Lecture Notes in Modern Geometry

RUI WANG

The content of this note mainly follows John Stillwell's book *geometry of surfaces*.

1 The euclidean plane

1.1 Approaches to euclidean geometry

Our ancestors invented the geometry over euclidean plane. Euclid [300 BC] understood euclidean plane via points, lines and circles. A motivation of Euclid's method was to answer the question that what can be done with ruler and compass only. Euclid's geometry is based on logic deductions from axiom system. (The rigorous axiom system was given by Hilbert [1899].) The proofs are usually tricky and simple but quite isolated from other branches of mathematics.

The viewpoint of modern geometry is to study euclidean plane (and more general, euclidean geometry) using sets and numbers. This idea dates back to Descartes (1596-1650) and is referred as analytic geometry. On one side, this brings an effective way in understanding geometry; on the other side, the intuition from geometry stimulates solutions of problems purely from algebras. (A famous example might be Fermat's last theorem which was solved by Andrew Wiles in 1995 using the most advanced algebraic geometry.) From this point of view, modern geometry successfully makes mathematics as a whole, which is the spirit of the math from 20 century's.

In fact, starting from Euclid's time, people are trying to ask whether one can remove the parallel axiom from the axiom system and set up all results from euclidean geometry. The answer turns to be negative. People found that there are three different types of geometry based on different assumption as replacements for parallel axiom. It was Riemann [1840] who clarified the basic viewpoints and opened the chapter of modern geometry. Riemann's idea basically includes:

- consider points in a n-dimensional space as n-tuple of numbers;
- consider the distance between two points as a distance function;
- introduce the concept of curvature which reflects the geometry of the space.

Different choices of metrics correspond to different geometry. From Riemann's point of view, the eulidean plane corresponds to a curvature zero metric over $\mathbb{R}^2$.

Though in general curvature is defined from point to point, if we add another assumption that the curvature is a constant, we will see that the situation gets much simplified. More concretely, the geometry of spaces now is completely reflected by its isometries. The idea of understanding geometry by studying its isometries dates back to Klein [1872]. In particular, this builds up a bridge between classical euclidean geometry (Euclid's method) and Riemannian geometry of constant curvatures. Our lectures will take this point of view.

1.2 Isometries

Consider the set of pair of real numbers

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

The euclidean distance is a function $d : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ defined as

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

This function represents Pythagorean distance of two points in the plane as what we know from Euclid's method.

Definition 1.1 An euclidean isometry for $\mathbb{R}^2$ is a map $f : \mathbb{R}^2 \to \mathbb{R}^2$ satisfying that for any $(x_1,y_1),(x_2,y_2) \in \mathbb{R}^2$,

$$d(f(x_1,y_1),f(x_2,y_2)) = d((x_1,y_1),(x_2,y_2)).$$

We use $\text{Iso}(\mathbb{R}^2,d)$ to denote the set of all euclidean isometries for $\mathbb{R}^2$.

Example 1.2 We give three important examples of euclidean isometries.

(1) Translation by (α,β).

$$t_{(\alpha,\beta)} : \mathbb{R}^2 \to \mathbb{R}^2, \quad (x,y) \mapsto (x+\alpha, y+\beta).$$

(2) Reflection about x-axis.

$$Rf_x : \mathbb{R}^2 \to \mathbb{R}^2, \quad (x,y) \mapsto (x,-y).$$

(3) Rotation around the origin by θ-angle counter-clockwise.

$$r_{O,\theta} : \mathbb{R}^2 \to \mathbb{R}^2, \quad (x,y) \mapsto (x\cos\theta - y\sin\theta, x\sin\theta + y\cos\theta).$$

Check: These are all isometries.

These isometries lists in previous Example 1.2 have nice representations via complex numbers.

Example 1.3 (1) Translation by $z_0 = \alpha + i\beta$.

$$t_{z_0} : \mathbb{C} \to \mathbb{C}, \quad z \mapsto z + z_0.$$

(2) Reflection about x-axis.

$$Rf_x : \mathbb{C} \to \mathbb{C}, \quad z \mapsto \bar{z}.$$

(3) Rotation around the origin by θ-angle counter-clockwise.

$$r_{O,\theta} : \mathbb{C} \to \mathbb{C}, \quad z \mapsto e^{i\theta}z.$$

Using complex numbers, the euclidean distance can be expressed as

$$d(z_1,z_2) = |z_1 - z_2|.$$

Use it, check again these three maps are isometries.

Exercise 1.4 (1) Assume f and g are two isometries of the euclidean plane. Prove that the composition $g \circ f$ is also an isometry.

(2) Prove that the following two definitions for a line in the euclidean plane are equivalent.

(a) A line in the euclidean plane is a set

$$\{(x,y)\in\mathbb{R}^2|ax+by+c=0\}$$

for some $a,b,c\in\mathbb{R}$ with $a^2+b^2\neq 0$;

(b) A line in the euclidean plane is a set

$$L_{(a_0,b_0),(a_1,b_1)}:=\{(x,y)\in\mathbb{R}^2|d((x,y),(a_0,b_0))=d((x,y),(a_1,b_1))\}$$

for some $(a_0,b_0),(a_1,b_1)\in\mathbb{R}^2$ with $(a_0,b_0)\neq(a_1,b_1)$. Here d denotes the euclidean distance.

(3) Use the second definition in (2) to prove: an isometry maps a line to a line. (Remark: In later lectures, we are going to introduce distance functions other than the euclidean distance. Then the first way of defining a line turns out to be not good any more because a line defined in that way is not preserved under isometries. However, the second definition still makes sense: notice in this definition, we only use the distance function.)

(4) Prove the three isometries given in Example 1.2 are all one-to-one and onto. Find their inverses.

Intuition tells us, not only the reflection about x-axis, a reflection about any line is an isometry; Not only the rotation around the origin, a rotation around any point in $\mathbb{R}^2$ is an isometry.

Example 1.5 (1) Assume $p=(\alpha,\beta)\in\mathbb{R}^2$. Denote by $r_{p,\theta}:\mathbb{R}^2\to\mathbb{R}^2$ the rotation around p by θ-angle counter-clockwise. Then

$$r_{p,\theta}=t_{(\alpha,\beta)}\circ r_{O,\theta}\circ t^{-1}_{(\alpha,\beta)}.$$

(2) Assume L is a line in $\mathbb{R}^2$. Denote by $Rf_L:\mathbb{R}^2\to\mathbb{R}^2$ the reflection about L (How to define a reflection?). Notice that if we map L to x-axis, then the reflection will be the standard one that about x-axis. For this, we need to be a little careful for the following two cases:

- Case: L intersect x-axis at some point $p=(\alpha,0)$ via θ as the angle from the positive direction of x-axis to L. Then we can rotate L to x-axis by $r_{p,-\theta}=r^{-1}_{p,\theta}$, and hence

$$Rf_L=r_{p,\theta}\circ Rf_x\circ r^{-1}_{p,\theta}.$$

Further, we can express $r_{p,\theta}$ using (1).

- Case: L is parallel to x-axis. Assume L can be written as $y=\beta$. Then we can translate L to x-axis via $t_{(0,-\beta)}=t^{-1}_{(0,\beta)}$. Similarly, for this case,

$$Rf_L=t_{(0,\beta)}\circ Rf_x\circ t^{-1}_{(0,\beta)}.$$

We are familiar with these expressions of the form $\phi\circ\psi\circ\phi^{-1}$ which is called conjugation, from linear algebra or more general from group theory. In general, the appearance of this form indicates we are doing some coordinate change.

Exercise 1.6 Represent Rf_L and $r_{p,\theta}$ using $\mathbb{C}$ and check your answers via examples.

From the expressions of Rf_L and $r_{p,\theta}$, we see that they are compositions of translations, reflections about x-axis and rotations around origin. In fact, we are going to prove any euclidean isometry can be written as compositions of these three.

1.3 Reflections

Take two lines L_1, L_2 in the euclidean plane. They either intersect or parallel (i.e. not intersect). Let's first see:

(1) If L_1 intersects L_2 at some point p, then $Rf_{L_2} \circ Rf_{L_1}$ is the rotation around p for the double of the oriented angle from L_1 to L_2.

(2) If $L_1 \cap L_2 = \emptyset$, then $Rf_{L_2} \circ Rf_{L_1}$ is some translation $t_{(\alpha,\beta)}$ with the amount $\sqrt{\alpha^2+\beta^2}$ as double of the oriented distance from L_1 to L_2.

Conversely, we prove the following result.

Theorem 1.7 (1) *Any rotation $r_{p,\theta}$ can be decomposed as*

$$r_{p,\theta} = Rf_{L_2} \circ Rf_{L_1}$$

for any two lines $L_1 \cap L_2 = \{p\}$ with the oriented angle from L_1 to L_2 as $\frac{1}{2}\theta$.

(2) *Any translation $t_{(\alpha,\beta)}$ can be decomposed as*

$$t_{(\alpha,\beta)} = Rf_{L_2} \circ Rf_{L_1}$$

for any two lines $L_1 \cap L_2 = \emptyset$ with the oriented distance from L_1 to L_2 as $\frac{1}{2}\sqrt{\alpha^2+\beta^2}$.

You can definitely check the proof directly by doing some calculation. However, a more geometric way is to follow the scheme:

(1) Prove the results for the simplest cases: $r_{O,\theta}$ and $t_{(0,\beta)}$.

(2) Prove that general cases can be reduced to these simplest cases using conjugation by isometries.
Let's take $t_{(\alpha,\beta)}$ as an example for how this works:

Step1. Show that we can find some isometry f so that

$$t_{(\alpha,\beta)} = f \circ t_{(0,\beta)} \circ f^{-1}.$$

Step2. Using the results for the simplest case $t_{(0,\beta)}$ to write

$$t_{(0,\beta)} = Rf_{L_2} \circ Rf_{L_1}.$$

Step3. Then we have

$$\begin{aligned} t_{(\alpha,\beta)} &= f \circ t_{(0,\beta)} \circ f^{-1} \\ &= f \circ Rf_{L_2} \circ Rf_{L_1} \circ f^{-1} \\ &= (f \circ Rf_{L_2} \circ f^{-1}) \circ (f \circ Rf_{L_1} \circ f^{-1}) \\ &= Rf_{f(L_2)} \circ Rf_{f(L_1)}. \end{aligned}$$

The rotation case is exactly the same and is left to you to finish.

Exercise 1.8 Prove the set of translations and rotations is closed under composition. (Closed means for any two maps of translations or rotations, their composition is still a translation or a rotation.)

1.4 The three reflections theorem

We have seen that both translation and rotation can be written as compositions of reflections. In this section, we prove an even stronger result that any euclidean isometry can be decomposed into reflections, which makes reflection play an essential role in understanding isometries (we are going to discuss more on this point of view along our lectures). More concretely, let's prove

Theorem 1.9 (The three reflections theorem) *Any euclidean isometry can be written as compositions of one or two or three reflections.*

To prove this theorem, the first difficulty needs to be overcome is we need to give a way to characterize any euclidean isometry. For this, we first show that

Lemma 1.10 *Any a euclidean isometry is uniquely determined by the image of three points which are not in a line.*

Proof Let's take A, B, C in the plane $\mathbb{R}^2$ and assume that they are not in a line. Assume f is an isometry. If there is another isometry f' so that

$$f(A) = f'(A), \quad f(B) = f'(B), \quad f(C) = f'(C).$$

Let's prove $f = f'$.

For this, we need to show for any $p \in \mathbb{R}^2$, we have $f(p) = f'(p)$. Assume this is not the case, i.e., we can find some p so that $f(p) \neq f'(p)$.

Notice that $d(f(z), f(p)) = d(f(z), f'(p))$ for any $z = A$ or B or C. This requires $f(A), f(B), f(C)$ live in the line $L_{f(p),f'(p)}$ determined by $f(p)$ and $f'(p)$ (see the definition of a line in Exercise 1.4). Now we show that this will require A, B, C live in the same line which contradicts with our assumption.

To see this, WLOG we can assume $d(f(A), f(B)) + d(f(B), f(C)) = d(f(A), f(C))$. Then using the fact that f is an isometry, we have

$$d(A,B) + d(B,C) = d(A,C).$$

A moment of calculation tells us, A, B, C must satisfy some linear relation as they all in a set $\{(x,y)|ax+by+c = 0\}$ for some a, b not both zero. This shows they are in a line.

Then we are done with this proof. □

Now with the help of this lemma, we can write any isometry f by A, B, C as three points not in a line and their images A', B', C' under f. Let's show then f can be decomposed into at most reflections.

We still start from the simplest case:

- Case $A = A', B = B', C = C'$: By the proof of Lemma 1.10, this map must be the identity map. By picking any line L, we can write

$$\mathrm{id}_{\mathbb{R}^2} = Rf_L \circ Rf_L.$$

- Case $A = A', B = B', C \neq C'$: By the proof of Lemma 1.10, this map must be the reflection about the line going through A, B. Hence we are done.

- Case $A = A', B \neq B', C \neq C'$: We have two possibilities for this case: Let's denote by $L_{BB'}$ and $L_{CC'}$ the two lines determined by B, B' and C, C' respectively (notation as in the definition of a line in Exercise 1.4). They both go through A.
  (1) If $L_{BB'} = L_{CC'} =: L$, we show this isometry must be the reflection about L.
  (2) If $L_{BB'} \neq L_{CC'}$. We show this isometry is a rotation around A. Recall that we have shown that a rotation can be written as a composition of two reflections, we are done for this case too.
- Case $A \neq A', B \neq B', C \neq C'$: For this case, we first take a translation to map A to A'. Then we reduce it to one of the first three cases we have done. There is only one trouble that we need to take care of in addition. For a translation, we know it can be decomposed into two reflections. In case after translation, we get the case (2) above, then we have four reflections to composite but we want up to three. To resolve this issue, notice that we can take a common line, say L, for both translation and rotation. Then this map is a composition $Rf_{L_2} \circ Rf_L \circ Rf_L \circ Rf_{L_1}$ which is the same as $Rf_{L_2} \circ Rf_{L_1}$ since reflection twice about the same line is just the identity map. Then we are done with the proof of this theorem.

1.5 Orientation-preserving isometries

For any simple loop in $\mathbb{R}^2$ without self intersection point, it divides $\mathbb{R}^2$ into two connected regions. (In fact, this is a nontrivial result called the Jordan curve theorem. But since it matches our intuition and more important in our case, we don't consider general loops, we just accept it without a proof.) Given an isometry f, we have shown in Lemma 1.10 that f is uniquely characterized by A, B, C not in a line and $A' := f(A), B' := f(B), C' := f(C)$. Consider the loop $ABCA$, it divides $\mathbb{R}^2$ into two regions, we call the bounded region the interior of the loop $ABCA$ and denote it by $\mathrm{Int}(ABCA)$. Now imagine we stand on a point of the loop $ABCA$, with our head towards the positive direction of z-axis (x, y, z-axises satisfy the right-hand-rule.), we define the positive direction of the loop $ABCA$ as the direction so that the interior $\mathrm{Int}(ABCA)$ is on our left-hand side. It is either $\overline{ABCA}$ or $\overline{ACBA}$. (Here by the notation $ABCA$, we don't consider the order of A, B, C, A but we use $\overline{ABCA}$ to denote the direction from A to B to C and back to A.)

Then we consider the loop $A'B'C'A'$, and define the orientation for it in the above way. We call f is an orientation-preserving isometry, if $\overline{A'B'C'A'}$ and $\overline{ABCA}$ are both positive or both negative directions; Otherwise, we call f is an orientation-reversing isometry. (In fact, this definition is not complete, because we need to prove that such defined orientation-preserving or orientation-reversing is independent of choices of the points A, B, C. This is left to you as an exercise.)

The following result is a basic property.

Proposition 1.11 (1) *A composition of two orientation-preserving isometries is orientation-preserving;*

(2) *A composition of two orientation-reversing isometries is orientation-preserving;*

(3) *A composition of an orientation-preserving isometry and an orientation-reversing isometry is orientation-reversing.*

Remark 1.12 One may denote an orientation-preserving isometry by 1 and denote an orientation-reversing isometry by -1, then the pattern above is just like $1 \cdot 1 = 1$; $(-1) \cdot (-1) = 1$; $1 \cdot (-1) = -1 = (-1) \cdot 1$. In fact, this is saying that we can construct a surjective group homomorphism from $\mathrm{Iso}(\mathbb{R}^2, d)$ to $\mathbb{Z}_2 \cong (\{\pm 1\}, \cdot)$, which maps orientation-preserving isometries to 1 and maps orientation-reversing isometries to -1. By this, we also see the set of orientation-preserving isometries, which we denote by $\mathrm{Iso}^+(\mathbb{R}^2,d)$ is a normal subgroup of $\mathrm{Iso}(\mathbb{R}^2,d)$ because it is the kernel of this group homomorphism. Using the fundamental theorem of group homomorphism, the quotient group $\mathrm{Iso}(\mathbb{R}^2,d)/\mathrm{Iso}^+(\mathbb{R}^2,d) \cong \mathbb{Z}_2$ and hence the subgroup $\mathrm{Iso}^+(\mathbb{R}^2,d)$ has index 2.

1.6 Classification of euclidean isometries

Now using the three reflections theorem, we prove the classification theorem for euclidean isometries.

For this, we first introduce a terminology called glide reflection. Assume L is a line, a glide reflection is a translation in L direction composite with the reflection Rf_L. Notice that the order of the translation and the reflection doesn't matter, i.e., these two maps commute. It is easy to check that glide reflection is either a reflection about L or three reflections about L, M, N with M, N both perpendicular to L. Now we prove the reverse part: if an isometry is a reflection or a composition of three reflections, then it is a glide reflection.

Clearly, after proving this, we will conclude the following classification result for euclidean isometries for $\mathbb{R}^2$.

Theorem 1.13 *A euclidean isometry of $\mathbb{R}^2$ is either a translation, a rotation or a glide reflection. Moreover, an orientation-preserving isometry is either a translation or a rotation; an orientation-reversing isometry must be a glide reflection.*

The case of reflection is obviously a glide reflection with trivial translation. We now prove that any isometry $Rf_N \circ Rf_M \circ Rf_L$ is a glide reflection.

Proof **Case 1.** $M \cap N = \{p\}$**.** Denote by M' the unique line going through p and perpendicular to L. Denote by $M' \cap L = \{q\}$, which are going to use later. Next we replace N by some N' going through p so that

$$Rf_N \circ Rf_M = Rf_{N'} \circ Rf_{M'}.$$

(Recalling from Theorem 1.7 that such N' is uniquely determined.) It follows $Rf_N \circ Rf_M \circ Rf_L = Rf_{N'} \circ Rf_{M'} \circ Rf_L$.
Now we take L' as the unique line going through q and perpendicular to N'. Take M'' as the unique line going through q so that

$$Rf_{M'} \circ Rf_L = Rf_{M''} \circ Rf_{L'}.$$

This follows

$$Rf_N \circ Rf_M \circ Rf_L = Rf_{N'} \circ Rf_{M'} \circ Rf_L = Rf_{N'} \circ Rf_{M''} \circ Rf_{L'}.$$

In particular, from the construction, we notice that both N' and M'' are perpendicular to L', which makes $Rf_{N'} \circ Rf_{M''} \circ Rf_{L'}$ a glide reflection about L'. We are done with this case.

Case 2. $M \cap N = \emptyset$**.** This one is left to you as a homework problem and here is a hint: Consider two subcases:

(1) $M \cap L = \emptyset$. For this case, these three lines are parallel and their composition is just a reflection.

(2) $M \cap L = \{p\}$. For this case, you can work out a similar construction as Case 1 starting with replacing M by the line M' going through p and perpendicular to N.

□

1.7 The group structure of Iso$(\mathbb{R}^2, d)$

We introduce a terminology called group. A group is a set G together with a binary operation $\cdot$ so that

(1) The binary operation is associative;

(2) There is an identity element;

(3) Every element has an inverse element.

For our case, the set G is taken as Iso$(\mathbb{R}^2, d)$ and the binary operation is taken as composition. With respect to composition, Iso$(\mathbb{R}^2, d)$ is a group and called the euclidean isometry group.

In general a subgroup of a group $(G, \cdot)$ is a subset $H \subset G$ which is closed under the binary operation $\cdot$ and the H with respect to the binary operation restricted to it is a group. A normal subgroup of a group G is a subgroup H so that every element of the form $g^{-1} \cdot h \cdot g \in H$ for each $g \in G$, $h \in H$. We don't want to spend too much time on the abstract definition, but you should have the following examples in your mind.

Example 1.14 (1) Assume L is a line, $\{\mathrm{id}_{\mathbb{R}^2}, Rf_L\}$ is a group with respect to composition. In fact this group is isomorphic to $\mathbb{Z}_2$;

(2) The set of translations $\{t_{(\alpha,\beta)} | (\alpha, \beta) \in \mathbb{R}^2\}$ forms a group with respect to composition. In fact this group is isomorphic to $(\mathbb{C}, +)$.

The group Iso$(\mathbb{R}^2, d)$ has many subgroups:

(1) For any $p \in \mathbb{R}^2$, all rotations around p forms a subgroup, which is isomorphic to $(U(1), \cdot)$;

(2) For any line L in $\mathbb{R}^2$, $\{\mathrm{id}_{\mathbb{R}^2}, Rf_L\}$ forms a subgroup which is isomorphic to $\mathbb{Z}_2$;

(3) All translations form a subgroup which is isomorphic to $(\mathbb{C}, +)$;

(4) All translations and rotations forms a normal subgroup, which is Iso$^+(\mathbb{R}^2, d)$.

In general, for a group G, any subset $S \subset G$ can generate a subgroup of G by including all possible multiplications from elements in S and their inverses. We denote this subgroup as $< S >$. In particular, if S contains only one element, say g, then the subgroup generated by g is a cyclic subgroup with g as a generator. We use $< g >$ to denote this group.

Example 1.15 (1) Take $z_0 \in \mathbb{C}$, the translation t_{z_0} generates a subgroup $< t_{z_0} >$ whose elements are of the forms t_{nz_0}, $n \in \mathbb{Z}$. This subgroup is isomorphic to $\mathbb{Z}$.

(2) The take $\theta = \frac{2\pi}{n}$ for some $n \in \mathbb{Z}^+$. The rotation $r_{p,\theta}$ generates a finite cyclic group which is isomorphic to $\mathbb{Z}_n$.

The following statement is main reason we introduce the concept of group here. Assume Γ is a subgroup of Iso$(\mathbb{R}^2, d)$. Define a relation over $\mathbb{R}^2$ as

$$p \sim_\Gamma q \quad \text{if and only if} \quad q = f(p) \text{ for some } f \in \Gamma.$$

Lemma 1.16 *The relation $\sim_\Gamma$ is an equivalence relation over $\mathbb{R}^2$ whenever Γ is a subgroup of* Iso$(\mathbb{R}^2, d)$*. As a result, the set of equivalence classes $\mathbb{R}^2/ \sim_\Gamma$ is well-defined.*

Proof (1) Reflexive: Take any $p \in \mathbb{R}^2$, notice that $p = \mathrm{id}_{\mathbb{R}^2}(p)$. Since Γ is a subgroup of $\mathrm{Iso}(\mathbb{R}^2, d)$, the identity element $\mathrm{id}_{\mathbb{R}^2} \in \Gamma$, then $p \sim_\Gamma p$.

(2) Symmetric: Assume $p \sim_\Gamma q$. Then we can find some $f \in \Gamma$ so that $q = f(p)$. Since Γ is a subgroup of $\mathrm{Iso}(\mathbb{R}^2, d)$, the inverse map f^{-1} as the inverse of f is also in Γ. We write

$$p = f^{-1}(f(p)) = f^{-1}(q),$$

which shows that $q \sim_\Gamma p$.

(3) Transitive: Assume $p \sim_\Gamma q$ and $q \sim_\Gamma r$. Then we can find some $f, g \in \Gamma$ so that

$$q = f(p), \quad r = g(q).$$

By writing $r = g(f(p)) = (g \circ f)(p)$ and noticing that $g \circ f \in \Gamma$ since Γ is a subgroup of $\mathrm{Iso}(\mathbb{R}^2, d)$, this proves $p \sim_\Gamma r$.

□

To simplify notation, we denote $\mathbb{R}^2/\sim_\Gamma$ by $\mathbb{R}^2/\Gamma$.

In next section, we are going to understand the geometry over $\mathbb{R}^2/\Gamma$ when Γ is "good".

Example 1.17 Take $\Gamma =< t_{(1,0)} >$. As a set, $\mathbb{R}^2/\Gamma$ can be identified with the strip $[0, 1) \times R$. More precisely, every point in this strip corresponds to a representative from a equivalence class.

Our next question is, how to give a natural distance function to this set $\mathbb{R}^2/\Gamma$?

2 Euclidean surfaces

2.1 Metric spaces

Assume X is a set, a function

$$d_X : X \times X \to \mathbb{R}$$

is called a distance function, if d_X satisfies the following three properties:

(1) $d_X(x, y) = d_X(y, x)$;

(2) $d_X(x, y) \geq 0$ for any $x, y \in X$, and $d_X(x, y) = 0$ if and only if $x = y$;

(3) $d_X(x, z) \leq d_X(x, y) + d_X(y, z)$ for any $x, y, z \in X$.

A set with a distance is called a metric space and we use the pair (X, d_X) to denote.

Example 2.1 (1) The euclidean distance is a distance function over $\mathbb{R}^2$.

(2) Similar as for $\mathbb{R}^2$, we can introduce the euclidean distance for $\mathbb{R}^n$ as

$$d((x_1, x_2, \cdots, x_n), (y_1, y_2, \cdots, y_n)) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2}.$$

(3) Consider the previous example $\mathbb{R}^2/\Gamma$ with $\Gamma =< t_{(1,0)} >$. Denote by $[p]$ the equivalence class on p. Define a function

$$d_\Gamma : \mathbb{R}^2/\Gamma \times \mathbb{R}^2/\Gamma \to \mathbb{R}$$

as

$$d_\Gamma([p],[q]) = \min_{p' \in [p], q' \in [q]} d(p', q').$$

In fact, we should take the notation of inf instead of min to define the distance d_Γ in general. The difference between inf and min can be seen from the following example:

$$\inf\{x > 0 | \frac{1}{x}\} = 0$$

but there is no $x > 0$ so that $1/x = 0$. For this case, it is not proper to use notation min. In another word, when we use min, we need to make sure that the minimal value can be taken by some element in the set we are considering. (The same explanation goes to the difference between sup and max.) However, for our current case of $\Gamma =< t_{(1,0)} >$, it is ok to use min, because we can prove that there must be some $p' \in [p]$ and $q' \in [q]$ so that

$$d(p', q') = \inf_{p' \in [p], q' \in [q]} d(p', q'). \tag{1}$$

The proof is left to you as an exercise.
Before we check d_Γ is a distance function by definition, you should convince yourselves that this definition matches our intuition for the geometry of cylinder.

Lemma 2.2 $(\mathbb{R}^2/\Gamma, d_\Gamma)$ *is a metric space.*

Proof We only need to check the three properties for a distance function.

(a) By definition, $d_\Gamma([p],[q]) = \inf_{p' \in [p], q' \in [q]} d(p', q') \geq 0$. Clearly, if $[p] = [q]$, then $d_\Gamma([p],[q]) = 0$. Now, notice that we can take $p' \in [p]$ and $q' \in [q]$ so that

$$d(p', q') = d_\Gamma([p],[q]).$$

Then $d_\Gamma([p],[q]) = 0$ indicates $p' = q'$ and then $[p] = [p'] = [q'] = [q]$.

(b) $d_\Gamma([p],[q]) = \inf_{p' \in [p], q' \in [q]} d(p', q') = \inf_{p' \in [p], q' \in [q]} d(q', p') = d_\Gamma([q],[p])$.

(c) Take three equivalence classes $[p],[q],[r] \in \mathbb{R}^2/\Gamma$, and pick $p' \in [p]$, $r' \in [r]$ so that

$$d_\Gamma([p],[r]) = d(p', r').$$

Notice that we can further take $q', q'' \in [q]$ so that $d_\Gamma([p],[q]) = d(p', q')$ and $d_\Gamma([q],[r]) = d(q'', r')$. (Why?)
Since both q', q'' are in $[q]$, we can find some $t \in \Gamma$ so that $q'' = t(q')$. Then

$$d(q'', r') = d(t(q'), t(t^{-1})(r')) = d(q', t^{-1}(r')).$$

To simplify notation, let's denote $r'' = t^{-1}(r')$. Now we have estimates

$$\begin{aligned} d_\Gamma([p],[r]) &= d(p', r') \leq d(p', r'') \leq d(p', q') + d(q', r'') = d(p', q') + d(q'', r') \\ &= d_\Gamma([p],[q]) + d_\Gamma([q],[r]), \end{aligned}$$

and the proof is done.

□

Definition 2.3 Assume (X_i, d_i), $i = 1, 2$, are two metric spaces. A bijective map $f : X_1 \to X_2$ is called an isometry if it satisfies that

$$d_1(x_1, y_1) = d_2(f(x_1), f(y_1)), \qquad \text{for any } x_1, y_1 \in X_1.$$

Clearly, if f is an isometry, then f^{-1} is also an isometry.

The set of metric spaces has a relation $\sim$ defined by isometries in this way:

$$(X_1, d_1) \sim (X_2, d_2) \quad \text{if and only if there exists some isometry from } X_1 \text{ to } X_2.$$

Check that such relation is an equivalence relation over metric spaces. We say two metric spaces (X_1, d_1) and (X_2, d_2) are isometric, if $(X_1, d_1) \sim (X_2, d_2)$.

Example 2.4 In calculus class, we have learned that a cylinder (of radius 1) can be defined as a surface in $\mathbb{R}^3$

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

The cylinder C posses a distance function defined as the smallest arc length along all paths over C connecting the two points, i.e.,

$$d_C(p, q) = \inf_{\gamma \text{ is smooth and } \gamma(0)=p, \gamma(1)=q, \gamma \subset C} \int_0^1 |\dot{\gamma}(t)| dt.$$

Exercise 2.5 (1) Check d_C is a distance function on C;

(2) Try to prove (C, d_C) and $(\mathbb{R}^2/\Gamma, d_\Gamma)$ with $\Gamma =< t_{(2\pi,0)} >$ are isometric.

Proof For this, we first introduce a bijective map between C and $\mathbb{R}^2/\Gamma$. Denote the map

$$\phi : C \to \mathbb{R}^2, \quad (x, y, z) \mapsto (\arctan \frac{y}{x}, z)$$

and $\bar{\phi} = [\phi] : C \to \mathbb{R}^2/\Gamma$. (The natural domain of the function arctan is $(-\infty, \infty)$ with range $(-\pi, \pi)$. We extend it by defining $\phi(-1, 0) = (-\pi, 0)$.) It is not hard to check the so-defined map $\bar{\phi}$ is a bijection.

For any $p, q \in C$, a smooth path connecting p, q over C can be written as a parametrized curve

$$r(t) = (x(t), y(t), z(t)),$$

with $r(0) = p$, $r(1) = q$ and $x(t)^2 + y(t)^2 = 1$ for any $t \in [0, 1]$. The distance between p and q by definition is the shortest arc length. Hence we don't need to consider all curves connecting p, q but can focus on the ones whose image under ϕ live in some fundamental region with

$$|\phi(r(t_2)) - \phi(r(t_1))| \le \pi \qquad \text{for any } t_1, t_2 \in [0, 1].$$

(You need some argument to show this. Though a rigorous proof is not required, you need to convince yourselves by geometry intuition.) Now WLOG, let's assume $p = (-1, 0, 0)$ and $q = (\cos\theta_0, \sin\theta_0, z_q)$ with $\theta_0 \in [-\pi, 0]$. We calculate the arc length of r for this case using the cylinder coordinates: $\tilde{r}(\theta) = (\cos\theta, \sin\theta, \tilde{z}(\theta))$, $\theta \in [-\pi, \theta_0]$, as follows

$$\begin{aligned} L_{\mathbb{R}^3}(r) &= \int_{-\pi}^{\theta_0} \sqrt{(\cos\theta)'^2 + (\sin\theta)'^2 + \tilde{z}'(\theta)^2}\, d\theta \\ &= \int_{-\pi}^{\theta_0} \sqrt{1 + \tilde{z}'(\theta)^2}\, d\theta \end{aligned}$$

On the other hand, the image of r under ϕ is the parametrized curve $\phi(\tilde{r}(\theta)) = (\theta, \tilde{z}(\theta))$, $\theta \in [-\pi, \theta_0]$, in $\mathbb{R}^2$. Notice that the arc length of this curve is calculated as

$$L_{\mathbb{R}^2}(\phi(\tilde{r})) = \int_{-\pi}^{\theta_0} \sqrt{1 + \tilde{z}'(\theta)^2}\, d\theta.$$

Hence from the calculation above,

$$L_{\mathbb{R}^3}(r) = L_{\mathbb{R}^2}(\phi(\tilde{r})).$$

Now combine with the definition

$$\begin{aligned} d_C(p,q) &:= \inf_r L_{\mathbb{R}^3}(r) \\ &= \inf_r L_{\mathbb{R}^2}(\phi(\tilde{r})) \\ &= d(\phi(p), \phi(q)) \\ &= d_\Gamma(\bar{\phi}(p), \bar{\phi}(q)). \end{aligned}$$

where the last second equality uses the fact that on the euclidean plane $\mathbb{R}^2$, the shortest arc length connecting two points is the line segment and the shortest arc length is the euclidean distance.

This concludes that d_C is a distance function since d_Γ is a distance function as we have shown. Moreover, (C, d_C), $(\mathbb{R}^2/\Gamma, d_\Gamma)$ are isometric via the map $\bar{\phi}$.

□

Exercise 2.6 Assume (X_i, d_i), $i = 1, 2$ are two metric spaces and $\phi : X_1 \to X_2$ is an isometry. Prove that ϕ maps a circle in X_1 to a circle in X_2 with the same circumference. Here we define the circumference of a circle as the supremum of the sum of all line segment connecting adjacent division points on the circle.

2.2 Locally euclidean surfaces

Assume (X, d_X) is a metric space. Take $\epsilon > 0$ and a point $x \in X$. A disk centered at p with radius ϵ is defined as

$$U_{(X,d)}(x;\epsilon) = \{y \in X | d_X(x,y) \leq \epsilon\}.$$

If metric space is the euclidean plane $(\mathbb{R}^2, d)$, we denote such disk by $D(x;\epsilon)$.

When we restrict the metric $U_{(X,d)}(x;\epsilon)$, we obtain a metric space $(U_{(X,d)}(x;\epsilon), d_X)$. For the case in euclidean plane, we call it a euclidean disk.

Definition 2.7 A metric space (S, d_S) is called locally euclidean surface, if for any $p \in S$, there exists some $\epsilon > 0$, so that the ϵ disk centered at p is isometric to some euclidean disk.

Example 2.8 (1) The euclidean plane $(\mathbb{R}^2, d)$ is a locally euclidean plane: For each $p \in \mathbb{R}^2$, we can take $\epsilon = 1$ and the translation t_{-p} is the isometry from $D(p;1)$ to $D(0;1)$.

(2) An open subset in $(\mathbb{R}^2, d)$ is a subset $U \subset \mathbb{R}^2$ so that each $x \in U$, there exists some disk $D(x;\epsilon)$ centered at x inside U. Any open subset U of $\mathbb{R}^2$ with respect to the distance function d is locally euclidean.

(3) $(\mathbb{R}^2/\Gamma, d_\Gamma)$ with Γ generated by a translation is locally euclidean.

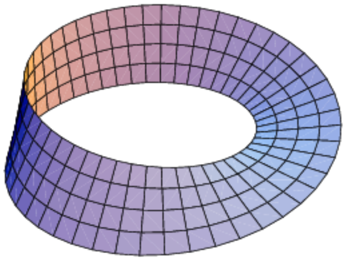
Proof WLOG, we take $\Gamma =< t_{(1,0)} >$. Pick any point $[p] \in \mathbb{R}^2/\Gamma$ and $p' = (x_0, y_0) \in [p]$. Notice that $U_{(\mathbb{R}^2/\Gamma, d_\Gamma)}(p'; \frac{1}{4})$ is isometric to the euclidean disk centered at (x_0, y_0) with radius $\frac{1}{4}$ by the definition of d_Γ. This shows that $\mathbb{R}^2/\Gamma$ is locally euclidean.

□

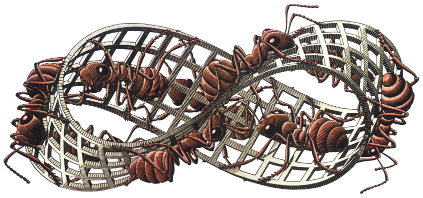
2.3 More examples

In this section, we construct more examples in a similar way as we construct the cylinder $\mathbb{R}^2/<t_{(1,0)}>$.

Example 2.9 The twisted cylinder. Assume f is a proper glide reflection (i.e., not a reflection). $\mathbb{R}^2/<f>$ is a twisted cylinder. It is also a locally euclidean surface. A portion of a twisted cylinder is called a Möbius band whose picture is as given below:



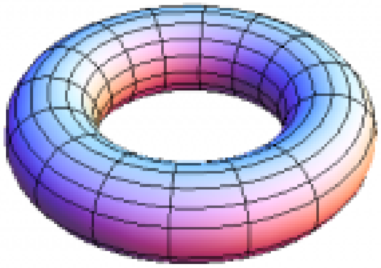
Remark 2.10 (1) A cylinder is orientable while a twisted cylinder is not orientable.



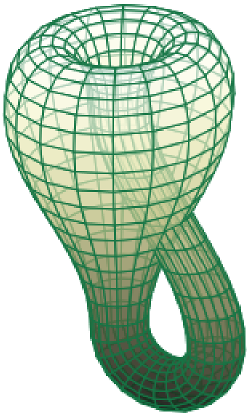
(Over a Möbius band, when the ant comes back to the same point for the first time, its antennae pointing to the opposite direction as it starts off. Viewing the surface in $\mathbb{R}^3$ (which is orientable), this description is equivalent to state the Möbius band is one-sided.)

(2) Both a cylinder and a twisted cylinder are unbounded (not compact).

Example 2.11 (1) The torus. Take $\Gamma =< t_{(1,0)}, t_{(0,1)} >$. Then $\mathbb{R}^2/\Gamma$ is the torus.



(2) The Klein bottle. Take $\Gamma =< t_{(1,0)},f >$, where f is a glide reflection along the y-axis with translation $t_{(0,1)}$. Then $\mathbb{R}^2/\Gamma$ is the Klein bottle.



Remark 2.12 (1) A torus is orientable while a Klein bottle is not orientable. Try to find a Möbius band in a Klein bottle.

(2) Both torus and Klein bottle are compact locally euclidean surfaces.

So far, we have constructed four types of locally euclidean surfaces.

Now let's see two 'bad' examples which are either not a metric space or not locally euclidean.

Example 2.13 (1) Consider the subgroup Γ generated by two translations as follows:

$$\Gamma =< t_{(0,1)}, t_{(0,\sqrt{2})} > .$$

Over the quotient space $\mathbb{R}^2/\Gamma$, we try to define a distance function d_Γ as

$$d_\Gamma([p],[q]) = \inf_{p'\in[p],q'\in[q]} d(p',q').$$

We take two point $[(0,0)], [(0,1)] \in \mathbb{R}^2/\Gamma$. Notice that

$$[(0,0)] = \{(0, m+\sqrt{2}n)|m,n\in\mathbb{Z}\}, \quad [(0,1)] = \{(0, m+\sqrt{2}n+1)|m,n\in\mathbb{Z}\}.$$

Notice that for any $\epsilon > 0$, we can always find $m,n\in\mathbb{Z}$ so that

$$|m+\sqrt{2}n-1| < \epsilon.$$

(In fact, $\{m+\sqrt{2}n|m,n\in\mathbb{Z}\}$ is dense in $\mathbb{R}$.) Hence

$$d_\Gamma([(0,0)],[(0,1)]) = 0$$

but $[(0,0)] \neq [(0,1)]$. This shows that d_Γ is not even a metric.

(2) $\Gamma =< r_{O,\frac{\pi}{4}} >$. ($\mathbb{R}^2/\Gamma$ is an orbifold.) We show now $(\mathbb{R}^2/\Gamma, d_\Gamma)$ is not locally euclidean. Consider the point $[(0,0)] \in \mathbb{R}^2/\Gamma$. Assume there exists some local isometry

$$\phi : \mathbb{R}^2/\Gamma \to \mathbb{R}^2.$$

WLOG, we can assume $\phi([(0,0)]) = (0,0) \in \mathbb{R}^2$. Take a circle γ in $\mathbb{R}^2/\Gamma$ centered at $[(0,0)]$ with radius ϵ. Recalling from Exercise 2.6, $\phi(\gamma)$ should be a circle in $\mathbb{R}^2$ with radius ϵ since ϕ is local isometry and preserves the circumference. The circumference of γ is $\frac{2\pi\epsilon}{8} = \frac{\pi\epsilon}{4}$, while the circumference of $\phi(\gamma)$ is $2\pi\epsilon$. We get contradiction and there is no such local isometry ϕ.
Notice in this example, there exists some point $p \in \mathbb{R}^2$ (the origin for this example), which is a fixed point of the subgroup Γ. In general, when this happens, we can not expect locally euclidean property.

In next section, we abstract some terminology to exclude the phenomenons show up in these two bad examples.

2.4 Conditions on group actions

We have shown that any subgroup Γ of Iso($\mathbb{R}^2, d$) induces a set of equivalent classes, which we denote by $\mathbb{R}^2/\Gamma$. Moreover, there is a natural surjective map

$$\pi : \mathbb{R}^2 \to \mathbb{R}^2/\Gamma, \quad p \mapsto [p].$$

From the viewpoint of group action, each equivalence class $[p]$ is a Γ-orbit through p. Hence the space $\mathbb{R}^2/\Gamma$ consists of (disjoint) orbits of Γ. For this reason, we also call $[p]$ a Γ-orbit through p when we want to emphasize the group Γ-action on $\mathbb{R}^2$.

Definition 2.14 Assume V is a subset of $\mathbb{R}^2$. Call $q \in \mathbb{R}^2$ is a limiting point of V, if every euclidean disk centered at q contains infinitely many points in V.

From the definition, a subset V has no limiting point means that every point in $q \in \mathbb{R}^2$, we can find a disk center at q, so that the disk contains only at most one point from V. As a result, a subset containing only finite points has no limiting point. However, be careful that a subset without limiting point is not necessary finite.

Example 2.15 (1) $\mathbb{Z} \times \mathbb{Z} \subset \mathbb{R}^2$ has no limiting point.

(2) Every point in $\mathbb{R}^2$ is a limiting point of $\mathbb{Q} \times \mathbb{Q} \subset \mathbb{R}^2$.

Definition 2.16 A subgroup Γ of Iso($\mathbb{R}^2, d$) acts on $\mathbb{R}^2$ discountinuously, if every orbit has no limiting point.

Example 2.17 (1) The subgroups that we use to construct cylinder, twisted cylinder, torus and the Klein bottle all act on $\mathbb{R}^2$ discountinuously.

(2) (This is Example 2.13 (1).) Consider the subgroup Γ generated by two translations as follows:

$$\Gamma =< t_{(0,1)}, t_{(0,\sqrt{2})} > .$$

Consider the orbit through $(0, 0) \in \mathbb{R}^2$, which is

$$[(0, 0)] = \{(0, m + \sqrt{2}n)|m, n \in \mathbb{Z}\}.$$

Notice that the subset $\{m + \sqrt{2}n|m, n \in \mathbb{Z}\}$ is dense in $\mathbb{R}$, hence Γ action on $\mathbb{R}^2$ is not discontinuous.

Next, we introduce another concept from group actions. Assume Γ is a subgroup of Iso($\mathbb{R}^2, d$). For each point $p \in \mathbb{R}^2$, the following subset of Γ

$$\Gamma_p := \{g \in \Gamma|g(p) = p\}$$

is a subgroup of Γ and is called the isotropy group of Γ at p.

Definition 2.18 We say a subgroup Γ of Iso($\mathbb{R}^2, d$) acts on $\mathbb{R}^2$ freely, if for every point $p \in \mathbb{R}^2$, the isotropy group Γ_p contains only the identity map.

Example 2.19 (1) The subgroups that we use to construct cylinder, twisted cylinder, torus and the Klein bottle all act on $\mathbb{R}^2$ discountinuously.

(2) Consider the subgroup generated by a reflection Rf_L (which contains only two elements). Then every point on L is fixed by the two elements and hence this group action is not free.

(3) (This is Example 2.13 (2).) Consider the subgroup generated by a rotation $r_{p,\theta}$ with $\theta \neq 0$. Then p is fixed by the whole subgroup hence this action is not free.

Proposition 2.20 *Assume Γ is a proper subgroup of $\mathrm{Iso}(\mathbb{R}^2, d)$. The quotient surface $(\mathbb{R}^2/\Gamma, d_\Gamma)$ is a locally euclidean metric space if and only if the subgroup Γ acts on R^2 discontinuously and freely.*

Proof We have seen from examples that if the group action of Γ is not free or not discontinuous, $\mathbb{R}^2/\Gamma$ is not locally euclidean. This shows that the locally euclidean property implies the discontinuous and free properities.

Conversely, assume Γ acts on $\mathbb{R}^2$ discontinuously and freely, then we can show that Γ must be generated by one translation or a glide reflection, or by two linear independent translations or glide reflections. For these four cases, we have shown that they are locally euclidean. (In the proof here, we used the three reflections theorem which avoids the using of assumption of properness for the group action.)

□

From the proof, in fact we have derived the following result.

Corollary 2.21 *Assume Γ is a subgroup of $\mathrm{Iso}(\mathbb{R}^2, d)$. The quotient surface $(\mathbb{R}^2/\Gamma, d_\Gamma)$ is a locally euclidean metric space if and only if it is one of the following four metric spaces:*

(1) *a cylinder. For this case, Γ is generated by a translation;*

(2) *a twisted cylinder. For this case, Γ is generated by a glide reflection;*

(3) *a torus. For this case, Γ is generated by two linearly independent translations;*

(4) *a Klein bottle. For this case, Γ is generated by two linearly independent glide reflections;*

In next section, we are going to consider more general locally euclidean surfaces and explain the corresponding Killing-Hopf theorem.

2.5 The Killing-Hopf theorem

To introduce the Killing-Hopf theorem, we need two mild conditions which we introduce now.

Definition 2.22 A metric space (X, d_X) is called path connected, if for any two points $x, y \in X$, there exists a continuous path in X joining x and y.

For example, the whole euclidean plane is path connected, but the euclidean plane excluding x-axis is not path connected. For a non path connected metric space, we can just look at its path connected components instead. For a metric space, if it is path connected then any both open and closed subset is either empty or the whole set. (The latter condition that any both open and closed subset is either empty or the whole set is called connectness. In general, a path connected space must be connected, but a connected space may not be path connected.)

Definition 2.23 A metric space (X, d_X) is called complete, if every Cauchy sequence converges to some point in X.

This definition is too analytic and not easy to use in geometry. Luckily, mathematicians prove that for a metric space which is locally euclidean (in fact, much more general than locally euclidean), it is complete if and only if every line segment (which is called geodesic for general cases) can be continued indefinitely. This is the property for a complete metric space that we are going to use. (Usually, this property is referred as geodesic completeness.)

Theorem 2.24 (Killing-Hopf theorem for euclidean case) *Each complete, connected locally euclidean surface is one of the following four metric spaces of the form* $\mathbb{R}^2/\Gamma$*:*

(1) *a cylinder;*

(2) *a torus;*

(3) *a twisted cylinder;*

(4) *a Klein bottle.*

Before we give a proof, let's see the importance of this theorem: The locally euclidean condition is a local condition, while the classification of types of four quotient surfaces is a global topology condition. A mathematical statement becomes extremely important once it indicates that certain local information could lead to a global result. For example, using the Killing-Hopf theorem, we can immediately conclude:

Example 2.25 (1) The topological 2-sphere can not be locally euclidean.

(2) The surfaces with more than one hole can not be locally euclidean.

We explain the key idea in the proof.

Step 1. We construct a covering $\pi : \mathbb{R}^2 \to S$, i.e., we look for a surjective map π so that it is also local isometry. First we take an arbitrary point $o_S \in S$ as the image of $(0, 0) \in \mathbb{R}^2$ by π. Then by the locally euclidean property, there exists a small disk $D(\epsilon)$ in $\mathbb{R}^2$ centered at $(0, 0)$ and a small disk $U_S(o_S; \epsilon)$ in S centered at o_S which are isometric to each other. We denote this isometry by $\pi : D(\epsilon) \to U_S(o_S; \epsilon)$.
Notice that the image of every ray in $D(\epsilon)$ by π is a line segment in S. Every point in $\mathbb{R}^2$ belongs to a unique ray. By extending the corresponding image segment, we define a map from $\mathbb{R}^2$ to S. The definition is based on the completeness assumption for S. Next we need to use the interplay of connected, completeness and locally euclidean property to show that π is a local isometry and surjective. Since the proof is a little subtle and beyond this class, we omit here. (Interested readers can refer the textbook.) This map is called the pencil map and was constructed first by Hopf [1925].

Step 2. First, we look at the set of euclidean isometries which are compatible with the covering map π:

$$\Gamma := \{f \in \mathrm{Iso}(\mathbb{R}^2, d) | \pi \circ f = \pi\}.$$

It is not hard to see Γ is a subgroup of $\mathrm{Iso}(\mathbb{R}^2, d)$. Next, we can show that $\mathbb{R}^2/\Gamma$ is isometric to S and this finishes the proof.

3 The sphere

3.1 The sphere $\mathbb{S}^2$ and its isometry group

In Section 1, we have studied the metric space $(\mathbb{R}^2, d)$ which is called the euclidean plane. In this section, we introduce an analogue, the sphere, and derive parallel results as we have obtained for the euclidean plane. To do this, the first thing we need to do is to define a distance function on the sphere.

A unit sphere is a surface in $\mathbb{R}^3$ defined as

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

and a distance function for $\mathbb{S}^2$ is defined as the shorter arc length between two points through the great circle. Equivalently, we can write the distance as

$$d_{\mathbb{S}^2}(p, q) = 2 \arcsin \frac{1}{2} d(p, q),$$

where d is the euclidean distance for $\mathbb{R}^3$. In this section, when we mention the sphere $\mathbb{S}$, we always mean the metric space $(\mathbb{S}^2, d_{\mathbb{S}^2})$. (If we just want to refer the topological sphere, we use the notation S^2 instead.)

Exercise 3.1 (1) Check $(\mathbb{S}^2, d_{\mathbb{S}^2})$ is a metric space.

(2) Show that this distance $d_{\mathbb{S}^2}$ is the same as the distance defined as the shortest path connecting the two points over the sphere. (It is easier to use spherical coordinates to calculate arc length.)

(3) Show that $(\mathbb{S}^2, d_{\mathbb{S}^2})$ is not locally euclidean. (Hint: Try to calculate the circumference of a circle with small radius and compare it with the one of its image in the euclidean plane.)

As for the euclidean plane, we now study the isometry group for $\text{Iso}(\mathbb{S}^2, d_{\mathbb{S}^2})$. We can also define rotation and reflection for the current situation, but since we have shown the three reflection theorem for euclidean plane, we might focus on the key concept first: Let's define reflections over $\mathbb{S}^2$. Recalling that to describe a euclidean reflection, we only need a line which is defined in Exercise 1.4 (4) using distance only. We define a line in $\mathbb{S}^2$ as the set

$$C_{p,q} := \{s \in \mathbb{S}^2 | d_{\mathbb{S}^2}(p, s) = d_{\mathbb{S}^2}(q, s)\},$$

for two distinct points $p, q \in \mathbb{S}^2$. This definition directly leads to the fact that an isometry of $\mathbb{S}^2$ maps a line to a line.

Exercise 3.2 Check that a line in $\mathbb{S}^2$ is a great circle in the sphere. As a result, an isometry of $\mathbb{S}^2$ maps a great circle to a great circle.

From this, we also see that any two distinct points on the sphere $\mathbb{S}^2$, there exists a unique line (i.e., a great circle) going through them. This is the same as the euclidean plane. However, now for $\mathbb{S}^2$, the parallel axiom no longer holds: Given a line and a point which is not on the line, there is no great circle going through the point and without intersecting the given line. This causes the difference of geometry between the euclidean plane and the sphere.

Now we define the reflection. Assume C is a line (i.e., a great circle) in $\mathbb{S}^2$. A reflection about C which we denote by Rf_C maps every point $p \in \mathbb{S}^2$ to the point $q \in \mathbb{S}^2$ so that the line determined by p, q as $C_{p,q}$ defined as above is the given line C.

Exercise 3.3 Prove that map Rf_C is well-defined. For this, you need to prove that such q exists and is unique. Moreover, check that a reflection is an isometry over $\mathbb{S}^2$ and $Rf_C \circ Rf_C = \mathrm{id}_{\mathbb{S}^2}$.

Assume ℓ is an axis through the origin with a positive direction being chosen. We can define the rotation around ℓ for points on $\mathbb{S}^2$ as rotations on the plane which going through the point the perpendicular to ℓ. We use $r_{\ell,\theta}$ to denote a rotation with θ-angle.

For isometries of the euclidean plane, we also consider translations and rotations and proved that a translation or a rotation can be written as compositions of two reflections. Whether the composition of two refections gives a translation or a rotation is determined by whether the two lines have intersection or not. For the sphere $\mathbb{S}^2$, every two distinct lines (great circles) must intersect at two points which are antipodal to each other, i.e., the axis connecting these two points must go through the origin. Their composition is a rotation around this axis with the double angle between these two great circles. Conversely, any rotation around an axis through the origin can be decomposed into two reflections. In another word, for the sphere $\mathbb{S}^2$, there are no translations but only rotations. This will bring a different result for the corresponding spherical Killing-Hopf theorem.

We now state the three reflections theorem for the sphere whose proof is a strict analogue of the corresponding result for the euclidean plane.

Theorem 3.4 (The three reflections theorem for the sphere) *Any isometry for $\mathbb{S}^2$ can be written as compositions of one or two or three reflections.*

Exercise 3.5 Prove the three reflections theorem for the sphere case.

Among these isometries, the orientation-preserving ones and orientation-reversing ones are defined in the same way as for the case of $\mathrm{Iso}(\mathbb{R}^2, d)$. Similarly, we use the notation $\mathrm{Iso}^+(\mathbb{S}^2)$ to denote the set of orientation-preserving isometries, but now it contains only rotations. The same as the euclidean case, $\mathrm{Iso}^+(\mathbb{S}^2)$ is a normal subgroup of $\mathrm{Iso}(\mathbb{S}^2)$ with index 2. People also use $SO(3)$ to denote $\mathrm{Iso}^+(\mathbb{S}^2)$. We are going to discuss more on it in Section **??**.

For the orientation-reversing isometries, we can similarly prove that any composition of three reflections is a glide reflection, but now, a glide reflection means a rotation composited with a reflection along the great circle perpendicular to the rotation axis.

3.2 Locally spherical surfaces

Theorem 3.6 (The Killing-Hopf theorem for spherical surfaces) *A connected complete locally spherical surface is either a sphere or a projective plane ($\mathbb{R}P^2$).*

We explain the key difference between the current Killing-Hopf theorem and the one for the euclidean plane. The point is, a nontrivial subgroup Γ of $Iso(\mathbb{S}^2)$ which discountinuously and freely acts on $\mathbb{S}^2$ can only be the one generated by the glide reflection f_π whose rotation angle is π.

Exercise 3.7 Explain the above point in details. In particular, explain why the glide reflection with other angles can not generate a discountinuous and free group action.

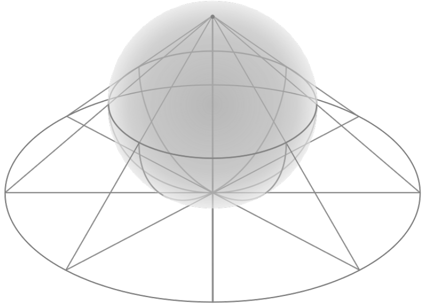
The quotient surface $\mathbb{S}^2/ < f_\pi >$, which is called the projective plane (because it can be understood as the space of lines in $\mathbb{R}^3$), can be understood as gluing the antipodal points on a sphere. As the Klein bottle, it is compact, without boundary and not orientable (one-sided surface). It is not embedded in $\mathbb{R}^3$ either.

Exercise 3.8 Find a fundamental region for the projective plane and try to find a Möbius band in the fundamental region.

3.3 Relation to the complex plane

For the euclidean plane, we know as a set, it is bijective to the complex plane $\mathbb{C}$ and we can represent basic isometries by maps over it. For the sphere, we can also do it (as a set only) using the so-called stereographic projection. Of course, the sphere and the complex plane are not isometric but they are conformal in the sense that the angles are preserved.

We explain the construction of stereographic projection first. Let's denote the north pole $(0,0,1)$ by N. For any point $p \in \mathbb{S}^2$ which is not N, we can find a unique line in $\mathbb{R}^3$ going through N and p. This line intersection with the $\{z=-1\}$-plane at a unique point, which we define as the image of p under the stereographic projection. By this way, every point on $\mathbb{S}^2 \setminus \{N\}$ one-to-one corresponds to point on the $\{z=-1\}$-plane which can be2 identified with the complex plane $\mathbb{C}$. We can formally assign the infinity as the image of the north pole. Denote by $\overline{\mathbb{C}}$ the complex plane adding the infinity, then the stereographic projection is a bijection from $\mathbb{S}^2$ to $\overline{\mathbb{C}}$.



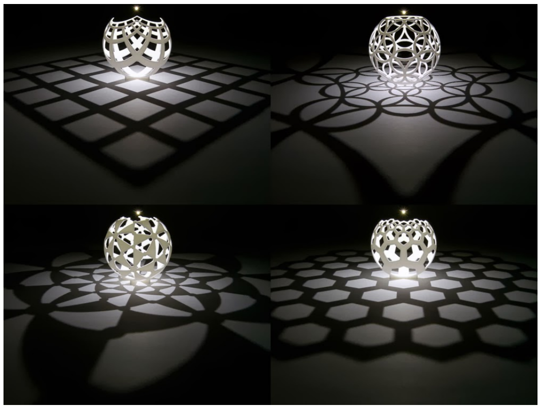
Explicitly, the stereographic projection pr has the expression

$$pr(x,y,z) = (\frac{2x}{1-z}, \frac{2y}{1-z}) \in \overline{\mathbb{C}}.$$

Its inverse map

$$pr^{-1}(u,v) = (\frac{4u}{u^2+v^2+4}, \frac{4v}{u^2+v^2+4}, \frac{u^2+v^2-4}{u^2+v^2+4}) \in \mathbb{S}^2.$$

The stereographic projection maps the great circle $\{(x,y,0) \in \mathbb{S}^2 | x^2+y^2=1\}$ to the circle $\{(u,v,-1)|u^2+v^2=4\}$. From this, clearly it is not an isometry (Why?). Instead, it assigns $\overline{\mathbb{C}}$ a different metric space structure from the euclidean plane. One important property for stereographic projection is that it is conformal, i.e., it preserves angles. Moreover it maps a circle on $\mathbb{S}^2$ to a circle in the $\{z=-1\}$-plane.



Exercise 3.9 Assume C is the great circle in $\mathbb{S}^2$ determined by the plane $ax+by+cz=0$ with $a^2+b^2+c^2$. Then its stereographic projection is the circle in the plane $\{(x,y,z)\in\mathbb{R}^3|z=-1\}$ with center $(-\frac{2a}{c},-\frac{2b}{c},-1)$ and radius $\frac{2}{|c|}$.

3.4 The inversions

Consider a circle in the complex plane centered at $z_0\in\mathbb{C}$ with radius R as $C_{z_0,R}:=\{w\in\mathbb{C}||w-z_0|=R\}$. A inversion about $C_{z_0,R}$ is defined as a map

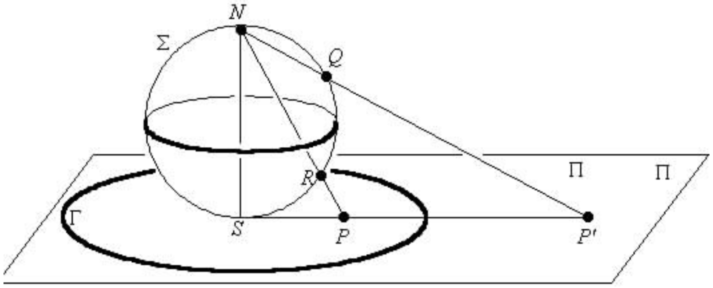
$$\mathrm{Inv}_{C_{z_0,R}}:\mathbb{C}\setminus\{z_0\}\to\mathbb{C}\setminus\{z_0\}$$

as

$$\mathrm{Inv}_{C_{z_0,R}}(w)=\frac{z_0\bar{w}+(R^2-|z_0|^2)}{\bar{w}-\bar{z_0}}.$$

(We can also include z_0 by extending $\mathbb{C}$ to $\overline{\mathbb{C}}=\mathbb{C}\cup\{\infty\}$.) The geometric meaning of inversion is as shown in the following picture that the inversion of the point P is P' with the relation that

$$|SP|\cdot|SP'|=R^2.$$



In fact, the same picture also shows the relation between the inversion and the reflection about a great circle in the sphere $\mathbb{S}^2$.

Proposition 3.10 *Assume C is a great circle in $\mathbb{S}^2$ whose image under the stereographic projection is denoted as C'. We know from last section that C' is a circle in the complex plane $\mathbb{C}\cong\{(x,y,z)\in\mathbb{R}^3|z=-1\}$. There is the relation between the reflection about a great circle C and the inversion about the circle C' as*

$$\mathrm{pr}\circ Rf_C=\mathrm{Inv}_{C'}\circ\mathrm{pr}.$$

Here $\mathrm{pr}:\mathbb{S}^2\to\mathbb{C}$ denotes the stereographic projection.

Further combine with Exercise 3.9, we get the following result.

Proposition 3.11 *Assume C is the great circle in $\mathbb{S}^2$ determined by the plane $ax+by+cz=0$ with $a^2+b^2+c^2=1$. Then the inversion about the stereographic projection of C is*

$$\mathrm{Inv}_{C'}(w)=\frac{2(a+bi)\bar{w}-4c}{-c\bar{w}-2(a-bi)}.$$

Recall that a rotation is a composition of two reflections, from Proposition 3.11 we obtain the following expression for a rotation under the stereographic projection.

Proposition 3.12 *Any rotation r over the sphere $\mathbb{S}^2$ has the following expression*

$$\mathrm{pr}\circ r\circ \mathrm{pr}^{-1}(w)=\frac{2uw+4v}{-\bar{v}w+2\bar{u}},$$

with some $u,v\in\mathbb{C}$ with $|u|^2+|v|^2=1$.

Proof We sketch the proof and leave the details to be checked by you. Assume r is decomposed into two reflections as $Rf_{C_2}\circ Rf_{C_1}$ whose inversions under stereographic projections are

$$\mathrm{pr}\circ Rf_{C_j}\circ \mathrm{pr}^{-1}=\frac{2\lambda_j\bar{w}-4\nu_j}{-\nu_j\bar{w}-2\lambda_j},\quad j=1,2.$$

Here $\lambda_j\in\mathbb{C}$ and $\nu_j\in\mathbb{R}$ with $|\lambda_j|^2+\nu_j^2=1$.

Then we calculate their composition and get

$$\mathrm{pr}\circ r\circ \mathrm{pr}^{-1}(w)=(\mathrm{pr}\circ Rf_{C_2}\circ \mathrm{pr}^{-1})\circ(\mathrm{pr}\circ Rf_{C_1}\circ \mathrm{pr}^{-1})(w)=\frac{2uw+4v}{-\bar{v}w+2\bar{u}}$$

where

$$u=\bar{\lambda_1}\lambda_2+\nu_1\nu_2,\quad v=\lambda_1\nu_2-\lambda_2\nu_1.$$

Direct checking shows that $|u|^2+|v|^2=1$.

□

We remark that the coefficients 2 and 4 will disappear if we do the stereographic projection to the xy-plane instead.

This expression is interesting in the following sense: Consider the map which maps the map $\frac{2uw+4v}{-\bar{v}w+2\bar{u}}$ to the matrix $\begin{bmatrix}u & v\\ -\bar{v} & \bar{u}\end{bmatrix}$. Then we find that the composition of maps commutes with the matrix multiplications.

The image set of 2×2 complex matrices forms a subgroup which is called the group of special unitary matrices

$$SU(2):=\{\begin{bmatrix}u & v\\ -\bar{v} & \bar{u}\end{bmatrix}|u,v\in\mathbb{C},|u|^2+|v|^2=1\}.$$

On the other hand, a rotation over $\mathbb{S}^2$ corresponds to an orientation-preserving linear map which preserves norms for the linear space $\mathbb{R}^3$. From linear algebra class, we know such matrices form a subgroup of 3×3 real matrices which is called the group of special orthogonal matrices

$$SO(3):=\{A\in GL_3(\mathbb{R})|AA^T=I,\det(A)=1\}.$$

By this way, the above construction in fact gives a map

$$\Phi: SU(2)\to SO(3),\quad \begin{bmatrix}u & v\\ -\bar{v} & \bar{u}\end{bmatrix}\to r(\cdot)=\frac{2u\cdot(\cdot)+4v}{-\bar{v}\cdot(\cdot)+2\bar{u}}.$$

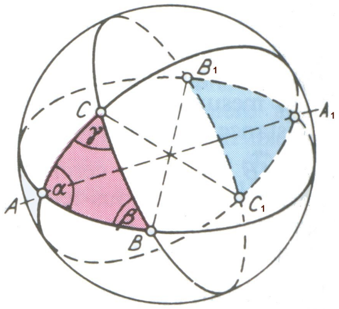
The above construction has shown that the map Φ is surjective. However, it is not injective but a 2-to-1 map.

Remark 3.13 From our interpretation for the group $SO(3)$, it is clear to see that it has a smooth manifold structure which is diffeomorphic to $\mathbb{R}P^3$ since it corresponds to equivalence classes by antipodal points in $\mathbb{C}^2\cong\mathbb{R}^4$. It is connected, but not simply-connected.

The non-contractible loop corresponds to the Balinese plate trick https://en.wikipedia.org/wiki/Plate_trick.

3.5 The Gauss-Bonnet formula

For three points A, B, C which are not in a line on the euclidean plane, it is well-known that the sum of interior angles of the triangle ΔABC is π. Now we ask the same question for three points A, B, C which are not in a great circle on the unit sphere $\mathbb{S}^2$. A spherical triangle is a region in $\mathbb{S}^2$ formed by three great circular arcs intersecting pairwise in three vertices. For example, in the following picture, the ΔABC is a spherical triangle.



An interior angle of a spherical triangle is defined as the angle between the two planes where the two adjacent great circular arcs live in. Then we can ask, what is the sum of the three interior angles in a spherical triangle? Some special examples show us that this sum must be great than π but is no longer a constant. Is this possible for us to obtain an equality for it then? We now try to derive a formula for the general case.

Consider the spherical triangle ΔABC as in the above picture and notice that

$$\text{Area}(ACA_1B) = \text{Area}(\Delta ABC) + \text{Area}(\Delta BCA_1).$$

Here A_1 is the antipodal point of A and thus $\text{Area}(ACA_1B) = \frac{\alpha}{2\pi} \cdot 4\pi R^2 = 2\alpha R^2$, where R is the radius of the sphere. Similarly, we also have

$$\begin{aligned} \text{Area}(\Delta ABC) + \text{Area}(\Delta ACB_1) &= 2\beta R^2 \\ \text{Area}(\Delta ABC) + \text{Area}(\Delta ABC_1) &= 2\gamma R^2. \end{aligned}$$

Sum them, we get

$$2(\alpha + \beta + \gamma)R^2 = 3\text{Area}(\Delta ABC) + (\text{Area}(\Delta BCA_1) + \text{Area}(\Delta ACB_1) + \text{Area}(\Delta ABC_1)).$$

Further notice that

$$\text{Area}(\Delta BCA_1) + \text{Area}(\Delta ACB_1) + \text{Area}(\Delta ABC_1) = 2\pi R^2 - \text{Area}(\Delta ABC).$$

Put these together, we obtain the following formula

$$\alpha + \beta + \gamma = \pi + \frac{1}{R^2}\text{Area}(\Delta ABC).$$

This formula not only shows that the sum of the three interior angles in a triangle is greater than π, but also explicitly relate the difference with the area of the triangle.

In fact, the constant $\frac{1}{R^2}$ also has a geometric meaning, which is nothing but the Gaussian curvature of the sphere with radius R. In general, the Gaussian curvature of a surface at a point is defined as the multiplication of the maximum and minimum values of the curvatures of all curves in the surface through the point. (Euler proved in 1760 that the curves in the surface which take the maximum and minimum values are perpendicular to each other at the point.)

A more general formula for triangles in a constant curvature surface is then given as

$$\alpha+\beta+\gamma=\pi+\kappa\cdot\text{Area}(\Delta ABC) \tag{2}$$

where κ is the Gaussian curvature. The euclidean plane has zero Gaussian curvature and then this formula is just the one that states the sum of interior triangles in a euclidean triangle is π. In later lectures, we are going to study the hyperbolic plane whose Gaussian curvature is constant -1 and then the sum of angles of a hyperbolic triangle is π minus the hyperbolic area of the triangle.

The formula (2) is referred as (a simple version of) the famous Gauss-Bonnet formula.

Exercise 3.14 (1) Prove the following formula for any spherical polygon

$$\text{Area(polygon)}+\text{sum of exterior angles}=2\pi.$$

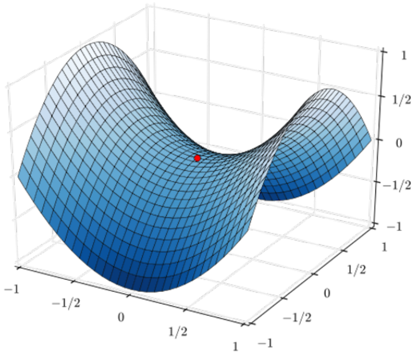
(Notice that this formula is independent of the number of vertices of the polygon.)

(2) Try to prove that two spherical triangles with the same angles must be isometric. (This result shows another difference between the spherical geometry and the euclidean geometry.)

4 The hyperbolic plane

4.1 The pseudosphere

As we have seen that the sphere is a surface of constant positive curvature. The nonzero curvature of the sphere make it has different geometry from the euclidean plane. Next, it is natural to look for some surface with negative constant curvature and then the geometry on this surface is expected to be opposite to the sphere. Locally, we know that a point on a surface with a negative curvature is a saddle point.

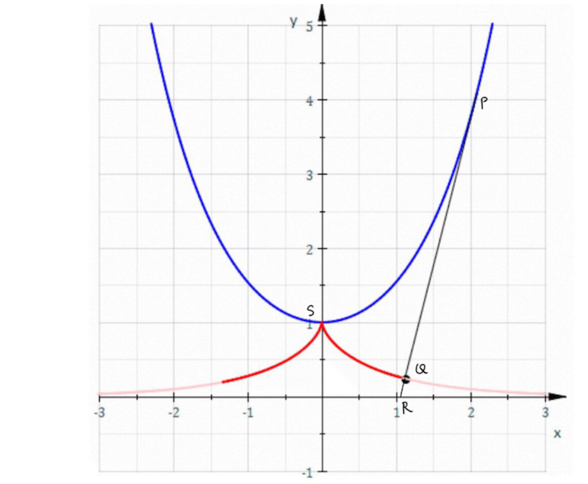


Let's now introduce example which has saddle points everywhere.

Consider the curve $v=\cosh u=\frac{e^u+e^{-u}}{2}$ in the uv-plane. Then we draw its involute curve, which is called tractrix, as the red curve shown in the following picture. The involute curve is determined by the condition that

$$|PQ|=\text{arc length } PS$$

and that PQ is tangent to the hyperbolic consine curve at P.



Next, we obtain the revolution of the tractrix about the u-axis. The surface we obtain in this way is called a pseudosphere. The Gaussian curvature of the pseudosphere at Q can be calculated as

$$\kappa = -\frac{1}{|PQ|} \cdot \frac{1}{|QR|}.$$

We assume P has coordinate $(u, \cosh u)$ in the uv-plane. Then

$$|PQ| = \text{arc length } PS = \int_0^u \sqrt{1+(\frac{dv}{du})^2 du} = \sinh u = \frac{e^u - e^{-u}}{2}.$$

On the other hand, use the slope of the tangent line PR at P, we can calculate out that R has coordinate $(u - \coth u, 0)$, where

$$\coth u := \frac{\sinh u}{\cosh u} = \frac{e^u - e^{-u}}{e^u + e^{-u}}.$$

It follows $|QR| = |PR| - |PQ| = \frac{1}{\sinh u}$. As a result, we have finished showing that the Gaussian curvature is constant -1 over the pseudosphere.

We end this section by discussing the defect of the pseudosphere.

(1) This surface is not smooth at $u = 0$ as we can see that the point $(0, 1)$ in the uv-plane is sharp which makes the whole circle from rotation become singularities. If we remove this circle from the surface, the resulting surface is not complete.

(2) It is not easy to describe 'lines' for the pseudosphere.

In fact, Hilbert in 1901 proved that any surface of constant negative curvature smoothly embedded in $\mathbb{R}^3$ can not be complete. This in fact break our dreams to look for a model as $\mathbb{S}^2$ for the negative curvature case. For this reason, we have to get rid of the requirement that the metric for the surface is induced from the euclidean metric $\mathbb{R}^3$.

4.2 The hyperbolic plane

The hyperbolic plane is a metric space whose set of points can be taken as the upper half plane

$$\mathbb{H}^2 := \{(x, y) \in \mathbb{R}^2 | y > 0\}.$$

We use the metric on the pseudosphere induces a metric on $\mathbb{H}^2$ as follows: define x as the rotation angle around u-axis and τ the arc length from S to Q. We can use (x,τ) to parametrize the pseudosphere. Then we calculate the infinitesimal distance between points (x,τ) and $(x+dx,\tau+d\tau)$ over the pseudosphere as

$$\begin{aligned} ds^2 &= (vdx)^2+(d\tau)^2 \\ &= v^2(dx)^2+(d\tau)^2. \end{aligned}$$

Here (u,v) is the uv-coordinate for the point (x,τ). A little calculation shows that $v=e^{-\tau}$. Hence the infinitesimal distance is

$$ds=\sqrt{e^{-2\tau}(dx)^2+(d\tau)^2}.$$

We change parameter again by defining $y=e^{\tau}$, then $dy=e^{\tau}d\tau$ and then this infinitesimal distance can be written as

$$ds=\frac{\sqrt{dx^2+dy^2}}{y}.$$

Notice that, though by change of variable, our parameters (x,y) has range $[0,2\pi)\times(1,\infty)$, we can safely extend the domain of (x,y) to the whole upper half plane $\mathbb{H}^2$ and regard it as an infinitesimal distance on $\mathbb{H}^2$. From now on, we use the notation $ds_{\mathbb{H}^2}$ to denote it.

Using $ds_{\mathbb{H}^2}$, we define a metric over the upper half plane $\mathbb{H}^2$ as the following way: For any two points $p,q\in\mathbb{H}^2$, the hyperbolic distance is defined as

$$\begin{aligned} d_{\mathbb{H}^2}(p,q) &= \inf_{\gamma:[0,1]\to\mathbb{H}^2,\gamma(0)=p,\gamma(1)=q}\int_\gamma ds_{\mathbb{H}^2} \\ &= \inf_{\gamma:[0,1]\to\mathbb{H}^2,\gamma(0)=p,\gamma(1)=q}\int_0^1\frac{\sqrt{x'(t)^2+y'(t)^2}}{y(t)}dt \end{aligned}$$

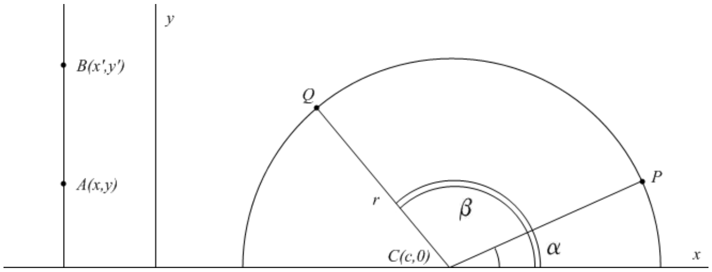
By the construction, we have the following conclusion.

Lemma 4.1 $(\mathbb{H}^2,d_{\mathbb{H}^2})$ *is a metric space. It is called the hyperbolic plane.*

In fact, we are going to show later that the path which realizes the minimal distance is either lines parallel to y-axis or half circles with centers at x-axis.

To give an intuition for the difference between the hyperbolic distance and the euclidean distance, we see the following example.

Example 4.2 Consider the half circle with center $C=(c,0)$ and radius r on the upper half plane. Let's calculate the arc length from P to Q along this half circle as shown in the following picture.



We can parametrize the curve as

$$x = c + r\cos\theta, \quad y = r\sin\theta, \quad \theta \in [\alpha, \beta].$$

We calculate the hyperbolic arc length as

$$\begin{aligned} & \int_\alpha^\beta \frac{\sqrt{x'(\theta)^2 + y'(\theta)^2}}{y(\theta)} d\theta \\ = & \int_\alpha^\beta \frac{1}{\sin\theta} d\theta \\ = & \frac{1}{2}\log|\frac{(\cos\beta - 1)(\cos\alpha + 1)}{(\cos\beta + 1)(\cos\alpha - 1)}|. \end{aligned}$$

We notice that, this arc length is independent of the radius r but only depends on angles.

For the points $A = (x, y)$ and $B = (x', y') = (x, y')$, we can calculate the hyperbolic distance as

$$\int_y^{y'} \frac{\sqrt{0^2 + 1^2}}{y} dy = \log y' - \log y.$$

4.3 The Poincaré disk

Now we introduce another model for the hyperbolic plane. Consider the map from $\overline{\mathbb{C}} \to \overline{\mathbb{C}}$ defined as

$$\phi(z) = \frac{iz + 1}{z + i}$$

The map ϕ is a bijective, orientation-preserving map. It maps 1 to 0, maps 0 to $-i$, maps ∞ to i. Moreover, it maps the upper half plane to the open unit disk

$$\mathbb{D}^2 := \{z \in \mathbb{C} | |z| < 1\},$$

with the real axis to the unit circle. We can calculate out the inverse map which is

$$\psi(w) = \frac{w + i}{iw + 1}.$$

Using the complex expression of the infinitesimal distance

$$ds_{\mathbb{H}^2} = \frac{\sqrt{dx^2 + dy^2}}{y} = \frac{|dz|}{\mathrm{Im} z},$$

we can formally calculate the corresponding infinitesimal distance in the unit disk $\mathbb{D}^2$

$$ds_{\mathbb{D}^2} = \frac{|d\frac{w+i}{iw+1}|}{\mathrm{Im}(\frac{w+i}{iw+1})} = \frac{2|dw|}{1 - |w|^2}.$$

This calculation in fact shows that the metric space $(\mathbb{D}^2, d_{\mathbb{D}^2})$ with the distance function induced from $ds_{\mathbb{D}^2}$ is isometric to $(\mathbb{H}^2, d_{\mathbb{H}^2})$. In anther word, these two models are different coordinate system for the same geometric space. We are going to interchange the use of these two models freely according to our different purposes.

There are two important properties for the $\mathbb{D}^2$-model that we would like to point out:

(1) It maps the euclidean lines parallel to y-axis and half circles centered on x-axis in $\mathbb{H}^2$ to arcs which are perpendicular to the boundary circle of the disk $\mathbb{D}^2$ (including diameters). These are in fact $\mathbb{D}^2$-lines. In particular, it maps the y-axis in $\mathbb{H}^2$ to the y-axis in $\mathbb{D}^2$ and maps the unit circle center at origin in $\mathbb{H}^2$ to the x-axis in $\mathbb{D}^2$.

(2) The map $\phi : \mathbb{C} \to \mathbb{C}$ in fact is the reflection about y-axis composed with the inversion about the circle centered at $-i$ with radius $\sqrt{2}$. From this, we can at least see that the map ϕ preserves angles. (In fact, all such kind of maps which are only in z (i.e., with no $\bar{z}$) preserve angles.)

4.4 Hyperbolic isometries

In this section, we first see five examples of hyperbolic isometries. Some of them have simple expressions in the $\mathbb{H}^2$-model and some have simple expressions in the $\mathbb{D}^2$-model. Then we see that these are isometries are plenty enough to move every hyperbolic triangle into some favored positions. Using these, we are able to obtain the three reflections theorem just as for the euclidean case.

Before we start, we give the following lemma whose proof we skip.

Lemma 4.3 *A smooth map $f = (f_1, f_2) : \mathbb{H}^2 \to \mathbb{H}^2$ is a hyperbolic isometry if and only if*

$$\frac{\sqrt{(df_1)^2 + (df_2)^2}}{f_2} = \frac{|dz|}{\mathrm{Im} z}.$$

Now let's see the five examples and the check of they are hyperbolic isometries using Lemma 4.3 is left to you.

The first three examples have simple expression in the upper half plane model $\mathbb{H}^2$.

(1) For any $\alpha \in \mathbb{R}$, define $t_\alpha : \mathbb{H}^2 \to \mathbb{H}^2$ as

$$t_\alpha(z) = z + \alpha.$$

This map is called a limit rotation. The reason to regard it as a rotation is it is a composition two reflections with the two reflection axises intersecting at infinity.

(2) For any $\rho \in \mathbb{R}^+$, define the dilatation

$$d_\rho(z) = \rho z.$$

This is map is in fact a translation, because it is a composition of two reflections about half circles centered at the origin.

(3) We can express the reflection about the y-axis by the following map

$$Rf_y(z) = -\bar{z}.$$

Exercise 4.4 Try to express the above three isometries in the Poincaré disk model $\mathbb{D}^2$.

The next two examples of isometries have simple expression in the disk model.

(1) For any $\theta \in \mathbb{R}$, define the rotation $r_\theta : \mathbb{D}^2 \to \mathbb{D}^2$ as

$$r_\theta(z) = e^{i\theta} z.$$

It can be decomposed into two reflections about diameters.

(2) We can express the reflection about the x-axis by the following map

$$Rf_x(z) = \bar{z}.$$

Exercise 4.5 Try to express the above two isometries in the upper half plane model $\mathbb{H}^2$. In particular, the reflection about the x-axis in $\mathbb{D}^2$-model is in fact the inversion about the unit disk in $\mathbb{H}^2$ model, i.e., the map $z \mapsto \frac{1}{\bar{z}}$.

Lemma 4.6 *For any two points A, B in $\mathbb{H}^2$, we can find construct a hyperbolic isometry using the above five basic isometries to map both A, B to the y-axis.*

Proof Assume $A = (\alpha_A, \beta_A)$ and $B = (\alpha_B, \beta_B)$. We first use the translation $t_{-\alpha_A}$ to map A to $A' = (0, \beta_A)$ and at the same time, B is mapped to the point $B' = (\alpha_B - \alpha_A, \beta_B)$. Then we can dilate and make the A' to i by multiplying $\frac{1}{\beta_A}$. The B' is mapped to some B''.

Then we move to the $\mathbb{D}^2$-model, in which the i in $\mathbb{H}^2$-model becomes the origin. So, we can rotate the corresponding image of B'' to some point on the y-axis. Notice that the preimage of the y-axis in the $\mathbb{H}^2$-model is also the y-axis, we are done with the construction. □

From now on, when we try to prove any property which is supposed to be invariant under isometries, we can apply this lemma to simplify the case we are dealing with.

As an application, let's show now the result we mentioned before.

Proposition 4.7 *For any two points A, B in $\mathbb{H}^2$, the shortest hyperbolic distance is realized by either of the following two cases:*

(1) *the euclidean line connected AB if AB is parallel to the y-axis;*

(2) *the arc of a circle centered on x-axis which joins A, B if AB is not parallel to the y-axis.*

Proof We start with the simplest case.

(1) If A, B are both on the y-axis. WLOG, we can assume any path connecting AB is parametrized as $(x(y), y)$ with $y \in [a, b]$, where a, b are the y-coordinates of A' and B'. Then we calculate

$$\int_a^b \frac{\sqrt{x'(y)^2 + 1}}{y} dy \geq \int_a^b \frac{1}{y} dy,$$

and the equality can be taken if and only if the latter is just the hyperbolic arc length of the line segment AB. We are done for this case.

(2) If only A is on the y-axis.

(a) If $A = i$. Then we can take the rotation on the $\mathbb{D}^2$-model around the center 0 to map B to the y-axis. Then apply (1), we know the shortest distance is realized by lines over y-axis. Then we only need to understand the image of y under rotation. Notice that the rotation of θ-angle on the $\mathbb{H}^2$-model can be expressed as

$$z \mapsto e^{i\theta} \frac{iz + 1}{z + i} \mapsto \frac{e^{i\theta} \frac{iz+1}{z+i} + i}{ie^{i\theta} \frac{iz+1}{z+i} + 1}.$$

Hence when z in on the y-axis, i.e., $z = yi$, we can do some calculation and show that every point on the image has equal euclidean distance to the point $\cot\theta$. Then this shows that it maps the y-axis to the half circle centered at $\cot\theta$ with radius $\csc\theta$. We are done for this case too.

(b) If A is not i. We can do dilatation to map A to i. Then we only need to show that half circle centered at x-axis is still some half circle centered at x-axis after dilatation, which is obvious since the image of the circle at $x_0 \in \mathbb{R}$ is $x_0 + re^{i\theta}$ and its image after dilatation ρ is $\rho x_0 + \rho re^{i\theta}$, which is still a half circle centered at x-axis. We are done with this case too.

(3) If neither A nor B on the y-axis. We can use limit rotation, i.e., real translation to move A to y-axis. Then this reduces to one of the cases are done. We only need to argue that a half circle centered at x-axis is still some half circle centered at x-axis after real translation, which is definitely true. Hence all cases are done.

□

In general, such lines which realize the shortest distance are called geodesic lines for the corresponding metric. For example, the geodesic lines for the euclidean metric are just the lines is usual sense. The geodesic lines for the spherical metric over $\mathbb{S}^2$ are great circles.

All these three types of geometry share the following important property that a geodesic line can be characterized using the equidistance property for two points.

Proposition 4.8 *A hyperbolic geodesic line can be expressed as the set of hyperbolic equidistance to two points in $\mathbb{H}^2$.*

Proof Similar as previous treatment, we can deal with the simplest case with A, B share equal euclidean distance to the y-axis and AB parallel to the x-axis in the euclidean sense. For this case, it is not hard to see that the set of hyperbolic equidistance to A, B is just the y-axis.

For general case, we can use the basic hyperbolic isometries to transfer any A, B to the previous case. The details are left to you. □

As a consequence, we can define the reflection about a hyperbolic geodesic line just as what we did for the spherical case. (See the definition and discussion around Exercise 3.3.) Further using the basic isometries, we mimic the proof for the euclidean case and are able to obtain the three reflections theorem again.

Theorem 4.9 *Any hyperbolic isometry can be written as compositions of one or two or three reflections.*

As a corollary, every hyperbolic isometry is surjective and we obtain

Corollary 4.10 $\mathrm{Iso}(\mathbb{H}^2)$ *forms a group.*

4.5 Hyperbolic isometries under complex coordinates

From the three reflections theorem for the hyperbolic case, we know that to understand hyperbolic isometries under complex coordinates, the first step is to express reflections under complex coordinates. For this, we have the following lemma.

Lemma 4.11 *In the $\mathbb{H}^2$-model, the reflection about the hyperbolic line C, i.e, a semicircle centered on x-axis, is the inversion map. Here, we also include the case that C is a half line parallel to the y-axis and for this case, the reflection is the same as the euclidean reflection.*

Proof For the special case that C is a half line parallel to the y-axis, it is clear from euclidean geometry to see that a hyperbolic reflection is the same as the euclidean reflection.

We now look at general cases.

(1) If C is the half unit circle centered at the origin. Notice that for this case, the image of C in the $\mathbb{D}^2$-model is the x-axis, and the reflection in the $\mathbb{D}^2$-model is conjugate map. Hence, the reflection about C in the $\mathbb{H}^2$-model is the composition of the following maps

$$z \mapsto \phi(z) \mapsto \overline{\phi(z)} \mapsto \psi(\overline{\phi(z)}),$$

where $\phi : \mathbb{H}^2 \to \mathbb{D}^2$ is $\phi(z) = \frac{iz+1}{z+i}$ and $\psi(w) = \frac{w+i}{iw+1}$ is its inverse map. Direct calculation shows that it maps z to $\frac{1}{\bar{z}}$, i.e., the inversion about the unit circle C.

(2) If C is a radius r half circle centered at the origin. We can use the dilatation by $\frac{1}{r}$ to map C to the unit circle C_1. The reflection then is the composition of the following maps

$$z \mapsto \frac{z}{r} \mapsto Rf_{C_1}(\frac{z}{r}) \mapsto r \cdot Rf_{C_1}(\frac{z}{r}) = \frac{r^2}{\bar{z}}.$$

This is the inversion about the circle C.

(3) For any radius r half circle C centered at some $x_0 \in \mathbb{R}$. We can use limit rotation t_{-x_0} to map it to a circle center at the origin. Then this is the case (2). It follows the reflection is the composition of the following maps

$$z \mapsto z - x_0 \mapsto \frac{r^2}{\overline{z - x_0}} \mapsto \frac{r^2}{\overline{z - x_0}} + x_0 = \frac{x_0\bar{z} + (r^2 - x_0^2)}{\bar{z} - x_0}.$$

Clearly, this is an inversion about the circle C.

□

Then by direct calculation, we will further obtain the following statement which was first discovered by Poincaré.

Theorem 4.12 (Poincaré) *The $\mathbb{H}^2$-isometries are either of the form*

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

where $a, b, c, d \in \mathbb{R}$ and $ad - bc = 1$ or of the form

$$\bar{f}(z) = \frac{-a\bar{z}+b}{-c\bar{z}+d},$$

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

The former are orientation-preserving isometries and the latter are orientation reversing isometries.

In the $\mathbb{D}^2$-model, the above theorem has the following interpretation.

Theorem 4.13 *The $\mathbb{D}^2$-isometries are either of the form*

$$f(z) = \frac{az+b}{\bar{b}z+\bar{a}},$$

where $a, b \in \mathbb{C}$ and $|a|^2 - |b|^2 = 1$ or of the form

$$\bar{f}(z) = \frac{az+b}{\bar{b}z+\bar{a}},$$

where $a, b \in \mathbb{C}$ and $|a|^2 - |b|^2 = 1$.

The former are orientation-preserving isometries and the latter are orientation reversing isometries.

We end this section by the following remark.

Remark 4.14 We can construct a map from the orientation preserving hyperbolic isometries $\mathrm{Iso}^+(\mathbb{H}^2)$ to the group of projective special linear group $PSL_2(\mathbb{R})$ as

$$\mathrm{Iso}^+(\mathbb{H}^2) \to PSL_2(\mathbb{R}): \quad f(z) = \frac{az+b}{cz+d} \mapsto \begin{bmatrix} a & b \\ c & d \end{bmatrix}.$$

This is a group isomorphism.

On the other hand, the group $PSL_2(\mathbb{R})$ is also the group $\mathrm{Aut}(\mathbb{H}^2)$ of biholomorphic maps from $\mathbb{H}^2$ to itself (This can be proved by the Schwarz lemma). This is an important fact which indicates that the hyperbolic geometry is closely related to complex analysis. (The euclidean isometry group doesn't have such connection to $\mathrm{Aut}(\mathbb{C})$.)

4.6 Geometric properties of hyperbolic isometries

By the three reflections theorem, all hyperbolic isometries are among the three classes: reflection, two reflections and three reflections. Moreover, compositions of two reflections are orientation preserving hyperbolic isometries. Let's first understand the geometry of such isometries.

When we consider compositions of two reflections in euclidean geometry, whether it is a rotation or a translation completely determined by the relation of two reflection lines. Here we have similar results, but the relation of two geodesic lines in hyperbolic geometry is more complicated.

Definition 4.15 Assume C_1 and C_2 are two hyperbolic geodesic lines. We say they

(1) intersect, if $C_1 \cap C_2 \neq \emptyset$;

(2) ultraparallel, if after extended to the boundary in the $\mathbb{D}^2$-model, C_1 and C_2 have no intersection;

(3) asymptotic, if C_1 and C_2 have no intersection in $\mathbb{D}^2$ but intersect at boundary.

The three cases divide orientation preserving hyperbolic isometries into three types:

(1) If $C_1 \cap C_2 \neq \emptyset$, we can see from euclidean geometry that the intersection is a unique point in the hyperbolic plane. Then the composition of two reflections about them is a rotation around this point for double angle of their intersection angle. In particular, if both C_1 and C_2 are diameters in the $\mathbb{D}^2$-model or equivalently, intersect at i in the $\mathbb{H}^2$-model, their composition is just the standard rotation about the origin in $\mathbb{D}^2$.

(2) If C_1 and C_2 are ultraparallel, then the composition of two reflections about them is a translation of double hyperbolic distance between them. In particular, if both C_1 and C_2 are center at 0, their composition is just the dilatation.

(3) If C_1 and C_2 are asymptotic, then the composition of two reflections about them is a limit rotation. In particular, if C_1 intersects C_2 at i in the $\mathbb{D}^2$-model, this is t_c where c is the double of the euclidean distance between C_1, C_2.

This leads to the following classification of hyperbolic isometries.

Theorem 4.16 (Classification of hyperbolic isometries) *Each $\mathbb{H}^2$-isometry is either a*

(1) *rotation;*

(2) *limit rotation;*

(3) *translation; or*

(4) *glide reflection.*

Proof We first deal with orientation preserving case. Any orientation preserving isometry can be written as a composition of two reflections. Let's use L, M to denote the two reflection hyperbolic lines. We need to consider the following three cases:

(1) $L \cap M \neq \emptyset$. From the euclidean geometry, two semicircles centered at x-axis, or a semicircle and a line parallel to y-axis intersect at only one point. We can show that this isometry must be the rotation with respect to this intersection point.

(2) L and M are asymptotic, which means that L and M has no intersection in $\mathbb{H}^2$, but if we look at them in the $\mathbb{D}^2$-model, they intersect at the boundary of the unit disk. For this case, we can show that their composition is a limit rotation.

(3) L and M has no intersection even with taking infinity into consideration. For this case, we can show that under good coordinates, it is a dilatation and hence a $\mathbb{H}^2$-translation.

For the orientation reversing case, we can show that it must be a glide reflection.

□

4.7 The Gauss–Bonnet formula for the hyperbolic case

5 Hyperbolic surfaces

No time to finish, so I only sketch some main results.