curves satisfying certain properties which characterize the straight lines in Euclidean geometry. There are at least two possible characterizations we might use.

(1) Straight lines are curves which minimize distance between its points.

This is in many ways an attractive definition. It is obviously intrinsic and it can be studied on arbitrary Riemannian surfaces, and we have seen that it is satisfied by the hyperbolic lines in the hyperbolic plane. But one has to find a formulation such that *closed* curves are allowed, as on S^2. This can be done locally, leading to a variational problem which can be solved. However, even in the case of $\mathbb{R}^2$ this requires a significant amount of work.

(2) Straight lines are curves which never change direction.

What do we mean by this? The direction of a curve $t \mapsto \beta(t)$ in $\mathbb{R}^3$ at $\beta(t)$ is the direction of the derivative $\beta'(t)$ at that point. To measure change of direction we need the second derivative $\beta''(t)$, and in $\mathbb{R}^2$ or $\mathbb{R}^3$ a precise formulation of (2) could be

(2’) Straight lines are curves with parametrizations $\beta(t)$ such that $\beta'(t)$ and $\beta''(t)$ are linearly dependent.

This is a much simpler condition to deal with in $\mathbb{R}^2$, since it leads directly to differential equations without the detour through the calculus of variations (exercise 1). However, it does not apply as it stands to curves on other surfaces, even surfaces in $\mathbb{R}^3$, since $\beta''(t)$ in general will not be a tangent vector of the surface. But there is a way around this: replace $\beta''(t)$ by the component that ’can be seen from the surface’. This leads to the definition of the *covariant second derivative*, which is precisely the replacement for the second derivative we need. With $\beta''(t)$ replaced by the covariant second derivative, condition (2’) will define our ’lines’. Finding them is then reduced to solving ordinary differential equations.

This is the approach we choose. The relation with condition (1) will be discussed briefly at the end of the section.

Consider a regular surface $S \subset \mathbb{R}^3$ (as always with Riemannian structure inherited from $\mathbb{R}^3$), and let $\beta : (-\epsilon, \epsilon) \to S$ be a smooth curve. Then $\beta'(t)$ and $\beta''(t)$ are both defined as vectors in $\mathbb{R}^3$, and $\beta'(t) \in T_{\beta(t)}S$ for all t. In general $\beta''(t)$ does not lie in $T_{\beta(t)}S$, but there is a unique orthogonal decomposition

$$\beta''(t) = D\beta''(t) + P_N(\beta''(t)),$$

where $D\beta''(t) \in T_{\beta(t)}S$ is the tangential component of $\beta''(t)$ and $P_N(\beta''(t))$ is its normal component. In other words, $D\beta''(t)$ and $P_N(\beta''(t))$ are the images of the orthogonal projections of the vector $\beta''(t)$ on $T_{\beta(t)}S$ and the line perpendicular to it. (See Fig. 5.6.1.)

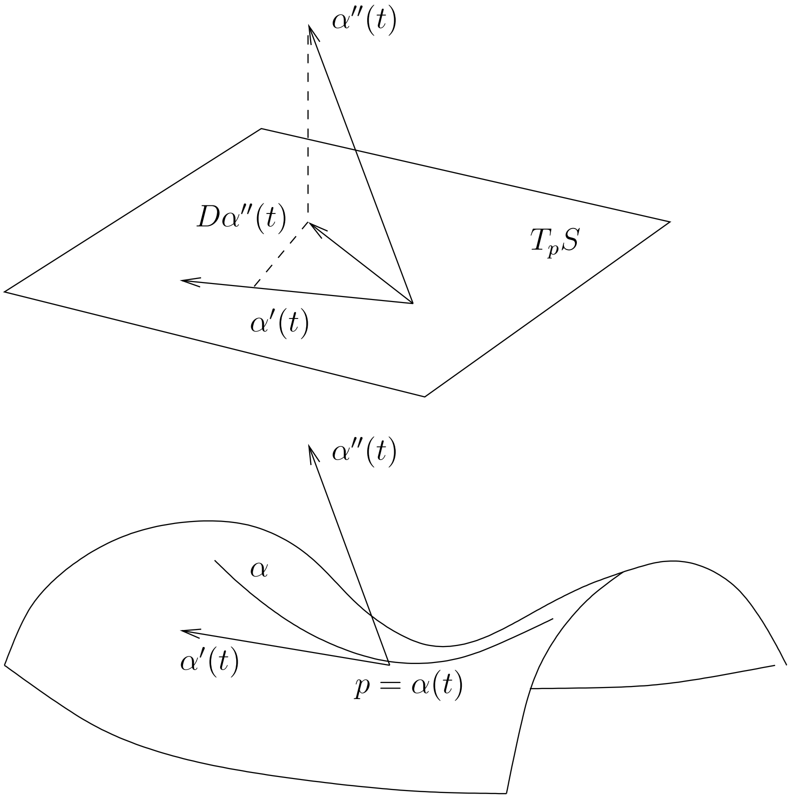


Fig. 5.6.1: Covariant second derivative

Definition 5.6.1. *$D\beta''$ is called* covariant second derivative *of β.*

We now use $D\beta''$ to define 'lines', but in order for the "direction of β" to make sense at every point, we need to require that $\beta'(t) \neq 0$ everywhere, i. e. the parametrization should be *regular.*

Definition 5.6.2. *The curve β is called a* geodesic *if $\beta'(t) \neq 0$ and $D\beta''(t)$ is a multiple of $\beta'(t)$ for all t.*

Remark 5.6.3. It is not hard to see that this condition is preserved if we reparametrize the curve, e. g. if we replace $\beta(t)$ by $\beta(h(t))$, where $h(t)$ is a diffeomorphism (exercise 2). Therefore this is really a property of the curve as a geometric object, and not of the particular parametrization.

To get a feeling for how this works we consider some examples, before proceeding with the theory.

Example 5.6.4. (1) A straight line in $\mathbb{R}^3$ is a geodesic in any surface that contains it, since the line can be parametrized by a linear function with vanishing second derivative.

(2) The difference between $\beta''(t)$ and $D\beta''(t)$ is a multiple of the surface normal $N(\beta(t))$. Hence the curve is a geodesic if and only if the three vectors $\beta'(t)$, $\beta''(t)$ and $N(\beta(t))$ are linearly dependent. This form of the condition is often easy to check on concrete examples, either geometrically or via the equation

$$\det \begin{pmatrix} \beta'(t) \\ \beta''(t) \\ N(\beta(t)) \end{pmatrix} = 0\,, \tag{5.6.1}$$

when the vectors are written on component form. An important special case is when the curve lies in the intersection between the surface and a plane which is not tangent to the surface. Then $\beta'(t)$ and $\beta''(t)$ both lie in the plane, and the curve will be a geodesic if the surface normal also lies in the plane. For example, on a surface of rotation with parametrization $r(u,v) = (g(u)\cos v, g(u)\sin v, h(u))$ all the generating curves $v =$ constant will be geodesics, and the circles $u =$ constant are geodesics if and only if $g'(u) = 0$.

Hence all great circles on standard spheres are geodesics.

Equation (5.6.1) is convenient for testing special curves, but far too complicated for finding geodesics in general. However, we shall see that if we require the parametrization to be of a special type, Definition (5.6.2) can be formulated as a much simpler differential equation.

Assume that $\beta'(t) \neq 0$ for all t. Then there is an orthogonal decomposition of the tangent space $T_{\beta(t)}S$ into the sum of the *tangent line* and the *normal line* of the curve. The tangent line is generated by $\beta'(t)$, and we let $T(t)$ denote the unit vector $T(t) = \frac{1}{\|\beta'(t)\|}\beta'(t)$. Hence we have a further decomposition

$$D\beta''(t) = a(t)T(t) + \nu(\beta(t))\,, \tag{5.6.2}$$

where $\langle T(t), \nu(\beta(t))\rangle = 0$. The coefficient $a(t)$ can be computed by taking inner product with $T(t)$:

$$a(t) = \langle D\beta''(t), T(t)\rangle = \frac{1}{\|\beta'(t)\|}\langle \beta''(t), \beta'(t)\rangle = \frac{1}{2\|\beta'(t)\|}\frac{d}{dt}(\|\beta'(t)\|^2). \tag{5.6.3}$$

It follows that $a(t)$, hence also the tangential component $a(t)T(t)$ of $D\beta''(t)$, is *intrinsic.* (In fact, although $D\beta''(t)$ so far is only defined on surfaces in $\mathbb{R}^3$, we could use formula (5.6.3) to define its tangential component for any regular curve on a general surface.)

Recall now that the arc length function $s(t)$ is determined (up to addition of a constant) by $\dfrac{ds}{dt} = \|\beta'(t)\|$, and since this is nonzero everywhere, we can invert $s = s(t)$ and write $t = t(s)$. Then $\alpha(s) = \beta(t(s))$ defines another parametrization of the same curve — *parametrization by arc length.* This parametrization has the property that $\|\alpha'(s)\| = 1$ for all s, hence the calculation above shows that the tangential component of $D\alpha''(s)$ vanishes. It follows that if a curve is parametrized by arc length, it is geodesic if and only if the covariant second derivative vanishes along the curve. Note that this is equivalent to saying that $\alpha''(s)$ is orthogonal to the surface.

Conversely, the same calculation gives that if a curve with a regular parametrization has vanishing covariant second derivatives, then $\dfrac{ds}{dt}$ is constant — i. e. the parametrization is by a multiple of arc length. We call such parametrizations *constant speed* parametrizations.

We summarize this discussion in

Proposition 5.6.5. *A curve parametrized by $\alpha(t)$ is a constant speed geodesic if and only if the covariant second derivative vanishes.*

In other words: the constant speed geodesics are the solutions of the second order differential equation

$$D\alpha''(s) = 0. \tag{5.6.4}$$

Example 5.6.6. Let $S \subset \mathbb{R}^3$ be a sphere of radius R and center at the origin. If $\alpha(s)$ is a curve on S parametrized by arc length, we have

$$D\alpha''(s) = \alpha''(s) - \frac{\alpha''(s)\cdot\alpha(s)}{R^2}\alpha(s)\,.$$

But $\alpha''(s)\cdot\alpha(s) = (\alpha'(s)\cdot\alpha(s))' - |\alpha'(s)|^2 = -1$. Hence α is a constant speed geodesic if and only if

$$\alpha''(s) + \frac{1}{R^2}\alpha(s) = 0\,.$$

The solutions of this vector equation have the form $\alpha(s) = \cos(s/R)A + \sin(s/R)B$, and it is easily seen that this is a curve on S if and only if A and B are orthogonal vectors on S. (Compute $|\alpha(s)|^2$.) But these curves are precisely the great circles on S.

In order to be able to work with equation (5.6.4) in general, we will express $D\alpha''(s)$ in local coordinates. (The property of being a geodesic is clearly a local property.) Assume that $\operatorname{Im}\beta \subset x(\mathcal{U})$ for a local parametrization $x : \mathcal{U} \to S$. We may then write $\beta(t) = x(u(t), v(t))$, and we have

$$\begin{aligned} \beta'(t) &= u'x_u + v'x_v, \\ \beta''(t) &= u''x_u + v''x_v + (u')^2x_{uu} + 2u'v'x_{uv} + (v')^2x_{vv}. \end{aligned} \tag{5.6.5}$$

We need expressions for Dx_{uu}, Dx_{uv} and Dx_{vv} — the orthogonal projections of x_{uu}, x_{uv} and x_{vv} on the tangent planes of S. First we write x_{uu}, x_{uv} and x_{vv} in terms of the basis (x_u, x_v, N), where N is the unit normal vector as in Proposition 5.2.2:

$$\begin{aligned} x_{uu} &= \Gamma^1_{11}x_u + \Gamma^2_{11}x_v + eN, \\ x_{uv} &= \Gamma^1_{12}x_u + \Gamma^2_{12}x_v + fN, \\ x_{vv} &= \Gamma^1_{22}x_u + \Gamma^2_{22}x_v + gN. \end{aligned} \tag{5.6.6}$$

The coefficients e, f and g are as in Proposition 5.5.3, as can be seen by taking inner product with N. Γ^k_{ij}, $i, j, k = 1, 2$ are called the *Christoffel symbols* of S with respect to the parametrization x. The effect of projecting to T_pS is just to remove the component along N.

Taking inner product of equations (5.6.6) with x_u and x_v and substitution of equalities derived in the proof of *Theorema egregium* yields a system of equations which can be written as

$$\begin{aligned} \begin{bmatrix} E & F \\ F & G \end{bmatrix} \begin{bmatrix} \Gamma^1_{11} \\ \Gamma^2_{11} \end{bmatrix} &= \begin{bmatrix} \frac{1}{2}E_u \\ F_u - \frac{1}{2}E_v \end{bmatrix}, \\ \begin{bmatrix} E & F \\ F & G \end{bmatrix} \begin{bmatrix} \Gamma^1_{12} \\ \Gamma^2_{12} \end{bmatrix} &= \begin{bmatrix} \frac{1}{2}E_v \\ \frac{1}{2}G_u \end{bmatrix}, \\ \begin{bmatrix} E & F \\ F & G \end{bmatrix} \begin{bmatrix} \Gamma^1_{22} \\ \Gamma^2_{22} \end{bmatrix} &= \begin{bmatrix} F_v - \frac{1}{2}G_u \\ \frac{1}{2}G_v \end{bmatrix}, \end{aligned} \tag{5.6.7}$$

or, more compactly:

$$\begin{bmatrix} E & F \\ F & G \end{bmatrix} \begin{bmatrix} \Gamma^1_{11} & \Gamma^1_{12} & \Gamma^1_{22} \\ \Gamma^2_{11} & \Gamma^2_{12} & \Gamma^2_{22} \end{bmatrix} = \begin{bmatrix} \frac{1}{2}E_u & \frac{1}{2}E_v & F_v - \frac{1}{2}G_u \\ F_u - \frac{1}{2}E_v & \frac{1}{2}G_u & \frac{1}{2}G_v \end{bmatrix}. \tag{5.6.8}$$

Substituting (5.6.6) in (5.6.5) and projecting to $T_{\beta(t)}S$, we get the following formula for the covariant second derivative in local coordinates:

$$\begin{aligned} D\beta''(t) = &\left(u'' + (u')^2\Gamma^1_{11} + 2u'v'\Gamma^1_{12} + (v')^2\Gamma^1_{22}\right) x_u \\ &+ \left(v'' + (u')^2\Gamma^2_{11} + 2u'v'\Gamma^2_{12} + (v')^2\Gamma^2_{22}\right) x_v. \end{aligned} \tag{5.6.9}$$

As a corollary we have the following analogue of *Theorema egregium*:

Theorem 5.6.7. *The covariant second derivative is intrinsic.*

Proof. This means that if $\Phi : S \to S'$ is an isometry between regular surfaces, then

$$d\Phi_{\beta(t)}(D\beta''(t)) = D(\Phi\beta)''(t). \tag{5.6.10}$$

("Covariant second derivatives are preserved by isometries.") Observe first that it follows from (5.6.8) that all Γ^k_{ij} can be expressed by E, F and G and their derivatives. Hence the Christoffel symbols are intrinsic.

Formula (5.6.9) is valid for any local parametrization. Hence, if $\Phi : S \to S'$ is an isometry, we get a similar formula for $D(\Phi\beta)''(t)$ using the parametrization Φx on S'. But then the functions $u(t)$ and $v(t)$ are the same as for β, and since Φ is an isometry, it follows that the Christoffel symbols also are preserved. Since $d\Phi_{\beta(t)}$ is linear and $(\Phi x)_u = d\Phi(x_u)$ and $(\Phi x)_v = d\Phi(x_v)$, (5.6.10) follows. □

Remark 5.6.8. The covariant second derivative is also independent of *direction*, in the sense that if we reverse the direction of the curve by replacing the parametrization $\alpha(t)$ by $\beta(s) = \alpha(c - s)$, for some constant c, then

$$D\beta''(s) = D\alpha''(c - s)$$

for every s, i. e. the covariant derivatives of the two curves are the same at every point of the curve. This follows easily from (5.6.9).

As a consequence of Proposition 5.6.5, formula (5.6.9) and Theorem 5.6.7 we now have

Theorem 5.6.9. (i) *Suppose $\alpha(s) = x(u(s), v(s))$ is a parametrization of a smooth curve. Then the curve is a constant speed geodesic if and only if $u(s)$ and $v(s)$ satisfy the following system of differential equations (the* geodesic equations*):*

$$\begin{aligned} u'' + (u')^2\Gamma^1_{11} + 2u'v'\Gamma^1_{12} + (v')^2\Gamma^1_{22} = 0, \\ v'' + (u')^2\Gamma^2_{11} + 2u'v'\Gamma^2_{12} + (v')^2\Gamma^2_{22} = 0. \end{aligned} \tag{5.6.11}$$

(ii) *Geodesics are preserved by isometries.*

Examples 5.6.10. (1) $\mathbb{R}^2$ with the Euclidean metric. Then E, F and G are constant, hence all $\Gamma^k_{ij} = 0$. Equations (5.6.11) reduce to $u'' = v'' = 0$, and the geodesic curves are the straight lines.

(2) S^2 with spherical coordinates: $x(u, v) = (\cos u \cos v, \sin u \cos v, \sin v)$. $x_u = (-\sin u \cos v, \cos u \cos v, 0)$, $x_v = (-\cos u \sin v, -\sin u \sin v, \cos v)$, and $E = \cos^2 v$, $F = 0$, $G = 1$. Equations (5.6.7) then yield $\Gamma^1_{12} = -\tan v$, $\Gamma^2_{11} = \cos v \sin v$, and $\Gamma^k_{ij} = 0$ otherwise. The geodesic equations are then

$$\begin{aligned} &u'' - 2\tan v\, u'v' = 0, \\ &v'' + \cos v \sin v (u')^2 = 0. \end{aligned}$$

With some effort these equations can be solved explicitly, but the more direct approach above is much better. However, it is also possible to argue as follows:

It is obvious that the equations have the solutions $u =$ constant, $v = as + b$. Hence all *meridians* are geodesics. But for any point p on the sphere we can find a linear isometry A taking $(0, 0, 1)$ to p , and A will map the meridians to the great circles through p. Since isometries map geodesics to geodesics, it follows that all great circles are geodesics.

This argument does not rule out the possibility that there could be more geodesics, but that will follow from the uniqueness in Proposition 5.6.11 below.

Note that our definition of covariant second derivative and the proof of Theorem 5.6.7 use the Euclidean geometry of $\mathbb{R}^3$ in essential ways. Hence the comments we made after the proof of Theorema egregium apply here, as well. We can define geodesics on surfaces which are locally isometric to regular surfaces as curves which map to geodesics by local isometries, and calculations can be done using Equations (5.6.11) — again referring to Proposition 5.4.3b.

But as with Brioschi's formula for curvature, it is possible to prove that the theory applies in full generality: we can use formula (5.6.9) with Christoffel symbols defined by (5.6.8) to *define* the covariant second derivative of curves on general Riemannian surfaces. Then Definition 5.6.2 and Proposition 5.6.5 immediately extends, and we can use Equations (5.6.11) to calculate geodesics. Then the whole theory will be intrinsic in the most general sense.

Examples 5.6.10 cont. (3) The upper half–plane model of the hyperbolic plane with coordinates $z = x+iy$: We have seen that $E = G = 1/y^2$, $F = 0$, and from (5.6.7) we get $\Gamma^1_{12} = \Gamma^2_{22} = -\Gamma^2_{11} = -1/y$ and $\Gamma^1_{11} = \Gamma^1_{22} = \Gamma^2_{12} = 0$.

The geodesic equations are now

$$\begin{aligned} x'' - \frac{2}{y} x' y' &= 0, \\ y'' + \frac{1}{y}(x')^2 - \frac{1}{y}(y')^2 &= 0. \end{aligned}$$

First we observe that the vertical lines $x = a$ are solutions. In fact, then the second equation reduces to $y'' - (y')^2/y = 0$, or $(y'/y)' = 0$. This equation has the solutions $y = be^{cs}$ for constants b and c, and if parametrization is by arc length, we must have $c = 1$. Hence $\alpha(s) = a + be^{s}i$ are parametrizations of the vertical $\mathbb{H}$–lines as geodesics. But since all $\mathbb{H}$–lines are images of vertical lines by hyperbolic isometries, it follows that all $\mathbb{H}$–lines are geodesics.

It follows from the uniqueness part of Proposition 5.6.11 below that these are all the geodesics, but we can also see this as follows: If $x' \neq 0$, the first equation is separable and can be written $\dfrac{x''}{x'} = \dfrac{2y'}{y}$. Integration gives $x' = Cy^2$, $C \neq 0$.

We could now try to substitute this in the second equation, but instead we use another trick which is often very useful. Namely, we will exploit the requirement that the geodesic should be parametrized by arc length, i. e. $(x')^2E + 2Fx'y' + (y')^2G = 1$.

In the present case this means that $(x')^2 + (y')^2 = y^2$. Now we substitute $x' = Cy^2$ into this equation and obtain $y' = \pm y\sqrt{1 - C^2y^2}$. This can be integrated explicitly, but instead we multiply by the expression $\pm Cy/\sqrt{1 - C^2y^2}$ on both sides and get

$$\frac{\pm Cyy'}{\sqrt{1 - C^2y^2}} = Cy^2 = x',$$

or, after integration:

$$\pm \frac{1}{C}\sqrt{1 - C^2y^2} = x - m,$$

for some constant m. Squaring, we obtain the equation for a circle with center on the real line:

$$(x - m)^2 + y^2 = \frac{1}{C^2}.$$

Hence we see that all geodesics are contained in $\mathbb{H}$–lines.

In concrete computations it is sometimes more convenient to use the general formulation $D\beta''(t) = \lambda(t)\beta'(t)$ than the equations (5.6.11), since this allows simpler parametrizations. An illustration of this in the case of the hyperbolic plane $\mathbb{H}$ is given in exercise 5.

In general it is of course hopeless to try to solve equations (5.6.11) explicitly. The real importance of the equations is that they allow us to prove general existence and uniqueness theorems for geodesic curves. In fact, (5.6.11) is a system of ordinary differential equations, and the following proposition follows from the general theory for such equations:

Proposition 5.6.11. *Every point of a Riemannian surface S has an open neighborhood $\mathcal{V}$ with the following property:*

There exist positive numbers ϵ and τ such that for every $q \in \mathcal{V}$ and $w \in T_qS$ with $\|w\| < \epsilon$, there is a unique constant speed geodesic $\gamma_w^q : (-\tau, \tau) \to S$ such that

$$\gamma_w^q(0) = q \quad \textit{and} \quad {\gamma_w^q}'(0) = w\,.$$

Moreover, $\gamma_w^q(t)$ depends smoothly on q, w and t.

The last sentence needs some explanation. Let $B_q(\epsilon) = \{w \in T_qS| \, \|w\| < \epsilon\}$, for $q \in \mathcal{V}$ and $\epsilon > 0$. Then $\eta(q, w, t) = \gamma_w^q(t)$ is defined on a subset of $\cup_q T_qS$, and we have not said what it means for a map from such a set to be smooth. What we mean is this:

We can assume that $\mathcal{V} \subset x(\mathcal{U})$ for some parametrization x. Then the mapping $\theta : \mathbb{R}^2 \times \mathcal{U} \to \cup_{q \in x(\mathcal{U})} T_qS$ defined by $\theta(a, b, u, v) = ax_u(u, v) + bx_v(u, v)$ is a bijection. The composition $\eta \circ \theta$ is defined on an open subset of $\mathbb{R}^2 \times \mathcal{U}$, and we say that η is smooth if this map is.

One can check that this does not depend on choice of local parametrization, and that it, in fact, defines a smooth manifold structure on $\cup_{p \in S} T_pS$ — the so–called *tangent bundle* of S, which plays a central role in differential geometry and topology.

The following observation is very useful: From the homogeneity of equations (5.6.11) it follows that if $\gamma(t)$ is a solution, then $\eta(t) = \gamma(ct)$ is also a solution. But $\eta(0) = \gamma(0)$ and $\eta'(0) = c\gamma'(0)$, so it follows that $\gamma_{cw}^q(t) = \gamma_w^q(ct)$. Using this and uniqueness, we can enlarge either one of the constants ϵ and τ in Theorem 5.6.11 at the expense of a reduction of the other. For example, we may assume $\tau > 1$ and define the *exponential map* $\exp_p : B_p(\epsilon) \to S$ by

$$\exp_p(w) = \gamma_w^p(1)\,,$$

for some $\epsilon > 0$. If $0 < s < 1$, we then have $\exp_p(sw) = \gamma_{sw}^p(1) = \gamma_w^p(s)$, hence $s \mapsto \exp_p(sw)$ is a geodesic and $\dfrac{d}{ds}\exp_p(sw)_{|s=0} = w$.

The following theorem states the most important local properties of the exponential map:

Theorem 5.6.12. *For every $p \in S$ there is an $\epsilon > 0$ such that*

(1) $\exp_p$ *is a diffeomorphism between $B_p(\epsilon)$ and a neighborhood of p.*

(2) *If ϵ is small enough, any two points in* $\exp_p(B_p(\epsilon))$ *can be joined by a unique geodesic of length less than 2ϵ.*

A neighborhood parametrized by the exponential map as in (1), is called a *normal neighborhood.* We come back to these in the next section. We end this section with some further remarks on geodesics, without proofs.

First we address the obviously very important problem of extending the exponential mapping to all of T_pS. Using the equality $\gamma^q_{cw}(t) = \gamma^q_w(ct)$ again, we see that this is equivalent to the following question: when can we extend constant speed geodesics infinitely in both directions?

We call S *geodesically complete* if $\exp_p$ is defined on all of T_pS, for all $p \in S$. It is trivial to construct examples which do not have this property: the simplest is to remove a point from a normal neighborhood. The famous *Hopf–Rinow* theorem gives exact conditions for a Riemannian structure to be complete. As mentioned earlier, a Riemannian metric determines a topological metric by $d(p, q) = \inf_\alpha l(\alpha)$ where α runs over all piecewise smooth curves from p to q and $l(\alpha)$ is the arc length of α. The topology defined by this metric is the given topology on the surface.

Theorem 5.6.13. (Hopf–Rinow) *A Riemannian surface is geodesically complete if and only it is complete in the metric d. Moreover, if S is complete (in either sense), then any two points p and q in S can be joined by a geodesic of length $d(p, q)$.*

Note that the last statement implies that the exponential map $\exp_p$ is *surjective* for every p.

Examples. (1) Since all compact metric spaces are complete, it follows that all compact Riemannian surfaces are geodesically complete.

(2) Regular surfaces S which are closed subsets of $\mathbb{R}^3$ are also geodesically complete. Note that the metric d is *not* the metric induced from the metric on $\mathbb{R}^3$ (except when S is a convex, open subset of a plane), but we clearly have the inequality $d(p, q) \geq |p - q|$ for all p and q. Therefore a Cauchy sequence with respect to the metric d is also Cauchy with respect to the induced metric. Hence it has a limit in S if S is closed as a subset of $\mathbb{R}^3$.

We conclude by stating an important result on the relationship between geodesics and “shortest length” curves. The proof will be given at the end of section 5.7.

We call a curve $\beta : (a,b) \to S$ locally *length-minimizing* if any $c \in (a,b)$ is contained in a smaller interval (a',b') such that $l(\beta|(a',b')) = d(\beta(a'),\beta(b'))$. A curve β from p to q such that $l(\beta) = d(p,q)$, is clearly locally length–minimizing, but the converse need not be true, as for example in the case of a great circle of a sphere.

Theorem 5.6.14. *A constant speed curve is locally length–minimizing if and only if it is geodesic.*

Thus, in this precise sense, the two characterizations of "lines" at the beginning of this section are indeed equivalent.

Exercises for 5.6

1. Let $t \mapsto \beta(t) \in \mathbb{R}^3$ be a smooth, regular curve such that $\beta'(t)$ and $\beta''(t)$ are everywhere linearly dependent. Show that the curve lies on a straight line.

2. Prove the statement of Remark 5.6.3.

3. Fill in the arguments for the claims in Example 5.6.4

4. Let S be the graph of the function $f(x,y) = 2x^2 - y^2$. Determine which intersections between S and planes through the z–axis are geodesics.

5. Show that the H–lines are geodesic by showing that $D\beta''(t)$ and $\beta'(t)$ are linearly dependent everywhere, for suitable parametrizations $\beta(t)$.

6. Let S be the cylinder with equation $x^2 + y^2 = 1$ in $\mathbb{R}^3$. Find infinitely many geodesics from $(1,0,0)$ to $(1,0,1)$ on S.

7. This exercise relies on the results and notation in exercise 5.4.5.

   Let w be an arbitrary point of the pseudosphere Σ. Show that there are infinitely many geodesics on Σ that go through w twice. Can a geodesic go through w more than twice?

8. Assume that S_1 and S_2 are two regular surfaces in $\mathbb{R}^3$ which intersect *tangentially* in a curve $\mathcal{C}$. Show that if $\mathcal{C}$ is a geodesic in S_1, then it is also geodesic in S_2.

9. Assume that S_1 and S_2 are two regular surfaces in $\mathbb{R}^3$ which intersect *orthogonally* in a curve $\mathcal{C}$. Show that if $\mathcal{C}$ is a geodesic in both S_1 and S_2, then it is a straight line in $\mathbb{R}^3$.

10. Compute the exponential map for the sphere S^2 at the point $p = (0,0,1)$, and show explicitly that it is defined on all of T_pS^2 and also that it is surjective.

    What is here the largest ϵ for which (1) and (2) of Theorem 5.6.12 are true?

11. Show that a rescaling of the metric by a constant factor (exercise 5.4.6) does not change the geodesics.

    Why does this prove that the set of geodesics does not determine the Riemannian structure up to isometry?

12. Assume that $f : S \to S'$ is a local isometry between geodesically complete surfaces, and let p be a point in S. Show that

    $$f(\exp_p(v)) = \exp_{f(p)}(df_p(v))$$

    for every $v \in T_pS$.

5.7 Geodesic polar coordinates

By Theorem 5.6.12 we can use exponential maps to define local parametrizations. These parametrizations are naturally suited to study local geometry, and they will play an essential rôle in the remaining sections. But first we have to choose coordinates on T_pS.

One possibility is to choose an orthonormal basis and use the associated coefficient vectors — i. e. an ordinary Cartesian coordinate system. But to study local behaviour around a point, it is often better to use *polar coordinates.* In fact, by comparing these two kinds of parameters, we shall see that we obtain a very precise description of the metric near a fixed point.

To fix notation, assume we have chosen an orthonormal basis for the tangent plane T_pS, giving rise to Cartesian coordinates (u,v). For $\theta \in \mathbb{R}$ we let $\alpha(\theta) \in T_pS$ be the point with Cartesian coordinates $(\cos\theta, \sin\theta)$. Then $\|\alpha(\theta)\| = 1$ and the curve $\theta \mapsto \alpha(\theta)$ parametrizes the 'unit circle' $\{w \in T_pS | \|w\| = 1\}$ in T_pS.

Recall that there is a natural identification of $T_w(T_pS)$ with T_pS for every $w \in T_pS$, giving T_pS (as a surface) a natural Riemannian structure. Then $\|\alpha'(\theta)\|_{\alpha(\theta)} = 1$, hence θ will also measure arc length (in radians).

Any $w \in T_pS$ can be written as $w = rv$, where $r = \|w\|_p$ and $\|v\|_p = 1$, and r and v are uniquely determined if $w \neq 0$. The map $(r, \theta) \mapsto r\alpha(\theta)$ from $\mathbb{R}^2$ to T_pS is a diffeomorphism when restricted to sets of the form $(0, \epsilon) \times J$, where J is an open interval of length $\leq 2\pi$. It follows that if $\epsilon > 0$ is small enough for $\exp_p$ to be a diffeomorphism on $B_\epsilon(0_p)$, the map

$$x(r, \theta) = \exp_p(r\alpha(\theta))\,, \ (r, \theta) \in (0, \epsilon) \times J$$

is a local parametrization of S. (r, θ) are then *geodesic polar coordinates.* Note that the formula is meaningful when $r \in (-\epsilon, \epsilon)$ and for θ in longer intervals J — for geodesically complete surfaces even for *all* r and θ — and it will often be useful to consider x as defined in this generality. But it is important to remember that x is a diffeomorphism only when the pair (r, θ) is restricted as stated.

A similar remark applies to the tangent vectors $x_r(r, \theta)$ and $x_\theta(r, \theta)$: they may be defined for all r and θ, but they do not form a basis everywhere. For example, $x_\theta(0, \theta) = 0$ always.

If we keep r or θ constant, $\exp_p(r\alpha(\theta))$ parametrizes *geodesic circles* or *geodesic radii.* These are the coordinate curves for geodesic polar coordinates. A word of warning, however: the geodesic radii are geodesics, but the geodesic circles are (usually) not!

Examples 5.7.1. (i) In $\mathbb{R}^2$ with $p = 0$, geodesic polar coordinates are just ordinary polar coordinates $x(r, \theta) = (r\cos\theta, r\sin\theta)$.

(ii) Around the point $p = (0, 0, 1) \in S^2$, geodesic polar coordinates coincide with spherical coordinates

$$x(r, \theta) = (\sin r\cos\theta, \sin r\sin\theta, \cos r).$$

(Cf. Exercise 5.6.10) The geodesic radii are great circles through p and geodesic circles are intersections of S^2 with horizontal planes.

(iii) For the hyperbolic plane we get the simplest formulas if we use the Poincaré disk model $\mathbb{D} \subset \mathbb{R}^2$ with $p = 0$. Then a point of hyperbolic distance r to p has Euclidean norm $\tanh(\frac{r}{2})$, and by the rotational symmetry of the disk model we get geodesic polar coordinates

$$x(r, \theta) = (\tanh(\frac{r}{2})\cos\theta, \tanh(\frac{r}{2})\sin\theta).$$

(I all three cases the fixed orthonormal basis is the standard basis for $\mathbb{R}^2$.)

We will now examine the functions $E(r,\theta) = \|x_r\|^2$, $G(r,\theta) = \|x_\theta\|^2$ and $F(r,\theta) = \langle x_r, x_\theta\rangle$ for geodesic polar coordinates.

E is the simplest: since $r \mapsto \exp_p(r\alpha(\theta))$ is a geodesic curve parametrized by arc length, $E = \langle x_r, x_r\rangle = 1$. The following *Gauss' lemma* tells us that $F = 0$, such that the metric has the form

$$ds^2 = dr^2 + G(r,\theta)d\theta^2. \tag{5.7.1}$$

Lemma 5.7.2. (Gauss' lemma.) *Let $\alpha : I \to T_pS$ be a regular curve such that $\|\alpha(t)\|$ is constant, and let $x(r,t) = \exp_p(r\alpha(t))$. Then $\langle x_r, x_t\rangle = 0$.*

Proof. Let Dx_{rr} denote the covariant second derivatives of the coordinate curves $r \mapsto \exp_p(r\alpha(t))$, t fixed. Then $Dx_{rr} = 0$, since these coordinate lines are by definition geodesics, parametrized by a constant multiple of arc length. But by equations (5.6.6), with r,t replacing u,v, we have

$$Dx_{rr} = \Gamma^1_{11}x_r + \Gamma^2_{11}x_t\,.$$

Thus $\Gamma^1_{11} = \Gamma^2_{11} = 0$, and

$$0 = \Gamma^1_{11}F + \Gamma^2_{11}G = F_r - \frac{1}{2}E_t\,.$$

Since $E = \langle x_r, x_r\rangle = \|\alpha(t)\|^2$ is constant, it follows that F is independent of r. But then, by continuity,

$$F(r,t) = \langle x_r(r,t), x_t(r,t)\rangle = \langle x_r(0,t), x_t(0,t)\rangle = 0\,.$$

□

Example 5.7.3. For use in the next section we examine what (5.7.1) looks like for the classical geometries. Clearly the metric will look the same at any point in any model we prefer, so we use the coordinates in Examples 5.7.1. Note that we only have to compute $G(r,\theta) = \langle x_\theta, x_\theta\rangle$.

$\mathbb{R}^2$: Geodesic polar coordinates are $x(r,\theta) = (r\cos\theta, r\sin\theta)$, giving $x_\theta = (-r\sin\theta, r\cos\theta)$. Hence

$$ds^2 = dr^2 + r^2 d\theta^2.$$

S^2: We get $x_\theta = (-\sin r\cos\theta, \sin r\cos\theta, 0)$. The metric is induced from the Euclidean metric in $\mathbb{R}^3$. Thus $G(r,\theta) = x_\theta \cdot x_\theta = \sin^2 r$, and

$$ds^2 = dr^2 + \sin^2 r d\theta^2.$$

$\mathbb{D}$: The metric in Cartesian coordinates is now

$$ds^2 = 4\frac{du^2 + dv^2}{(1 - u^2 - v^2)^2},$$

and the formula for geodesic polar coordinates in Example 5.7.1(iii) gives

$$x_\theta = (-\tanh(\frac{r}{2})\sin\theta, \tanh(\frac{r}{2})\cos\theta).$$

Writing $x_\theta = (a, b)$, this means that

$$\langle x_\theta, x_\theta\rangle = 4\frac{a^2 + b^2}{(1 - u^2 - v^2)^2},$$

or, since here $a^2 + b^2 = u^2 + v^2 = \tanh^2(\frac{r}{2})$,

$$\langle x_\theta, x_\theta\rangle = \frac{4\tanh^2(\frac{r}{2})}{(1 - \tanh^2(\frac{r}{2}))^2} = \sinh^2 r.$$

The last equality follows from a well–known relation between hyperbolic functions. (Exercise 2.7.8f.)

Formula (5.7.1) is valid throughout a normal neighborhood of p, except at the point p itself. In order to analyze the behaviour of $G(r, \theta)$ as we approach p — i. e. as $r \to 0$ — we compare with the Cartesian coordinates $u = r\cos\theta$, $v = r\sin\theta$, which also are valid near p. That is, we have a parametrization $y(u, v)$ defined in a neighborhood of 0, such that $x(r, \theta) = y(r\cos\theta, r\sin\theta)$.

In these coordinates, the metric has the form

$$ds^2 = \overline{E}du^2 + 2\overline{F}du\,dv + \overline{G}dv^2,$$

where now $\overline{E}$, $\overline{F}$ and $\overline{G}$ are smooth functions of (u, v). Moreover, at the point p we have $\overline{E} = \overline{G} = 1$ and $\overline{F} = 0$. Now use the relation

$$x_\theta = \frac{\partial u}{\partial\theta}y_u + \frac{\partial v}{\partial\theta}y_v = -r\sin\theta\, y_u + r\cos\theta\, y_v$$

to obtain

$$G = \langle x_\theta, x_\theta\rangle = r^2(\sin^2\theta\,\overline{E} - 2\sin\theta\cos\theta\,\overline{F} + \cos^2\theta\,\overline{G}) = r^2 L\,,$$

where $L = L(r,\theta)$ is a smooth function such that $L(0,\theta) = 1$. Hence we can write $G = h^2$, where $h(r,\theta)$ has an expansion

$$h(r,\theta) = r + a(\theta)r^2 + b(\theta)r^3 + r^3\mathcal{O}(r). \tag{5.7.2}$$

(Recall that $\mathcal{O}(r)$ is a generic notation for a term which for small r is bounded by a constant times r. Note also that $|h(r,\theta)| = |x_\theta(r,\theta)|$.)

At this point we recall the curvature formula from our proof of *Theorema egregium.* With $ds^2 = dr^2 + h^2 d\theta^2$ we obtain, after a little calculation

$$K = -\frac{h_{rr}}{h}\,, \tag{5.7.3}$$

which with the expression (5.7.2) for h yields

$$K = -\frac{2a(\theta) + 6b(\theta)r + r\mathcal{O}(r)}{r + a(\theta)r^2 + b(\theta)r^3 + r^3\mathcal{O}(r)}\,.$$

This formula is valid for $r \neq 0$, but since Gaussian curvature is continuous, this expression must approach $K(p)$ as $r \to 0$. This is only possible if $a(\theta) = 0$ and $6b(\theta) = -K(p)$ for all θ. Hence (5.7.2) reduces to

$$h(r,\theta) = r - \frac{K(p)}{6}r^3 + r^3\mathcal{O}(r)\,. \tag{5.7.4}$$

An immediate consequence is the following formula for Gaussian curvature:

$$K(p) = \lim_{r\to 0} 6\frac{r - |x_\theta|}{r^3}.$$

An even more interesting formula is obtained from the following computation of the circumference of a geodesic circle of radius ρ and center p. Such a circle is parametrized by $\beta_\rho(\theta) = x(\rho,\theta)$, $\theta \in [0, 2\pi]$, with geodesic coordinates as above. Then the circumference is

$$\begin{aligned} l_\rho(p) = l(\beta_\rho) &= \int_0^{2\pi} \|\beta_\rho'(\theta)\|\,d\theta = \int_0^{2\pi} h\,d\theta \\ &= \int_0^{2\pi} (\rho - \frac{K(p)}{6}\rho^3 + \cdots)\,dt = 2\pi\rho - \frac{K(p)\pi}{3}\rho^3 + \cdots\,. \end{aligned}$$

Corollary 5.7.4.

$$K(p) = \lim_{\rho\to 0} \frac{3(2\pi\rho - l_\rho(p))}{\pi\rho^3}.$$

Example 5.7.5. A hyperbolic circle of radius ρ has circumference $2\pi \sinh \rho$. Hence

$$K(p) = \lim_{\rho \to 0} \frac{3(2\pi\rho - 2\pi \sinh \rho)}{\pi \rho^3} = \lim_{\rho \to 0} \frac{6(\rho - (\rho + \frac{\rho^3}{6} + \cdots))}{\rho^3} = -1$$

for every p, as before.

Corollary 5.7.4 also gives an interpretation of the *sign* of the curvature: $K(p)$ is positive if the circumference of a small geodesic circle is smaller than that of a Euclidean circle of the same radius and negative if it is bigger.

Exercise 1 gives a similar result comparing *areas* of circles.

We now show that geodesic polar coordinates also provide the tools we need to prove Theorem 5.6.14, in the following form:

Theorem 5.7.6. *Every point p in a Riemannian surface S has a neighborhood V such that any point $q \in V - \{p\}$ can be connected to p by a unigue shortest curve, and this curve is a geodesic.*

Proof. (We use the same notation as at the beginning of this section.) Let $\epsilon > 0$ such that $\exp_p |B_p(\epsilon)$ is a diffeomorphism, and let $V = \exp_p(B_p(\epsilon))$. Then any $q \in V - \{p\}$ can be written as $q = x(\rho, \theta_0) = \exp_p(\rho\alpha(\theta_0))$ for a unique $\rho \in (0, \epsilon)$ and $\theta \in [0, 2\pi)$, and the curve $\gamma(s) = x(s, \theta_0),\ s \in [0, \rho]$ is a geodesic of length ρ from p to q.

Let $c : [0, a] \to S$ be another curve from p to q. We want to prove that the length of c at least ρ. Then we can assume that $c(t) \neq p$ for every t, since otherwise we can replace c with a shorter curve having this property.

Assume first that $c([0, a]) \subset V$. Then we can write $c(t) = x(r(t), \theta(t))$, for uniquely determined functions $r : [0, a] \to [0, \epsilon)$ and $\theta : [0, a] \to \mathbb{R}$, and where $r(0) = 0$ and $r(a) = \rho$. Then $c'(t) = r'(t)x_r + \theta'(t)x_\theta$, and $|c'(t)|^2 = (r') + (\theta')^2 h^2$. For the arclength of c we then have

$$\ell(c) = \int_0^a |c'(t)|\, dt \geqslant \int_0^a |r'(t)|\, dt \geqslant \int_0^a r'(t)\, dt = r(a) = \rho,$$

with equalities if and only if $\theta(t)$ is constant and $r(t)$ is non-decreasing. But then $c(t)$ is just a reparametrization of the geodesic γ.

Finally, suppose c is not contained in V. Then there is smallest number $t_0 \in [0, a]$ such that $c(t_0) \notin V$, and the argument above proves that the length of $c|[0, t_0]$ is at least ϵ, which is bigger than ρ. Hence c must certainly again be longer than γ.

□

Exercises for 5.7

1. Let A_ρ be the area of a geodesic circle of radius ρ and center p. Prove that
$$K(p) = \lim_{\rho \to 0} \frac{12(\pi\rho^2 - A_\rho)}{\pi\rho^4} \,.$$
Use this to give yet another proof that the curvature of the hyperbolic plane is constant equal to -1.

2. Let $ds^2 = dr^2 + G(r,\theta)d\theta^2$ be the metric in geodesic polar coordinates around the point $p \in S$. Now rescale the metric by a factor $\lambda = c^2$ as in Exercise 5.4.6, and show that in the new geodesic polar coordinates the metric is
$$ds^2 = dr^2 + c^2 G(\frac{r}{c}, \theta)d\theta^2 \,.$$

3. Discuss the relationship between formula (5.7.3) and the formula in exercise 5.5.3.

5.8 Riemannian surfaces of constant curvature

We now have the tools necessary for proving one of our main results: the characterization of geometries locally isometric to the classical geometries.

Suppose given a function K defined in a neighborhood of $p \in S$. If we want to find a local Riemannian metric with curvature function K, we have to find a solution of equation (5.7.3) of the form (5.7.4).

We are interested in the simplest case, namely the case when K is *constant.* Our calculations have shown that surfaces built on the classical geometries have this property, and if you did Exercise 5.5.5, you know that metrics obtained from these by *scaling* have the same property. We want to prove that these are the only ones. This relies on the observation in Proposition 5.4.3, saying that if two surfaces can be parametrized by the same subset of $\mathbb{R}^2$ in such a way that the functions E, F and G are the same, then they are locally isometric. Thus it suffices to show that a point of a surface of constant curvature has a neighborhood such that the metric in geodesic polar coordinates looks like one of the metrics in Example 5.7.3, possibly scaled by a constant as in Exercise 5.7.2.

So, consider the equation (5.7.3), or equivalently

$$h_{rr} + Kh = 0, \tag{5.8.1}$$

for K constant. The general solution of this equation is well known, and we distinguish between the three cases $K = 0$, $K > 0$ and $K < 0$.

(i) $K = 0$: The equation reduces to $h_{rr} = 0$, which has the solutions $h = Ar + B$, where A and B are functions of only θ. But the condition (5.7.4) gives $B = 0$ og $A = 1$. Hence the metric has the form $ds^2 = dr^2 + r^2 d\theta^2$, which is the same as for the Euclidean metric in polar coordinates. Hence a neighborhood of p in S is isometric to a neighborhood of the origin in the Euclidean plane.

(ii) $K = 1/R^2$, $R > 0$: Equation (5.8.1) is now $h_{rr} + \frac{1}{R^2}h = 0$, with the general solution

$$h = A\cos\frac{r}{R} + B\sin\frac{r}{R} = A\left(1 - \frac{1}{2}\left(\frac{r}{R}\right)^2 + \cdots\right) + B\left(\frac{r}{R} - \frac{1}{6}\left(\frac{r}{R}\right)^3 + \cdots\right).$$

Condition 5.7.4 gives $A = 0$ and $B = R$, and the metric becomes

$$ds^2 = dr^2 + R^2\sin^2(\frac{r}{R})\, d\theta^2 .$$

This is, by Example 5.7.3 and Exercise 5.7.2, the metric on the sphere S^2_R of radius R. It follows that S must be locally isometric to this sphere.

(iii) $K = -\frac{1}{\rho^2}$, $\rho > 0$: Now (5.8.1) becomes $h_{rr} - \frac{1}{\rho^2}h = 0$. The general solution is

$$h = Ae^{\frac{r}{\rho}} + Be^{-\frac{r}{\rho}} = A\left(1 + \frac{r}{\rho} + \frac{1}{2}\left(\frac{r}{\rho}\right)^2 + \cdots\right) + B\left(1 - \frac{r}{\rho} + \frac{1}{2}\left(\frac{r}{\rho}\right)^2 + \cdots\right).$$

Condition 5.7.4 give $A = \frac{\rho}{2}$ and $B = -\frac{\rho}{2}$. Consequently, $h = \rho\sinh\frac{r}{\rho}$ and the metric is

$$ds^2 = dr^2 + \rho^2\sinh^2\frac{r}{\rho}d\theta^2 . \tag{5.8.2}$$

But this is the metric of $\mathbb{D}$ in geodesic polar coordinates, scaled by the constant ρ.

Let us denote this scaled version of $\mathbb{D}$ by $\mathbb{D}_\rho$. We have now proved

Theorem 5.8.1. *Suppose S is a Riemannian surfaces such that the Gaussian curvature is constant. Then*

- *If $K = 0$, S is locally isometric to the Euclidean plane.*
- *If $K = 1/R^2$, S is locally isometric to a sphere of radius R.*
- *If $K = -1/\rho^2$, S is locally isometric to the hyperbolic plane $\mathbb{D}_\rho$.*

Remark 5.8.2. Clearly there is a scaled version $\mathbb{H}_\rho$ as well, and $\mathbb{H}_\rho$ is isometric to $\mathbb{D}_\rho$.

Exercises for 5.8

1. Show that $G(z) = \dfrac{iz+1}{z+i}$ is an isometry between $\mathbb{H}_\rho$ and $\mathbb{D}_\rho$ as Riemannian surfaces for every ρ.
2. Show that if the surface S has a geometric structure modeled on $\mathbb{H}$, then it also has one modeled on $\mathbb{H}_\rho$ for every $\rho > 0$.

5.9 The Gauss–Bonnet theorem

Until now, our study of differential geometry has been a study of *local* properties. In this last section we shall prove the celebrated *Gauss–Bonnet* theorem, relating local, geometric information to *global*, topological properties of surfaces. The theorem has a number of striking consequences, and it can be seen as the first example of an *index theorem* — a formula relating topological and analytic invariants.

Before we state the theorem, we need some more definitions. First, we will need to define integrals of smooth functions over curves and surface regions, generalizing the usual line and surface integrals in $\mathbb{R}^3$ to arbitrary Riemannian surfaces.

Line integrals. If f is a function defined on a curve $\mathcal{C}$ parametrized by a function $\alpha : [a, b] \to S$ with $\|\alpha'(t)\| \neq 0$ for all t, we define

$$\int_{\mathcal{C}} f ds = \int_a^b f(\alpha(t)) \frac{ds}{dt} dt = \int_a^b f(\alpha(t)) \|\alpha'(t)\|_{\alpha(t)} dt. \qquad (5.9.1)$$

It is not difficult to check that this is independent of the parametrization. If $\mathcal{C}$ is only piecewise smooth, we define the integral as the sum of integrals over the smooth pieces.

Surface integrals. Suppose f is defined on a compact region R bounded by a piecewise smooth curve on a Riemannian surface. If $R \subset x(\mathcal{U})$ for a parametrization x, we define

$$\iint_R f dA = \iint_{x^{-1}(R)} f(x(u,v))\sqrt{EG - F^2} dudv. \tag{5.9.2}$$

By exercise 1. this is independent of choice of parametrizations. If R is not included in a coordinate neighborhood, we can subdivide it into smaller pieces which are, and define the integral as a sum. This will be independent of the subdivision; cf. the discussion of area in section 5.3. In particular, we can define the integral over the whole surface S if it is compact.

(These integrals can, of course, also be defined directly as analogues of Riemann integrals ("limits of sums"). Note also that the formulas for arc length and area in Section 5.3 are special cases.)

The other new ingredient we need is *geodesic curvature.* Fix a Riemannian surface S, and assume that $\alpha(s)$ is a regular curve parametrized by arc length, with unit tangent vector $T(s) = \alpha'(s)$. For every s the curve has two possible unit normal vectors $n_\alpha(s)$ at $\alpha(s)$. A continuous choice of one of them we will call a *normal orientation* of the curve. (*Continuous* means that locally we can write $n_\alpha = \phi x_u + \psi x_v$, where ϕ and ψ are continuous functions.)

Definition 5.9.1. *Let $\alpha(s)$ be a normally oriented curve parametrized by arc length. The* geodesic curvature *is defined as the normal component of $D\alpha''(t)$:*

$$k_g(s) = \langle D\alpha''(s), n_\alpha(s)\rangle. \tag{5.9.3}$$

It follows from formula (5.6.3) that for curves on regular surfaces in $\mathbb{R}^3$ we have

$$D\alpha''(s) = k_g(s)\, n_\alpha(s). \tag{5.9.4}$$

One can show that (5.6.3) holds in general, but we will see that (5.9.4) is also a consequence of calculations in the proof of Gauss–Bonnet given below. It then follows that the curve is a geodesic if and only if $k_g(s) = 0$ for all s. Hence the function k_g can be thought of as a measure of how far the curve is from being a geodesic.

The choice of normal orientation only affects the sign of k_g. There are two situations where a normal orientation is naturally defined.

(i) Usually one considers only the case when the surface S is oriented, and then the normal vector n_α is chosen such that $(T(s), n_\alpha(s))$ is a positively oriented (orthonormal) basis for $T_{\alpha(s)}S$. For example, if S is a regular surface in $\mathbb{R}^3$ and N is the surface normal defining the orientation, we can set $n_\alpha(s) = N(\alpha(s)) \times T(s)$.

(ii) In the situation we will consider, however, the surface is not necessarily oriented, but the curves will all be regular boundary curves of regions $R \subset S$. Then we can choose n_α to be the unit normal vector pointing *into* R — i. e. each $n\alpha(s)$ has the form $\beta'(0)$, for some curve such that $\beta(t) \in R$ for $t > 0$. Note that for this definition of normal orientation, the direction of the boundary curve does not matter.

If $R \subset S$ is as in (ii) and S is oriented, the two orientations agree if we choose to traverse ∂R *counterclockwise* around R.

Let now S be a connected Riemannian surface, and suppose $R \subset S$ is a compact region bounded by a finite union of regular curves. Denote the boundary by ∂R. Note that ∂R can have many components, and each component is a piecewise smooth, closed curve. We are now in situation (ii) above, and each smooth piece is normally oriented by the inward normal vector.

At a non–smooth point p_i, R has a well defined interior angle $\eta_i \in [0, 2\pi]$, and we say that the boundary *changes direction by the angle* $\epsilon_i = \pi - \eta_i \in [-\pi, \pi]$ at p_i. Note that if S is oriented and we follow ∂R counterclockwise, the angle is positive if we turn to the left and negative is we turn to the right.

We call p_i a *cusp point* if ϵ_i is $\pm\pi$. This means that the two smooth boundary curves meeting at p_i have the same tangent there.

The Gauss–Bonnet theorem is the following formula:

Theorem 5.9.2. (Gauss–Bonnet)

$$\iint_R K dA + \int_{\partial R} k_g ds + \sum_i \epsilon_i = 2\pi\chi(R). \tag{5.9.5}$$

Here K and k_g are the Gaussian and geodesic curvatures, respectively, and $\chi(R)$ is the Euler characteristic of R. (Chapter 3.) The index i runs over all the non–smooth points of ∂R. The striking feature of this formula is that the left hand side depends entirely on geometric information, whereas the right hand side is purely topological. For example, If S is compact, we can apply the theorem to $R = S$, and the formula simplifies to

$$\iint_S K dA = 2\pi\chi(S). \tag{5.9.6}$$

Since the Euler characteristic determines S up to homeomorphism, it follows that the curvature function also determines S up to homeomorphism!

Here are some more immediate applications:

- For any metric on S^2 or P^2 there must be points where the curvature is positive.
- On orientable surfaces of genus at least 2, there has to be points where the curvature is negative.
- On a torus or a Klein bottle the curvature either vanishes everywhere, or it must take both positive and negative values.

In particular, The only compact surfaces that allow metrics of constant positive curvature are S^2 or P^2, the only ones with constant zero curvature are the torus and the Klein bottle, and neither of these can have metrics of constant negative curvature.

These are just a few of the many consequences of the theorem; more will follow after the proof, which will occupy the next five pages.

Proof of the Gauss–Bonnet theorem. Assume first that no ϵ_i is $\pm\pi$, i. e. ∂R has no *cusp point.* We take as a fact that any such region R can be *smoothly triangulated,* i. e. it can be written as a union of smooth, embedded triangles, where the intersection of two distinct triangles (if non–empty) is either a common side or a vertex. (Cf. Exercise 5.1.10. An *embedding* is a regular, injective map.) In fact, by further subdividing the triangles, if necessary, we may assume that every triangle is contained in a *normal neighborhood*, i. e. a neighborhood parametrized by geodesic polar coordinates, centered at a point in the interior of some triangle.

We shall see that the general case follows if we can prove the theorem for each such triangle.

Thus, let $R \subset W_p$ be the embedded image of a triangle, where W_p is parametrized by geodesic coordinates $x(r,\theta)$, $(r,\theta) \in [0,\rho) \times \mathbb{R}$, with center $p = x(0,\theta) \notin \partial R$. Although S is not assumed to be orientable, W_p is, and we now orient it such that (x_r, x_θ) is a positively oriented basis outside p. (This orientation extends uniquely also to p, by declaring that e. g. $(x_r(0,0), x_r(0,\pi/2))$ should be a positively oriented basis there.)

The boundary ∂R is the union of three smooth curves, and we choose parametrizations by arc length $\alpha_1(s)$, $\alpha_2(s)$ and $\alpha_3(s)$, traversing ∂R in counterclockwise direction, with respect to the orientation chosen on W_p. We may assume that if the length of α_i is l_i, then $\alpha_1(l_1) = \alpha_2(0)$, etc.

It also follows that $(\alpha_i'(l_i), \alpha_{i+1}'(0))$ is a positively oriented basis for $T_{\alpha_i(l_i)}S = T_{\alpha_i(l_i)}V_p$. Let $\epsilon_i \in (0, \pi)$ be the angle between $\alpha_i'(l_i)$ and $\alpha_{i+1}'(0)$, where i is counted mod 3. (Cf. also exercise 5.3.5.)

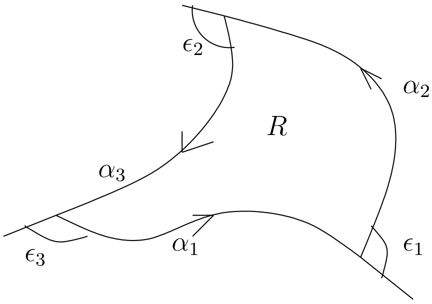


Fig. 5.9.1:

We have seen that Gaussian curvature in geodesic polar coordinates is given by $K = -\dfrac{h_{rr}}{h}$, where $ds^2 = dr^2 + h^2 d\theta^2$ is the metric. Let us now calculate the geodesic curvature k_g for the boundary curves using these coordinates.

Write the parametrization of the curve as $\alpha(s) = x(r(s), \theta(s))$. Then $\alpha' = r'x_r + \theta' x_\theta$ has norm 1. Moreover, $\|x_r\| = 1$ and $\|x_\theta\| = h$, hence $\alpha' = r'x_r + h\theta' \dfrac{x_\theta}{h}$ is α' expressed in an orthonormal basis. It follows that there is an angle $\phi(s)$, uniquely determined mod 2π, such that $r'(s) = \cos\phi(s)$ and $h\theta' = \sin\phi(s)$. $\phi(s)$ is the angle between the geodesic radius through $\alpha(s)$ and the curve, and it defines a smooth function of s, normalized e. g. by choosing $\phi(0) \in (-\pi, \pi]$. Moreover, the inward pointing unit normal n_α is then determined as

$$n_\alpha = -\sin\phi\, x_r + \cos\phi \frac{x_\theta}{h}\,.$$

(α' rotated $\pi/2$ in the positive direction.)

The Christoffel symbols for the metric $ds^2 = dr^2 + h^2 d\theta^2$ are computed from equation (5.6.8:

$$\begin{bmatrix} 1 & 0 \\ 0 & h^2 \end{bmatrix} \begin{bmatrix} \Gamma_{11}^1 & \Gamma_{12}^1 & \Gamma_{22}^1 \\ \Gamma_{11}^2 & \Gamma_{12}^2 & \Gamma_{22}^2 \end{bmatrix} = \begin{bmatrix} 0 & 0 & -hh_r \\ 0 & hh_r & hh_\theta \end{bmatrix}\,.$$

Formula (5.6.9) for the covariant second derivative now gives:

$$\begin{aligned} D\alpha''(s) &= (r'' - hh_r\theta'^2)\, x_r + (\theta'' + 2\,\frac{h_r}{h}r'\theta' + \frac{h_\theta}{h}\theta'^2)\, x_\theta \\ &= ((r')' - h_r\theta'(h\theta'))\, x_r + ((h\theta')' + h_r r'\theta')\frac{x_\theta}{h} \qquad (5.9.7) \\ &= (-\sin\phi\,\phi' - h_r\theta'\sin\phi)\, x_r + (\cos\phi\,\phi' + h_r\theta'\cos\phi)\frac{x_\theta}{h} \\ &= (\phi' + h_r\theta')\, n_\alpha(s). \end{aligned}$$

Thus we have the following expression for the geodesic curvature:

$$k_g(s) = \phi' + h_r\theta'. \qquad (5.9.8)$$

We now compute the integral

$$\int_{\partial R} k_g ds = \int_{\partial R} \phi'\, ds + \int_{\partial R} h_r\theta' ds.$$

(Strictly speaking, we should replace ∂R by the corresponding parameter set on the right hand side, but we choose to simplify the notation, believing that it is clear what we mean. This remark also applies at several points later.)

Consider first the term $\int_{\partial R}\phi' ds$:

$$\begin{aligned} \int_{\partial R}\phi' ds &= \int_{\alpha_1}\phi_1' ds + \int_{\alpha_2}\phi_2' ds + \int_{\alpha_3}\phi_3' ds \\ &= (\phi_1(l_1) - \phi_1(0)) + (\phi_2(l_2) - \phi_2(0)) + (\phi_3(l_3) - \phi_3(0)). \end{aligned}$$

The function ϕ_i is determined up to a constant multiple of 2π on each α_i, uniquely determined by $\phi_i(0)$. If we fix $\phi_1(0)$, then ϕ_2 and ϕ_3 are determined by $\phi_2(0) = \phi_1(l_1) + \epsilon_1$ and $\phi_3(0) = \phi_2(l_2) + \epsilon_2$. But coming back to $\alpha_3(l_3) = \alpha_1(0)$ we can only say that $\phi_1(0) \equiv \phi_3(l_3) + \epsilon_3 \,(\mathrm{mod}\, 2\pi)$. Hence

$$\int_{\partial R}\phi' ds = 2k\pi - \epsilon_1 - \epsilon_2 - \epsilon_3$$

for some integer k.

We state the following result without proof:

Claim: $k = 0$ if $p \in \operatorname{int} R$ and $k = 1$ if $p \notin R$. ("Hopf's Umlaufsatz".)

This may sound obvious, but the Claim is not at all trivial. Here is an indication of how it can be proved. The idea is to use the fact that $k = (\int_{\partial R}\phi' ds +$

$\epsilon_1 + \epsilon_2 + \epsilon_3)/2\pi$ is an integer — hence will remain constant under continuous modifications of the terms involved.

We apply two kinds of such modifications. First consider R as lying in $\mathbb{R}^2 \approx T_pS$ (via $\exp_p^{-1}$) with the metric $ds^2 = dr^2 + h^2 d\theta^2$, and deform the *metric* via $ds^2 = dr^2 + (tr^2 + (1-t))h^2 d\theta^2$, $t \in [0, 1]$ to the standard metric $ds^2 = dr^2 + r^2\theta^2$.

Then use the fact that ∂R is the boundary of an embedded triangle to deform to the case where R is a very small Euclidean triangle. (p may remain inside or outside R throughout this deformation.) But in this situation the claim is easy to verify.

Next we calculate the term $\int_{\partial R} h_r\theta' ds$. We must again distinguish between the two cases $p \notin R$ and $p \in R$ — i. e. the case when R is contained in the image of the parametrization x and the case when it is not.

Case 1: $p \notin R$. Then Green's theorem gives

$$\int_{\partial R} h_r\theta' ds = \int_{\partial R} h_r d\theta = \iint_R h_{rr} dr\, d\theta = \iint_R \frac{h_{rr}}{h} h\, dr\, d\theta = -\iint_R K dA.$$

Case 2: $p \in R$. Let R_δ be R minus the interior of a small geodesic disk of radius δ around p. Then R_δ is contained in the parametrized region, and Green's theorem now gives

$$\int_{\partial R} h_r\theta' ds = -\iint_{R_\delta} K dA + \int_{r=\delta} h_r\theta' ds\,.$$

Using formula (5.7.2) we see that the last line integral is 2π plus terms containing the factor δ^2. The curvature function K is defined and continuous in all of R. Therefore, taking limits as δ goes to 0, we get

$$\int_{\partial R} h_r\theta' ds = -\iint_R K dA + 2\pi.$$

Putting all this together shows that in both cases we have

$$\int_{\partial R} k_g ds = 2\pi - \iint_R K\, dA - \epsilon_1 - \epsilon_2 - \epsilon_3,$$

which completes the proof in the case where R is an embedded triangle contained in a normal neighborhood. Note that although we used an auxilliary orientation of R in the proof, the result is independent of which orientation we choose, since all the terms can be defined without an orientation.

In the general case (but still no cusp) we can triangulate R as $\cup_j R_j$, where each R_j has this form. Let the *interior* angles of R_j be $\eta_{ji} = \pi -$

ϵ_{ji}, $i = 1, 2, 3$. Then we can also write Gauss–Bonnet's theorem for R_j as

$$\iint_{R_j} K dA + \int_{\partial R_j} k_g ds = \eta_{j1} + \eta_{j2} + \eta_{j3} - \pi.$$

Summing over all R_j, we get

$$\iint_{R} K dA + \sum_j \int_{\partial R_j} k_g ds = \sum_{j,i} \eta_{ji} - T\pi\,, \tag{5.9.9}$$

where T is the total number of triangles. Denote the three smooth boundary curves (the "edges") of R_j by α_{ji}, $i = 1, 2, 3$, all normally oriented by our convention. The edges lying in the *interior* of R will then come in pairs with opposite normal orientations. Hence the geodesic curvatures have opposite signs, and the corresponding terms cancel in

$$\sum_j \int_{\partial R_j} k_g ds = \sum_{j,i} \int_{\alpha_{ji}} k_g ds.$$

The remaining edges are the ones in ∂R, and we have

$$\sum_j \int_{\partial R_j} k_g ds = \sum_{\alpha_{ji} \subset \partial R} \int_{\alpha_{ji}} k_g ds = \int_{\partial R} k_g ds.$$

It remains to analyze the right hand side of (5.9.9). Observe first that the sum of all the angles η_{ji} around an *internal* vertex of R is 2π, so summing all such angles, we get $2\pi(V - V_\partial)$, where V and V_∂ denote the numbers of vertices in R and ∂R.

The remaining angles η_{ji} are interior angles at vertices in ∂R. Note that each non–smooth point must be such a vertex, and there the angles sum up to $\pi - \epsilon_k$, where this defines the *external angle* $\epsilon_k \in (-\pi, \pi)$ at this vertex. At each smooth point the angles η_{ji} sum up to π.

It follows that the sum of all η_{ji} at all vertices in ∂R is $V_\partial\, \pi - \sum_k \epsilon_k$.

Let E be the total number of edges in R and E_∂ the number of edges in ∂R. Then clearly

$$E_\partial = V_\partial.$$

Since every triangle has three edges, of which interior edges lie in two triangles and boundary edges lie in one, we also have the relation

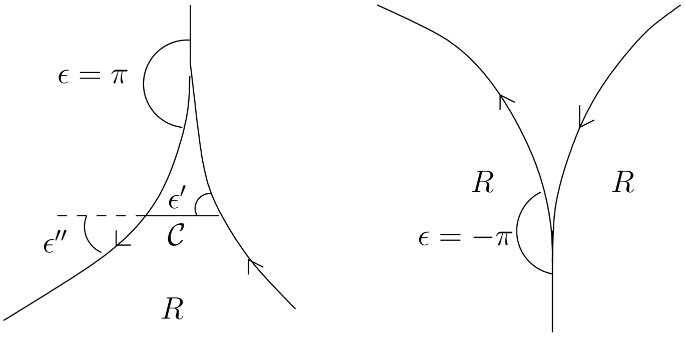
$$3T = 2E - E_\partial.$$

Substituting this, we obtain

$$\begin{aligned}\sum_{j,i}\eta_{ji}-T\pi &= 2\pi(V-V_\partial)+V_\partial\,\pi-\sum_k\epsilon_k-T\pi\\ &= 2\pi V-\pi V_\partial-\pi T-\sum_k\epsilon_k\\ &= 2\pi V-\pi E_\partial-\pi T-\sum_k\epsilon_k\\ &= 2\pi V-\pi(2E-3T)-\pi T-\sum_k\epsilon_k\\ &= 2\pi\chi(R)-\sum_k\epsilon_k\,.\end{aligned}$$

This proves Gauss–Bonnet for regions without any cusp points along the boundary, which suffices for most applications. But the theorem is also valid in the cusp case. In fact, the above proof applies as it stands if there are only *inward* cusps (cf. Fig. 5.9.2), since then R can be be triangulated as before. However, the case of outward cusps needs slightly more care.

We refer to Fig. 5.9.2a. Near the cusp point, ∂R looks like two curves becoming tangent at q. Now cut $\mathbb{R}$ along a curve curve $\mathcal{C}$ connecting the two curves near the cusp and remove the part containing the cusp. (The upper part in Fig. 5.9.2.)



a. Outward cusp b. Inward cusp

Fig. 5.9.2:

Do this for all the cusps and call the result R'. Then Gauss–Bonnet applies to R', to give

$$\int_{R'} K dA + \int_{\partial R'} k_g ds = 2\pi\chi(R') - \sum_{i\in I_o} \epsilon_i - \sum_{j\in I_c} (\epsilon'_j + \epsilon''_j)\,,$$

where the index set I_c corresponds to the outward cusp points and I_o to the other corners. For each cusp point $\partial R'$ has two corners with angles ϵ' and ϵ'', and as we let the curve $\mathcal{C}$ move toward the cusp point, we can arrange for the sum $\epsilon' + \epsilon''$ to tend to π. In the limit we get formula (5.9.5), where now $\epsilon_i = \pi$. □

We end with some more examples of applications of the Gauss–Bonnet theorem.

(1) Area formulas for geodesic triangles:

Let R be a triangle with geodesic sides and interior angles α, β and γ. Then $\iint_R K dA = \alpha + \beta + \gamma - \pi$. If K is constant, we get

$$KA(R) = \alpha + \beta + \gamma - \pi\,,$$

which gives the well–known area formulas in hyperbolic and spherical geometry. In Euclidean geometry it reduces to the relation $\alpha + \beta + \gamma = \pi$.

(2) A closed geodesic cannot bound a disk on a surface with curvature less than or equal to 0. For along such a curve $k_g = 0$, so if it bounds a region R homeomorphic to a disk, we would have $\iint_R K dA = 2\pi - \epsilon$, where $\epsilon \le \pi$. But then K cannot be non-positive on R.

(3) Similarly, on a surface of curvature ≤ 0 two geodesics cannot meet at two points so as to bound a region R homeomorphic to a disk, because we then would have $\iint_R K dA = 2\pi - \epsilon_1 - \epsilon_2$ for two angles $\epsilon_i < \pi$. This means, for example, that in a neighborhood parametrized by a disk, there can be only one geodesic joining two points.

(4) Two simple, closed geodesics on a compact surface S of everywhere positive curvature must have a point of intersection.

In fact, by Gauss–Bonnet, $\chi(S) > 0$, so $S \approx S^2$. If γ_1 and γ_2 are two geodesics which do not intersect, they must bound a region $R \approx S^1 \times I$, which has Euler characteristic 0. Hence $0 = 2\pi\chi(R) = \iint_R K > 0$, which is a contradiction.

(5) As an example of a completely different kind of application we get a new proof that the Euler characteristic is independent of triangulation. Note that there is no relation at all between triangulations and metrics locally.

Exercises for 5.9

1. Show that the definition of surface integrals is independent of choice of parametrization. (Use Exercise 5.3.4.)

2. Find a formula for the geodesic curvature in terms of a parametrization $\beta(t)$ which is not necessarily by arc length.

3. Verify "Hopf's Umlaufsatz" for a Euclidean triangle in $\mathbb{R}^2$.

4. Generalize the area formulas in the constant curvature cases to arbitrary geodesic n–gons.

5. Show that a compact, oriented Riemannian surface S with curvature everywhere 0 is homeomorphic to a torus.

   Show that if $K \geq 0$, but not 0 everywhere, then S is homeomorphic to a sphere.

6. Suppose S is a compact, oriented surface with constant curvature $K \neq 0$, and suppose that $S = S_1 \cup S_2$, where $S_1 \cap S_2$ is a simple, closed, geodesic. Show that the ratio between the areas of S_1 and S_2 is a rational number.

   What is this number if $K > 0$?

7. Assume that a surface homeomorphic to $T^2 \# T^2 \# T^2$ has a geometric structure modeled on standard hyperbolic geometry. What is its area?

8. Use the Gauss–Bonnet theorem to compute the geodesic curvature of circles on the standard S^2.

9. Let $\mathcal{C}$ be a hyperbolic circle of (hyperbolic) radius ρ. Explain why the geodesic curvature k_g of $\mathcal{C}$ is constant and depends only on ρ.

   Compute k_g.

10. Let p be a point in a Riemannian surface S. Let $\{\epsilon_n\}$ be a sequence of positive numbers converging to 0, and let $\{\mathcal{C}_n\}$ be a sequence of simple closed curves with no non-smooth points, such that $d(q,p) < \epsilon_n$ for every $q \in \mathcal{C}_n$. Prove that

$$\lim_{n\to\infty} \int_{\mathcal{C}_n} k_g ds = 2\pi.$$

11. Verify the claim in Remark 4.2.2.

12. Let $P \subset \mathbb{R}^3$ be a polyhedron homeomorphic to a compact surface. If v is a vertex P, define its *defect* $d(v)$ to be $2\pi - \alpha(v)$, where $\alpha(v)$ is the sum of the angles around v. (Hence $d(v)$ measures the failure of P to lie in a plane near v.) Define the total defect of P to be $D(P) = \Sigma_v d(v)$, where the sum is over all the vertices of P.

    Show that $D(P) = 2\pi\chi(P)$. (Descartes' theorem.)

Index