Classical geometry

Danny Calegari

5/4/2005, Version 0.14

Contents

List of Figures

Chapter 1

A crash course in group theory

1.1 Groups

1.1.1 Definition of a group

Mathematical objects are defined in terms of the properties that they satisfy. A *structure* on an object is a list of properties, or requirements for the object to satisfy. Sometimes this list of properties determines the object uniquely, sometimes not. Sometimes the same object can be described in many different ways.

The *symmetries* of an object are the transformations which preserve some (usually implicit) structure.

A *group* is an algebraic object which formalizes the mathematical notion of symmetry. We start with an abstract definition.

Definition 1.1.1. A *group* is a set G and an operation $m : G \times G \to G$ called *multiplication* with the following properties:

1. m is *associative*. That is, for any $a, b, c \in G$,

   $$m(a, m(b, c)) = m(m(a, b), c)$$

   and the product can be written unambiguously as abc.

2. There is a unique element $e \in G$ called *the identity* with the properties that, for any $a \in G$,

   $$ae = ea = a$$

3. For any $a \in G$ there is a unique element in G denoted a^{-1} called *the inverse* of a such that

   $$aa^{-1} = a^{-1}a = e$$

Given an object with some structural qualities, we can study the *group of all symmetries* of that object; namely, the set of transformations of the object to itself which preserve the structure in question, with composition as the group operation. Obviously, symmetries can be composed associatively, since the effect of a symmetry on the object doesn't depend on what sequence of symmetries we applied to the object in the past. Moreover, the transformation which does nothing preserves the structure of the object. Finally, symmetries are reversible — performing the opposite of a symmetry is itself a symmetry. Thus, the symmetries of an object (also called the *automorphisms* of an object) are an example of a group.

The power of the abstract idea of a group is that the symmetries can be studied by themselves, without requiring them to be tied to the object they are transforming. So for instance, the same group can act by symmetries of many different objects, or on the same object in many different ways.

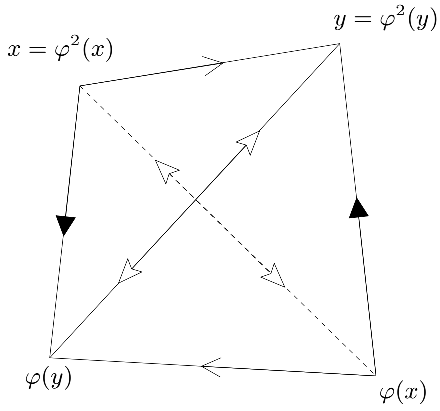

Figure 1.1: An element φ of the Klein 4–group interchanges two pairs of opposite edges, and flips the orientation of the other two edges.

Example 1.1.2. The group with only one element e and multiplication $e \times e = e$ is called the *trivial group*.

Example 1.1.3. The integers $\mathbb{Z}$ with $m(a, b) = a + b$ is a group, with identity 0.

Example 1.1.4. The positive real numbers $\mathbb{R}^+$ with $\mathrm{m}(a, b) = ab$ is a group, with identity 1.

Example 1.1.5. The group with two elements $even$ and odd and "multiplication" given by the usual rules of addition of even and odd numbers; here $even$ is the identity element. This group is denoted $\mathbb{Z}/2\mathbb{Z}$.

Example 1.1.6. The group of integers mod n is a group with $m(a, b) = a + b \mod n$ and identity 0. This group is denoted $\mathbb{Z}/n\mathbb{Z}$ and also by C_n, the *cyclic group of order* n.

Notice that the group $\mathbb{Z}/n\mathbb{Z}$ can also be thought of as the group of *rotations* of a regular n–gon.

Definition 1.1.7. If G and H are groups, one can form the *Cartesian product*, denoted $G \oplus H$. This is a group whose elements consist of the set of pairs

$$G \times H = \{(g, h) | g \in G, h \in H\}$$

where $m : (G \times H) \times (G \times H) \to G \times H$ is defined by

$$m((g_1, h_1), (g_2, h_2)) = (m_G(g_1, g_2), m_H(h_1, h_2))$$

The identity element is (e_G, e_H).

Example 1.1.8. Let S be a regular tetrahedron; label opposite pairs of edges by A, B, C. Then the group of symmetries which preserves the labels is $\mathbb{Z}/2\mathbb{Z} \oplus \mathbb{Z}/2\mathbb{Z}$. It is also known as the *Klein* 4*–group*. The action of a typical group element φ is shown in figure 1.1.

1.1.2 Abelian and nonabelian groups

In all of the examples above, $m(a, b) = m(b, a)$. A group with this property is called *commutative* or *Abelian*. Not all groups are Abelian!

Example 1.1.9. Let T be an equilateral triangle with sides A, B, C opposite vertices a, b, c in anticlockwise order. The symmetries of T are the reflections in the lines running from the corners to the midpoints of opposite sides, and the rotations. There are three possible rotations, through anticlockwise angles $0, 2\pi/3, 4\pi/3$ which can be thought of as e, ω, ω^2. Observe that $\omega^{-1} = \omega^2$. Let r_a be a reflection through the line from the vertex a to the midpoint of A. Then $r_a = r_a^{-1}$ and similarly for r_b, r_c. Then $\omega^{-1} r_a \omega = r_c$ but $r_a \omega^{-1} \omega = r_a$ so this group is *not commutative*. It is callec the *dihedral group* D_3 and has 6 elements.

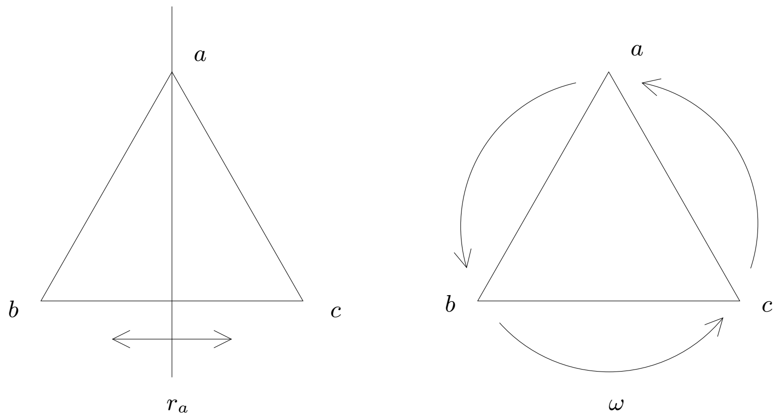

Figure 1.2: The composition $\omega^{-1} r_a \omega = r_c$, but $r_a \omega^{-1} \omega = r_a$, so D_3 is not commutative.

Example 1.1.10. If P is an equilateral n–gon, the symmetries are reflections as above and rotations. This is called the *dihedral group* D_n and has $2n$ elements. The elements are $e, \omega, \omega^2, \ldots, \omega^{n-1} = \omega^{-1}$ and $r_1, r_2, \ldots, r_n$ where $r_i^2 = e$ for all i, $r_i r_j = \omega^{2(i-j)}$ and $\omega^{-1} r_i \omega = r_{i-1}$.

Example 1.1.11. The symmetries of an "equilateral ∞–gon" (i.e. the unique infinite 2–valent tree) defines a group D_∞, the *infinite dihedral group*.

Example 1.1.12. The set of 2×2 matrices whose entries are real numbers and whose determinants do not vanish is a group, where multiplication is the usual multiplication of matrices. The set of *all* 2×2 matrices is *not* naturally a group, since some matrices are not invertible.

Example 1.1.13. The group of permutations of the set $\{1 \ldots n\}$ is called the *symmetric group* S_n. A permutation breaks the set up into subsets on which it acts by cycling the members. For example, $(3, 2, 4)(5, 1)$ denotes the element of S_5 which takes $1 \to 5, 2 \to 4, 3 \to 2, 4 \to 3, 5 \to 1$. The group S_n has $n!$ elements. A *transposition* is a permutation which interchanges exactly two elements. A permutation is *even* if it can be written as a product of an even number of transpositions, and *odd* otherwise.

Exercise 1.1.14. *Show that the symmetric group is not commutative for $n > 2$. Identify S_3 and S_4 as groups of rigid motions of familiar objects. Show that an even permutation is not an odd permutation, and vice versa.*

1.2 Subgroups, quotients, homomorphisms

Definition 1.2.1. A *subgroup* H of G is a subset such that if $h \in H$ then $h^{-1} \in H$, and if $h_1, h_2 \in H$ then $h_1 h_2 \in H$. With its inherited multiplication operation from G, H is a group. The *right cosets* of H in G are the equivalence classes $[g]$ of elements $g \in G$ where the equivalence relation is given by $g_1 \sim g_2$ if and only if there is an $h \in H$ with $g_1 = g_2 h$.

If G is the group of symmetries of a structure on an object X, a subgroup H of G corresponds to the group of symmetries of a more rigid structure on X, called a *stiffening*.

For example, the group D_3 of symmetries of an equilateral triangle contains the cyclic subgroup C_3 of *orientation preserving* symmetries of the triangle. The geometric structure of the triangle given by lengths and angles is stiffened by prescribing an *orientation* on the angles. We will see many more sophisticated examples in the sequel.

Exercise 1.2.2. *If H is finite, the number of elements of G in each (right or left) coset are equal to $|H|$, the number of elements in H. Consequently, if $|G|$ is finite, $|H|$ divides $|G|$.*

Exercise 1.2.3. *Show that the subset of even permutations is a subgroup of the symmetric group, known as the* alternating group *and denoted A_n. Identify A_5 as a group of rigid motions of a familiar object.*

Example 1.2.4. Given a collection of elements $\{g_i\} \subset G$ (not necessarily finite or even countable), the *subgroup generated by the g_i* is the subgroup whose elements are obtained by multiplying together *finitely* many of the g_i and their inverses in some order.

Exercise 1.2.5. *Why are only finite multiplications allowed in defining subgroups? Show that a group in which infinite multiplication makes sense is a trivial group. This fact is not as useless as it might seem . . .*

Definition 1.2.6. A group is *cyclic* if it is generated by a single element. This justifies the notation C_n for $\mathbb{Z}/n\mathbb{Z}$ used before.

Definition 1.2.7. A *homomorphism* between groups is a map $f : G_1 \to G_2$ such that $f(g_1)f(g_2) = f(g_1g_2)$ for any g_1, g_2 in G_1. The *kernel* of a homomorphism is the subgroup $K \subset G_1$ defined by $K = f^{-1}(e)$. If $K = e$ then we say f is *injective*. If every element of G_2 is in the image of f, we say it is *surjective*. A homomorphism which is injective and surjective is called an *isomorphism*.

Example 1.2.8. Every finite group G is isomorphic to a subgroup of S_n where n is the number of elements in G. For, let $b : G \to \{1, \dots, n\}$ be a bijection, and identify an element g with the permutation which takes $b(h) \to b(gh)$ for all h.

Definition 1.2.9. An *exact sequence* of groups is a (possibly terminating in either direction) sequence

$$\cdots \to G_i \to G_{i+1} \to G_{i+2} \to \dots$$

joined by a sequence of homomorphisms $h_i : G_i \to G_{i+1}$ such that the *image* of h_i is equal to the *kernel* of h_{i+1} for each i.

Definition 1.2.10. If $a, b \in G$, then bab^{-1} is called *the conjugate* of a by b, and $aba^{-1}b^{-1}$ is called *the commutator* of a and b. Abelian groups are characterized by the property that a conjugate of a is equal to a and every commutator is trivial.

Definition 1.2.11. A subgroup $N \subset G$ is *normal*, denoted $N \lhd G$ if for any $n \in N$ and $g \in G$ we have $gng^{-1} \in N$. A kernel of a homomorphism is normal. Conversely, if N is normal, we can define the *quotient group* G/N whose elements are equivalence classes $[g]$ of elements in G, and two elements g, h are equivalent iff $g = hn$ for some $n \in N$. The multiplication is given by $m([g], [h]) = [gh]$ and the fact that N is normal says this is well–defined. Thus normal subgroups are exactly kernels of homomorphisms.

Example 1.2.12. Any subgroup of an abelian group is normal.

Example 1.2.13. $\mathbb{Z}$ is a normal subgroup of $\mathbb{R}$. The quotient group $\mathbb{R}/\mathbb{Z}$ is also called *the circle group* S^1. Can you see why?

Example 1.2.14. Let D_n be the dihedral group, and let C_n be the subgroup generated by ω. Then C_n is normal, and $D_n/C_n \cong \mathbb{Z}/2\mathbb{Z}$.

Definition 1.2.15. If G is a group, the subgroup G_1 generated by the commutators in G is called the *commutator subgroup* of G. Let G_2 be the subgroup generated by commutators of elements of G with elements of G_1. We denote $G_1 = [G, G]$ and $G_2 = [G, G_1]$. Define G_i inductively by $G_i = [G, G_{i-1}]$. The elements of G_i are the elements which can be written as products of iterated commutators of length i. If G_i is trivial for some i — that is, there is some i such that every commutator of length i in G is trivial — we say G is *nilpotent*.

Observe that every G_i is normal, and every quotient G/G_i is nilpotent.

Definition 1.2.16. If G is a group, let $G^0 = G_1$ and define $G^i = [G^{i-1}, G^{i-1}]$. If G^i is trivial for some i, we say that G is *solvable*. Again, every G^i is normal and every G/G^i is solvable. Obviously a nilpotent group is solvable.

Definition 1.2.17. An isomorphism of a group G to itself is called an *automorphism*. The set of automorphisms of G is naturally a group, denoted $\mathrm{Aut}(G)$. There is a homomorphism from $\rho : G \to \mathrm{Aut}(G)$ where g goes to the automorphism consisting of *conjugation by* g. That is, $\rho(g)(h) = ghg^{-1}$ for any $h \in G$. The automorphisms in the image of ρ are called *inner automorphisms*, and are denoted by $\mathrm{Inn}(G)$. They form a normal subgroup of $\mathrm{Aut}(G)$. The quotient group is called the group of *outer automorphisms* and is denote by $\mathrm{Out}(G) = \mathrm{Aut}(G)/\mathrm{Inn}(G)$.

1.3 Semi–direct products

Definition 1.3.1. Suppose we have two groups G, H and a homomorphism $\rho : G \to \mathrm{Aut}(H)$. Then we can form a new group called *the semi–direct product of G and H* denoted $G \ltimes H$ whose elements are the elements of $G \times H$ and multiplication is given by

$$m((g_1, h_1), (g_2, h_2)) = (g_1 g_2, h_1 \rho(g_1)(h_2))$$

Observe that H is a normal subgroup of $G \ltimes H$, and there is an exact sequence

$$1 \to H \to G \ltimes H \to G \to 1$$

Example 1.3.2. The dihedral group D_n is equal to $\mathbb{Z}/2\mathbb{Z} \ltimes C_n$ where the homomorphism $\rho : \mathbb{Z}/2\mathbb{Z} \to \mathrm{Aut}(C_n)$ takes the generator of $\mathbb{Z}/2\mathbb{Z}$ to the automorphism $\omega \to \omega^{-1}$, where ω denotes the generator of C_n.

Example 1.3.3. The group $\mathbb{Z}/2\mathbb{Z} \ltimes \mathbb{R}$ where the nontrivial element of $\mathbb{Z}/2\mathbb{Z}$ acts on $\mathbb{R}$ by $x \to -x$ is isomorphic to the group of *isometries* (i.e. 1–1 and distance preserving transformations) of the real line. It contains D_∞ as a subgroup.

Exercise 1.3.4. *Find an action of $\mathbb{Z}/2\mathbb{Z}$ on the group S^1 so that D_n is a subgroup of $\mathbb{Z}/2\mathbb{Z} \ltimes S^1$ for every n.*

1.4 Groups defined by generators and relations

Definition 1.4.1. Let A be a set. The *free group generated by A*, denoted F_A, is defined by the following two properties:

1. There is an inclusion of sets $i : A \to F_A$.
2. For any group H and any map of sets $\varphi : A \to H$ there is a unique homomorphism

   $$F_\varphi : F_A \to H$$

   satisfying

   $$\varphi \circ i = F_\varphi|_A$$

Exercise 1.4.2. *Show that if such an F_A exists, the defining properties determine it uniquely up to isomorphism.*

Lemma 1.4.3. *F_A exists for any A.*

Proof. Let T be the set $A \times \pm 1$. By abuse of notation, we denote elements of T by $(a, 1) = a$ and $(a, -1) = a^{-1}$ for $a \in A$. Let T^* denote the set of finite words in the letters of T; that is, finite words in symbols $a \in A$ and $a^{-1}, a \in A$. We define an equivalence relation $\sim$ on T^* by declaring two words w, w' to be equal if one can be obtained from the other by inserting or removing a subword of the form aa^{-1} or $a^{-1}a$ for any $a \in A$. So, for example, if $a, b \in A$,

$$ab^{-1}aa^{-1}bb \sim ab^{-1}bb \sim ab$$

Then $F_A = T^*/\sim$ is a group, with concatenation of words as multiplication, and the empty word as the identity. For each equivalence class of word

$$w = a_1^{e_1} a_2^{e_2} \dots a_n^{e_n}$$

where $e_i = \pm 1$, the inverse is the equivalence class of

$$w^{-1} = a_n^{-e_n} a_{n-1}^{-e_{n-1}} \dots a_1^{-e_1}$$

One can check that this is a group, and that it satisfies the defining properties. The inclusion $i : A \to F_A$ sends each element $a \in A$ to the equivalence class of the word a. $\square$

Definition 1.4.4. If A is a set with n elements, for n a positive integer, then F_A is denoted F_n, and is determined up to isomorphism by n.

Notice that $F_1 = \mathbb{Z}$.

Definition 1.4.5. A group G is *finitely generated* if there is a finite subset of G which generates G.

Exercise 1.4.6. *Let G be an arbitrary finitely generated group. Show that there is a surjective homomorphism $F_n \to G$ for some n.*

Example 1.4.7. If we have an alphabet consisting of letters $a_1, \dots, a_n$ and their inverses, we can consider a collection of words in these letters $r_1, \dots, r_m$. If R denotes the subgroup of F_n generated by the r_i and all their conjugates, then R is a normal subroup of F_n and we can form the quotient F_n/R. This is denoted by

$$\langle a_1, \dots, a_n | r_1, \dots, r_m \rangle$$

and an equivalent description is that it is the group whose elements are words in the a_i and their inverses modulo the equivalence relation that two words are equivalent if they are equivalent in the free group, or if one can be obtained from the other by inserting or deleting some r_i or its inverse as a subword somewhere. The a_i are the *generators* and the r_i the *relations*. Groups defined this way are very important in topology. Notice that a *presentation* of a group in terms of generators and relations is far from unique.

Definition 1.4.8. A group G is *finitely presented* if it can be expressed as $\langle A | R \rangle$ for some *finite* set of generators A and relations R.

Exercise 1.4.9. *Let G be any finite group. Show that G is finitely presented.*

Exercise 1.4.10. *Let F_2 be the free group on generators x, y. Let $i : F_2 \to \mathbb{Z}$ be the homomorphism which takes $x \to 1$ and $y \to 1$. Show that the kernel of i is not finitely generated.*

Exercise 1.4.11. *(Harder). Let $i : F_2 \oplus F_2 \to \mathbb{Z}$ be the homomorphism which restricts on either factor to i in the previous exercise. Show that the kernel of i is finitely generated but* not *finitely presented.*

Definition 1.4.12. Given groups G, H the *free product of G and H*, denoted $G * H$, is the group of words whose letters alternate between elements of G and H, with concatenation as multiplication, and the obvious proviso that the identity is in either G or H. It is the unique group with the *universal property* that there are injective homomorphisms $i_G : G \to G * H$ and $i_H : H \to G * H$, and given any other group I and homomorphisms $j_G : G \to I$ and $j_H : H \to I$ there is a *unique* homomorphism c from $G * H$ to I satisfying $c \circ i_G = j_G$ and $c \circ i_H = j_H$.

Exercise 1.4.13. *Show that $*$ defines an associative and commutative product on groups up to isomorphism, and*

$$F_n = \mathbb{Z} * \mathbb{Z} * \cdots * \mathbb{Z}$$

where we take n copies of $\mathbb{Z}$ in the product above.

Exercise 1.4.14. *Show that $\mathbb{Z}/2\mathbb{Z} * \mathbb{Z}/2\mathbb{Z} \cong D_\infty$.*

Remark 1.4.15. Actually, one can extend $*$ to *infinite* (even uncountable) products of groups by the universal property. If one has an arbitrary set A and a partition of A into disjoint subsets A_α for $\alpha \in \Sigma$ some index set, then

$$*_{\alpha \in \Sigma} F_{A_\alpha} \cong F_A$$

In particular, any free group is the free product of some number (possibly uncountable) of copies of $\mathbb{Z}$.

Exercise 1.4.16. *(Hard). Every subgroup of a free group is free.*

Chapter 2

Fundamental ideas in geometry

2.1 Klein's "Erlanger Programm"

At an address at Friedrich–Alexander–Universitaet in Erlangen Germany on December 17 1872, Felix Klein proposed a program to unify the study of geometry by the use of algebraic methods, more specifically, by the use of *group theory*. In particular, the geometrical properties of a space can be understood and explored by a study of the *symmetries* of that space. These symmetries can be organized into a natural algebraic object — a group. Conversely, this group can often be given a natural geometric structure, and investigated in its turn as a geometric space! The interplay between geometry and algebra leads to an enrichment of both structures.

2.2 Category theory

Example 2.2.1. This is not really an example, but rather a *template* for the examples we will meet that fit into Klein's program. We are given a space X together with some sort of structure. A *structure–preserving map* from X to itself is called a *morphism*. The map from X to itself which does nothing is a distinguished morphism, the *identity morphism*, denoted $\mathbf{1}_X$. A morphism f is *invertible* if there is another morphism f^{-1} such that $f \circ f^{-1} = f^{-1} \circ f = \mathbf{1}_X$. The invertible morphisms are also called *automorphisms*. The set of automorphisms of X is a group called $\mathrm{Aut}(X)$, with $\mathbf{1}_X$ as the identity, and composition as multiplication. Observe that a structure on a space can be *defined* by the admissible morphisms. This is a simple example of what is known as a *category*; in particular, it is a category with one object X.

Definition 2.2.2. More generally, a *category* can be thought of as a collection of *objects* (denoted $\mathscr{O}$) and a collection of *morphisms* or admissible maps between objects (denoted $\mathscr{M}$). Every morphism m has a *source* object $s(m)$ and a *target* object $t(m)$, which might be the same object. For every object o, there is a special morphism called the *identity* morphism

$$\mathbf{1}_o : o \to o$$

which acts like the usual identity: i.e.

$$\mathbf{1}_x m = m \text{ for any } m \text{ with } t(m) = x$$

$$m\mathbf{1}_x = m \text{ for any } m \text{ with } s(m) = x$$

The composition of two morphisms is another morphism, and this composition is associative; composition can be expressed as a function $c : \mathscr{M} \times \mathscr{M} \to \mathscr{M}$. That is, c satisfies

$$c(m, c(n, r)) = c(c(m, n), r) \text{ for any } m, n, r \in \mathscr{M}$$

Sometimes a category is written as a 5–tuple $(\mathscr{O}, \mathscr{M}, s, t, c)$, but in practice it is frequently sufficient to specify the objects and the morphisms.

A category is something like a *class* in an object–oriented programming language like C++; one defines at the same time the *data types* (the objects in the category) and the *admissible functions* which operate on them (the morphisms).

Example 2.2.3. The category whose objects are all sets and whose morphisms are all functions between sets is a category called $\mathbf{SET}$. If X is an object in $\mathbf{SET}$ (i.e. a set) then Aut(X) is the group of permutations of X.

Example 2.2.4. The category whose objects are all groups and whose morphisms are all homomorphisms between groups is called $\mathbf{GROUP}$.

Example 2.2.5. A group is itself a category with one object in which every morphism is invertible. A category in which every morphism is invertible, but which may have more than one object, is called a *groupoid*.

A very readable introduction to category theory, with numerous exercises, are the notes by John Stallings [9].

2.3 Topology

Basic notions of incidence or connectivity are part of our fundamental geometric intution. Concepts such as "inside" and "outside", or "bounded" and "unbounded" are topological. Topology can be thought of as the study of the *qualitative* properties of a space that are left unchanged under continuous deformations of the space; that is, deformations which may bend or stretch the space but do not cut or tear it. An allowable morphism between topological spaces is just a continuous map; invertible morphisms are called *homeomorphisms*.

Continuous maps

Topologists frequently discuss very abstract spaces. The concept of continuity in this general context relies on the definition of the following structure on a space.

Definition 2.3.1. A *topology* on a set X is a collection of subsets of X

$$\mathscr{U} \subset \{U \subset X\}$$

with the following properties:

- The empty set and X are both in $\mathscr{U}$.
- If $U_1, \ldots, U_n$ are a finite collection of elements in $\mathscr{U}$ then $\cap_i U_i$ is in $\mathscr{U}$.
- If $\mathscr{V} \subset \mathscr{U}$ is an arbitrary collection of elements in $\mathscr{U}$, then $\cup_{V \in \mathscr{V}} V$ is in $\mathscr{U}$.

Thus, a topology is a system of subsets X which includes the empty set and X, and is closed under finite intersections and arbitrary unions. The sets in $\mathscr{U}$ are called the *open* sets in X. Their complements are called the *closed* sets. From the definition, finite unions and arbitrary intersections of closed sets are closed.

Remark 2.3.2. It suffices, when defining a topology, to give a set of subsets of the space which are supposed to be open, and then let the open sets be the smallest collection of subsets, including the given sets, which satisfy the axioms of a topology. We call the given colletion of sets a *basis* for the topology. Most of the spaces we will encounter have a *countable* basis.

Definition 2.3.3. Given a set $Y \subset X$, the *closure* of Y is the intersection of the closed subsets of X containing Y. Given a set Y, the *interior* of Y is the complement of the closure of the complement of Y. It is the union of all the open sets in X contained in Y.

Example 2.3.4. Let $Y \subset X$ be a subset. The *subspace* topology on Y is the topology whose open sets are the intersections $U \cup Y$ where U is open in X.

We will not discuss topological spaces in general in the sequel and stick only to some very concrete examples.

Let's suppose we have two spaces X and Y where we understand what the open sets are. For instance, in $\mathbb{R}$ the open sets are the unions of intervals of the kind (a, b), where we don't include the endpoints. In this context we can define abstractly what is meant by a continuous map.

Definition 2.3.5. A map from X to Y is *continuous* if the inverse of any open set is open.

Example 2.3.6. Let $\sim$ be an equivalence relation on X, and let $\pi : X \to X/\sim$ be the quotient map to the space of equivalence classes. The *quotient* topology on $X/\sim$ is the topology whose open sets are those $U \subset X/\sim$ such that $\pi^{-1}(U)$ is open in X. Thus, $X/\sim$ has as many open sets as it is allowed subject to the condition that π is continuous.

Definition 2.3.7. A homeomorphism from X to Y is a continuous map which is invertible and has a continuous inverse.

The category of topological spaces is denoted $\mathbf{TOP}$ and has as objects all topological spaces and as morphisms all continuous maps. The group of all self–homeomorphisms of a space is frequently very large and hard to understand.

We list some frequently encountered topological concepts:

Definition 2.3.8. A *neighborhood* of a point $p \in X$ is any open set $U \in X$ containing p.

Definition 2.3.9. A topological space is *Hausdorff* if for any two distinct points $p, q \in X$ there are neighborhoods of p and of q which are disjoint.

One of the most significant class of topological spaces are the *manifolds*.

Definition 2.3.10. A *manifold* is a Hausdorff topological space with a countable basis, such that every point has a neighborhood homeomorphic to an open subset of $\mathbb{R}^n$ for some (usually fixed) n.

Informally, a manifold is a topological space which looks locally like n–dimensional Euclidean space $\mathbb{R}^n$.

Example 2.3.11. Any open subset of $\mathbb{R}^n$ is a manifold.

Definition 2.3.12. A space X is *connected* if there are no proper nonempty subsets $U \subset X$ which are both *closed* and *open*. A space is *locally connected* if for any point p and any open set $O \subset X$ there is another $U \subset O$ such that U is connected, as a subspace of X. A space is *path connected* if for any two points $p, q \in X$ there is a continuous map $\varphi : I \to X$ where I is the unit interval, such that $\varphi(0) = p$ and $\varphi(1) = q$.

Definition 2.3.13. A subset $X \subset Y$ (perhaps all of Y) is *compact* if it is closed, and for every collection $\mathscr{U}$ of open sets in Y whose union contains X, there is a *finite subcollection* whose union contains X. A space X is *locally compact* if every $p \in X$ has a neighborhood whose closure is compact.

2.4 Metric geometry

One of our basic intuitions in geometry is that of *distance*. In fact the word geometry literally means "measuring the earth". Metric geometry is the study of the concept of distance, and its various generalizations and abstractions. A beautiful (but quite advanced) reference for this subject is [3].

Metric spaces

Definition 2.4.1. A *metric space* X, d is a set X together with a function

$$d : X \times X \to \mathbb{R}_0^+$$

where $\mathbb{R}_0^+$ denotes the non–negative real numbers, with the following properties:

1. d is *symmetric*. That is,
$$d(x, y) = d(y, x)$$

2. d is *nondegenerate* in the sense that
$$d(x, y) = 0 \text{ iff } x = y$$

3. d satisfies the *triangle inequality*. That is,

$$d(x, y) + d(y, z) \geq d(x, z)$$

for all triples $x, y, z \in X$.

Example 2.4.2. The real line $\mathbb{R}$ is a metric space with

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

Example 2.4.3. The plane $\mathbb{R}^2$ is a metric space with

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

Example 2.4.4. The plane $\mathbb{R}^2$ is a metric space with

$$d((x_1, y_1), (x_2, y_2)) = |x_1 - x_2| + |y_1 - y_2|$$

This metric is known as the *Manhattan metric*. Can you see why?

A metric d on a space X naturally induces a topology. A basis for the topology on X is given by the balls $B_r(p)$ of radius r, centered at arbitrary points $p \in X$. Here

$$B_r(p) = \{q \in X | d(p, q) < r\}$$

A map $f : X \to Y$ between metric spaces is *continuous at a point* $p \in X$, with respect to the topologies induced by the metrics, if for every $\epsilon > 0$ there is a $\delta > 0$ such that if $q \in X$ satisfies $d_X(p, q) < \epsilon$, then $d_Y(f(p), f(q)) < \delta$. f is *continuous* if it is continuous for every $p \in X$. However, the relationship between ϵ and δ may certainly depend on the choice of point p.

Not every topological structure on a space X is induced by some metric. A topology is said to be *metrizable* if it arises in this way. In general, a topological space with a countable basis is metrizable if for any closed set C in X or any point $q \in X$, and any point p not in C or equal to q, there are disjoint open sets in X, one containing C or q and one containing p. This statement is a little tortuous, since we cannot assume *a priori* that the points p of X are *closed*.

Definition 2.4.5. An *isometry* of a metric space X is a 1–1 and onto transformation of X to itself which preserves distances between points. The set of *isometries* of a space X is a group $\mathrm{Isom}(X)$, where multiplication in the group is composition of symmetries, and e is the trivial symmetry which fixes every x in X. This is an example of a group of the form $\mathrm{Aut}(X)$ where the relevant structure on X is that of the *category of metric spaces* $\mathbf{MET}$ whose objects are metric spaces and whose morphisms are isometries.

Isometries are continuous with respect to the induced topology. We say therefore that a metric structure is a *stiffening* of the topological structure it induces.

Definition 2.4.6. Isometries are frequently too restrictive for many circumstances; a typical metric space of study might admit no non–trivial isometries at all. We can enrich the structure by allowing as morphisms those maps which, though they don't literally *preserve* distances between points, at least don't increase distances between points by too much. Such a map is called a *Lipschitz* map, and metric spaces with these as morphisms define a category $\mathbf{LIP}$ which is in many ways a much more interesting object than $\mathbf{MET}$.

A map $f : X \to Y$ between metric spaces is *bilipschitz* if there is a $K > 1$ so that

$$\frac{1}{K} d_Y(f(x), f(y)) \leq d_X(x, y) \leq K d_Y(f(x), f(y))$$

One may think of this as a map which only distorts distances up to a bounded factor. A bilipschitz map is 1–1, since metrics are nondegenerate. An invertible Lipschitz map with Lipschitz inverse is bilipschitz, so that the automorphisms in the category $\mathbf{LIP}$ are bilipschitz. Moreover, the composition of two bilipshitz maps is bilipschitz.

A metric structure is a stiffening of a lipschitz structure.

Exercise 2.4.7. *Show that the bilipschitz self–maps of* $\mathbb{R}^2$ *with the Euclidean or the Manhattan metric are the same.*

2.5 Differential geometry

Differential geometry is the abstraction of calculus and analysis on n–dimensional Euclidean space to generalized geometric spaces called "smooth Riemannian manifolds". These are spaces which look like Euclidean space on a small scale, but on larger scales they are deformed or "curved". Einstein's theory of general relativity says that our own universe is a certain kind of curved space, where the curvature is proportional to the strength of the gravitational force; on the human scale it looks Euclidean, but near massive objects like neutron stars, the "curvature" of the space is evidenced by the bending of light rays. The concept of curvature is a very important connection between geometry and topology.

Since calculus and analysis are basically *local*, one can do calculus on such spaces, since on smaller and smaller scales they look more and more like $\mathbb{R}^n$ so that limits, derivatives, differentiability etc. all make sense, and the tools of multivariable calculus can be transplanted to this setting.

The morphisms which preserve the structure used to do differential geometry are the *smooth maps*, generalizations of differentiable functions.

2.5.1 Smooth Manifolds

A *manifold* is a space which, on a small scale, resembles Euclidean space of some dimension. The dimension is usually assumed to be constant over the space, and is called the *dimension of the manifold*.

A circle or a line is an example of a 1–dimensional manifold. A sphere or the surface of a donut is an example of a 2–dimensional manifold. Our universe, or the space outside a knot or link are examples of 3–dimensional manifolds.

Definition 2.5.1. A *smooth manifold* is a manifold on which one can do Calculus. One covers the space with a collection of little snapshots called "charts" which are meant to be all the different possible choices of local parameters for the space. Technically one has a collection of charts, which are subsets U_i of the manifold M, and a collection of ways of parameterizing these charts as subsets of Euclidean space; that is, maps $\phi_i : U_i \to V_i$ which are continuous and have continuous inverses, where V_i is some open region in some $\mathbb{R}^n$. These maps should be compatible, in the sense that if two charts U_i, U_j overlap, the map $\rho = \phi_j \phi_i^{-1}$ between the appropriate subsets of $\mathbb{R}^n$ (where ρ is defined) should be smooth (i.e. it should have continuous partial derivatives of all orders), and it should be *locally invertible*; that is, the matrix of partial derivatives

$$d\rho = \begin{bmatrix} \frac{\partial \rho(x_1)}{\partial x_1} & \frac{\partial \rho(x_2)}{\partial x_1} & \cdots & \frac{\partial \rho(x_n)}{\partial x_1} \\ \frac{\partial \rho(x_1)}{\partial x_2} & \frac{\partial \rho(x_2)}{\partial x_2} & \cdots & \frac{\partial \rho(x_n)}{\partial x_2} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial \rho(x_1)}{\partial x_n} & \frac{\partial \rho(x_2)}{\partial x_n} & \cdots & \frac{\partial \rho(x_n)}{\partial x_n} \end{bmatrix}$$

should be invertible at every point.

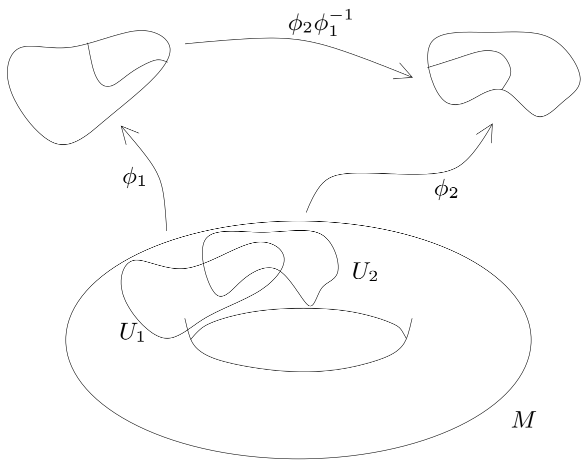

Figure 2.1: The composition $\rho = \phi_2\phi_1^{-1}$ on the overlap $\phi_1(U_1 \cap U_2)$ of charts should be smooth and locally invertible.

This definition seems bulky but it is actually quite elegant. When doing multi-variable calculus, we are used to switching back and forth between local co–ordinates which might only be defined on certain subsets of $\mathbb{R}^n$.

Example 2.5.2. In the plane $\mathbb{R}^2$, we might switch between x, y Cartesian co–ordinates and r, θ polar co–ordinates. Note that θ is not really a "co–ordinate" on the whole of $\mathbb{R}^2$, since its value is only well–defined up to multiples of 2π, and at the origin there is no sensible value for it. These co–ordinates are actually maps from subsets of the manifold $\mathbb{R}^2$ to the "standard" Euclidean space, which in this case also happens to be $\mathbb{R}^2$. The first chart U_1 can be taken to be all of $\mathbb{R}^2$, and the function ϕ_1 is just the identity $\phi_1 : (x, y) \to (x, y)$. The second function ϕ_2 is not defined on all of $\mathbb{R}^2$, and might be given in the chart $U_2 = \{x, y > 1\}$ for instance, by

$$\phi_2 : (x, y) \to \left(\sqrt{x^2 + y^2}, \arctan \frac{x}{y}\right)$$

Defining $\rho = \phi_2 \phi_1^{-1} = \phi_2$ as above, check that $d\rho$ is invertible everywhere in the overlap of the two charts.

The charts on a smooth manifold are just the collection of all the possible local co–ordinates for subsets of the manifold; this collection of charts is called an *atlas*.

Example 2.5.3. The spaces $\mathbb{R}^n$ are smooth manifolds for any n.

Example 2.5.4. An open subset of a smooth manifold is itself a smooth manifold, by restriction of charts and functions.

In differential geometry, the allowable morphisms are typically the *smooth maps*. A map $f : M^m \to N^n$ is *smooth* if for charts $U_i \subset M, U_j \subset N$, the composition $\phi_j \circ f \circ \phi_i^{-1}$ is a smooth map from the appropriate subset of $\mathbb{R}^m$ to the appropriate subset of $\mathbb{R}^n$. That is, the co–ordinate maps have continuous partial derivatives of all orders. The *category of smooth manifolds*, denoted $\mathbf{DIFF}$ has as objects smooth manifolds, and as morphisms smooth maps. An invertible smooth map is called a *diffeomorphism*; the group of diffeomorphisms of a smooth manifold is typically a huge, unmanageable object, but certain features of it can be studied.

2.5.2 Tangent spaces and vector fields

If $U \subset \mathbb{R}^n$ is an open set, for each $p \in U$ we may define the *tangent space T_pU to p* to be the set of vectors in $\mathbb{R}^n$ based at p. This vector space is naturally isomorphic to $\mathbb{R}^n$ itself, by translating the basepoint of each vector to the origin.

If $V \subset \mathbb{R}^n$ is another open set, and $\rho : U \to V$ is a smooth map, then the differential $d\rho$ takes vectors at p to vectors at $\rho(p)$. If ρ is locally invertible, $d\rho : T_pU \to T_qV$ is an *isomorphism* of vector spaces.

Thus it makes sense to define *the tangent space T_pM at a point $p \in M$* for M a smooth manifold, by defining it to be equal to the tangent space to p for any chart U containing p. The *tangent space to M*, denoted TM, is just the union of T_pM for each $p \in M$.

Exercise 2.5.5. *If M is a smooth n–dimensional manifold, show TM is a smooth $2n$–dimensional manifold, and show that the natural embedding $e : M \to TM$ sending*

$$e : p \to 0_p \in T_pM$$

is a smooth map. Show also that the natural projection $\pi : TM \to M$ sending

$$\pi : T_pM \to p$$

is smooth.

Example 2.5.6. A *section* $s : M \to TM$ is a smooth map such that

$$\pi \circ s = \mathrm{id}$$

The image is called a (smooth) *vector field* on M.

The image of M under the embedding e from exercise 2.5.5 is the 0 vector field.

Suppose $f : I \to M$ is a smooth map. Let U be a neighborhood of $f(0)$, and let $\varphi : U \to \mathbb{R}^n$ be a chart on U. Then

$$v_t = \frac{d(\varphi f)(t)}{dt}$$

is a vector in $\mathbb{R}^n$ based at $\varphi f(t)$. That is, it is an element of $T_{\varphi f(t)}\mathbb{R}^n$. Under the identification of $T_{f(t)}$ with $T_{\varphi f(t)}$ by $d\varphi$, we get a vector $w_t \in T_{f(t)}$ which is the *tangent vector to* f *at* t, and is denoted $f'(t)$.

Let $P_p(M)$ denote the space of smooth maps $f : I \to M$ satisfying $f(0) = p$. Then the correspondence $f \to f'(0)$ defines a map

$$P_p(M) \to T_p(M)$$

One may think of $T_p(M)$ as capturing the *tangential part* of a smooth path in M starting at p. Conversely, we can think of T_pM as the set of equivalence classes for a natural equivalence relation on smooth paths starting at p.

Exercise 2.5.7. *Let X be a smooth vector field on M. For each $p \in M$, show that there is a unique smooth map $f : I \to M$ such that $f(0) = p$, and such that $f'(t) = X_{f(t)}$ for all $t \in I$.*

2.5.3 Riemannian manifolds

Smooth manifolds are certainly metrizable. A particularly interesting class of metrics on manifolds are those which arise from *Riemannian structures*.

Definition 2.5.8. A *Riemannian structure* on a smooth manifold M is an assignment of a smoothly varying positive definite symmetric bilinear form g on the tangent space to M at each point. That is, for each $p \in M$, g_p defines a linear map

$$g_p : T_pM \times T_pM \to \mathbb{R}$$

which satisfies the following properties:

1. For each p, if $X_p, Y_p \in T_pM$,

   $$g_p(X_p, Y_p) = g_p(Y_p, X_p)$$

2. For each p,

   $$g_p(X_p, X_p) \geq 0$$

   with equality iff $X_p = 0$

3. If X_p, Y_p are smooth vector fields on M, then

$$f : p \to g_p(X_p, Y_p)$$

is a smooth function on M.

For two vectors $X, Y \in T_pM$, the value $g_p(X, Y)$ is called the *inner product* of X and Y.

Riemannian structures exist on every manifold. A Riemannian structure defines a metric as follows. For each smooth path $f : I \to M$, define the *length* of the path f to be the integral

$$|f| = \int_I \sqrt{g_{f(t)}(f'(t), f'(t))}$$

One can check that length satisfies the following *reparameterization property*: if $f = g \circ h$ for $h : I \to I$ some diffeomorphism, then

$$|f| = |g|$$

Definition 2.5.9. Let g denote a Riemannian structure on M. The *Riemannian metric* associated to g is defined by the following formula:

$$d(x, y) = \inf_f |f|$$

where the infimum is taken over all smooth paths $f : I \to M$ with $f(0) = x$ and $f(1) = y$.

It is not hard to check that this satisfies the axioms of a metric. Moreover, the topology on M induced by this metric agrees with the usual topology of M as a manifold. A Riemannian structure is therefore a stiffening of a smooth structure on a manifold. Unlike the group of diffeomorphisms of M, the group of *isometries* of M with respect to any Riemannian metric admits the structure of a *finite dimensional* smooth manifold.

Definition 2.5.10. With respect to a Riemannian structure g, we can define the *length* of a vector X to be equal to

$$\text{length}(X) = \sqrt{g(X, X)}$$

We also denote the length of X by $\|X\|$. Two vectors X, Y at the same point are *orthogonal* if

$$g(X, Y) = 0$$

An *orthonormal basis* for T_pM is a basis $X_1, X_2, \ldots, X_n$ for T_pM as a vector space, such that $\|X_i\| = 1$ for all i, and X_i, X_j are mutually orthogonal for all $i \neq j$.

Definition 2.5.11. In general, if $X, Y \in T_pM$, the *angle* between X and Y is defined, by analogy with the case of vectors in usual Euclidean space, by the formula:

$$\angle(X, Y) = cos^{-1} \frac{g(X, Y)}{\|X\|\|Y\|}$$

Notation 2.5.12. If g is understood, we denote the inner product of two vectors by

$$g(X, Y) = \langle X, Y \rangle$$

2.5.4 Topological groups

Definition 2.5.13. A *topological group* is a group which is also a space (i.e. we understand what continuous maps of the space are) such that $m : G \times G \to G$ and $i : G \to G$, the multiplication and inverse maps respectively, are *continuous*. If G is a *smooth manifold* and the maps m and i are smooth maps, then G is called a *Lie group*.

Remark 2.5.14. Actually, the usual definition of Lie group requires that G be a *real analytic manifold* and that the maps m and i be real analytic. A real analytic manifold is like a smooth manifold, except that the co–ordinate transformations between charts are required to be real analytic, rather than merely smooth. It turns out that any *connected, locally connected, locally compact* topological group actually admits the structure of a Lie group. In particular, the underlying topological space of the group is actually a manifold.

Example 2.5.15. If M, g is a Riemannian manifold, the group G of isometries of M is a Lie group.

2.6 Axiomatic definition of a geometry

Definition 2.6.1. A *geometry* is a topological space X together with a group G of continuous transformations of X.

In the sequel, we will be almost exclusively interested in the case where X is a smooth manifold, and G is a finite dimensional Lie group of diffeomorphisms of X.

Abstractly, a *stiffening* of a pair X, G is another pair X, H where H is a subgroup of G. Again, we will typically be interested in stiffenings where H is both a subgroup and a (closed) submanifold of G; in other words, that it is a *Lie subgroup*.

Chapter 3

Model geometries

3.1 Euclidean geometry

3.1.1 The Euclidean plane

Euclid's axioms

Notation 3.1.1. The Euclidean plane will be denoted by $\mathbb{E}^2$.

Euclid, who taught at Alexandria in Egypt and lived from about 325 BC to 265 BC, is thought to have written 13 famous mathematical books called the *Elements*. In these are found the earliest (?) historical example of the *axiomatic method*. Euclid proposed 5 postulates or axioms of geometry, from which all true statements about the Euclidean plane were supposed to inevitably follow. These axioms were as follows:

1. A straight line segment can be drawn joining any two points.
2. Any straight line segment is contained in a unique straight line.
3. Given any straight line segment, a circle can be drawn having the segment as radius and one endpoint as center.
4. All right angles are congruent.
5. One and only one line can be drawn through a point parallel to a given line.

The terms *point, line, plane* are supposed to be primitive concepts, in the sense that they can't be described in terms of simpler concepts. Since they are not defined, one is not supposed to use one's personal notions or intuitions about these objects to prove theorems about them; one strategy to achieve this end is to replace the terms by other terms (Hilbert's suggestion is *glass, beer mat, table*; Queneau's is *word, sentence, paragraph*) or even nonsense terms. The point is not that intuition is worthless (it is *not*), but that by proving theorems about objects by only using the properties expressed in a list of axioms, the proof immediately applies to any other objects which satisfy the same list of axioms, including collections of objects that one might not have originally had in mind. In this way, our ordinary geometric intuitions of space and movement can be used to reason about objects far from our immediate experience. One important remark to make is that, by modern standards, Euclid's foundations are far from rigorous. For instance, it is implicit in the statement of the axioms that *angles can be added*, but nowhere is it said what properties this addition satisfies; angles are *not* numbers, neither are lengths, but they have properties in common with them.

A closer look at the fourth postulate

Notice that Euclid does not define "congruence". A working definition is that two figures X and Y in a space Z are *congruent* if there is a transformation of Z which takes X to Y. But which transformations are allowed? By including certain kinds of transformations and excluding others, we can drastically affect the flavor of the geometry in question. If not enough transformations are allowed, distinct objects are incomparable and one cannot say anything meaningful about them. If too many transformations are allowed, differences collapse and the supply of distinct objects to investigate dries up. On first approximation, the fourth postulate says that space is *homogeneous*: that is, the properties of an object do not depend on where it is placed in space. But more than this, the postulate says that space is *isotropic*: not only does space look the same at every point, but it also looks the same in every *direction*. Most of the spaces we will encounter in the sequel will be isotropic.

We formalize these notions.

Definition 3.1.2. A geometry X, G is *homogeneous* if for any two points $p, q \in X$ there is a symmetry $\varphi \in G$ of X with $\varphi(p) = q$.

Definition 3.1.3. A geometry M, G where M is a Riemannian manifold is *isotropic* if for any two points p, q and orthonormal bases $X_1, \ldots, X_n \in T_pM$ and $Y_1, \ldots, Y_n \in T_qM$ there is a symmetry $\varphi \in G$ of M with $\varphi(p) = q$, and satisfying $d\varphi(X_i) = Y_i$ for all i.

Notice that homogeneity makes sense in a very general context, whereas isotropy makes sense only in the context of a Riemannian manifold, where lengths and angles of vectors are defined.

A closer look at the parallel postulate

The fifth axiom above is also known as the *parallel postulate*. To decode it, one needs a workable definition of parallel. The "usual" definition is that two distinct lines are parallel if and only if they do not intersect. So the postulate says that given a line l and a point p disjoint from l, there is a unique line l_p through p such that l_p and l are disjoint. Historically, this axiom was seen as unsatisfying, and much effort was put into attempts to show that it followed inevitably as a consequence of the other four axioms. Such an attempt was doomed to failure, for the simple reason that there are interpretations of the "undefined concepts" point, line, plane which satisfy the first four axioms but which do *not* satisfy the fifth. If we say that given l and p there is *no* line l_p through p which does not intersect l, we get *elliptic geometry*. If we say that given

l and p there are *infinitely many* lines l_p through p which do not intersect l, we get *hyperbolic geometry*. Together with Euclidean geometry, these geometries will be the main focus of this course.

Parallel transport

For more general Riemannian manifolds, the notion of parallelism is a local one. We can use a Riemannian structure to define straight lines, known in this context as geodesics, and a conformal structure to define angles between straight lines passing through the same point. If γ_1, γ_2 are two geodesics which make the same angle α with a third geodesic τ, then we can say that γ_2 is obtained by *parallel transport* of γ_1 along τ. Note that if τ' is another geodesic which intersects both γ_1 and γ_2, then in general the angle it makes with these geodesics will be different. One can define parallel transport along more general rectifiable paths by approximating them by piecewise geodesics and taking a limit.

Symmetries of $\mathbb{E}^2$

What are the "allowable" transformations in Euclidean geometry? That is, what are the transformations of $\mathbb{E}^2$ which preserve the geometrical properties which characterize it? These special transformations are called the *symmetries* (also called *automorphisms*) of $\mathbb{E}^2$; they form a *group*, which we will denote by Aut$(\mathbb{E}^2)$. A symmetry of $\mathbb{E}^2$ takes lines to lines, and preserves angles, but a symmetry of $\mathbb{E}^2$ does *not* have to preserve lengths. A symmetry can either preserve or reverse orientation. Basic symmetries include *translations, rotations, reflections, dilations*. It turns out that all symmetries of $\mathbb{E}^2$ can be expressed as simple combinations of these.

Exercise 3.1.4. *Let $f : \mathbb{E}^2 \to \mathbb{E}^2$ be orientation–reversing. Show that there is a unique line l such that f can be written as $g \circ r$ where r is a reflection in l and g is an orientation–preserving symmetry which fixes l, in which case g is either a translation parallel to l or a dilation whose center is on l. A reflection in l followed by a translation parallel to l is also called a* glide reflection.

Exercise 3.1.5. *Show that any homeomorphism $h : \mathbb{E}^2 \to \mathbb{E}^2$ which takes straight lines to straight lines must be* affine*; i.e. h is the composition of a linear map and a translation. In symbols, if v is a vector denoting a point in $\mathbb{E}^2$, then h is of the form*

$$h(v) = M(v) + w$$

for some invertible matrix M and some other vector w. The group of all affine transformations of $\mathbb{E}^2$ is denoted Aff$(\mathbb{E}^2)$, and algebraically we can write it in the form

$$\mathit{Aff}(\mathbb{E}^2) = GL(2, \mathbb{R}) \ltimes \mathbb{R}^2$$

Denote by Aut$^+(\mathbb{E}^2)$ the orientation–preserving symmetries, and by Isom$^+(\mathbb{E}^2)$ the orientation–preserving symmetries which are also distance–preserving.

Exercise 3.1.6. *Suppose $f : \mathbb{E}^2 \to \mathbb{E}^2$ is in Aut$^+(\mathbb{E}^2)$ but not in Isom$^+(\mathbb{E}^2)$. Then there is a unique point p fixed by f, and we can write f as $r \circ d$ where d is a dilation with center p and r is a rotation with center p.*

Exercise 3.1.7. *Suppose $f \in$ Isom$^+(\mathbb{E}^2)$. Then either f is a rotation or a translation, and it is a translation exactly when it does not have a fixed point. In either case, f can be written as $r_1 \circ r_2$ where r_i is a reflection in some line l_i. f is a translation exactly when l_1 and l_2 are parallel.*

These exercises show that any distance–preserving symmetry can be written as a product of at most 3 reflections. An interesting feature of these exercises is that they can be established *without using the parallel postulate*. So they describe true facts (where relevant) about elliptic and about hyperbolic geometry. So, for instance, a distance preserving symmetry of the hyperbolic plane can be written as a product $r_1 \circ r_2$ of reflections in lines l_1, l_2, and this transformation has a fixed point if and only if the lines l_1, l_2 intersect.

Exercise 3.1.8. *Verify that the group of* orientation–preserving *similarities of* $\mathbb{E}^2$ *which fix the origin is isomorphic to $\mathbb{C}^*$, the group of non–zero complex numbers with multiplication as the group operation. Verify too that the group of translations of $\mathbb{E}^2$ is isomorphic to $\mathbb{C}$ with addition as the group operation.*

Exercise 3.1.9. *Verify that the group Aut$^+(\mathbb{E}^2)$ of orientation–preserving similarities of $\mathbb{E}^2$ is isomorphic to $\mathbb{C}^* \ltimes \mathbb{C}$ where $\mathbb{C}^*$ acts on $\mathbb{C}$ by multiplication. In this way identify Aut$^+(\mathbb{E}^2)$ with the group of 2×2 complex matrices of the form*

$$\begin{bmatrix} \alpha & \beta \\ 0 & 1 \end{bmatrix}$$

and Isom$^+(\mathbb{E}^2)$ with the subgroup where $|\alpha| = 1$.

These exercises show that the group of (orientation–preserving) isometries of $\mathbb{E}^2$ is the semi–direct product of the circle group, and the subgroup of translations. There is a short exact sequence

$$0 \to \mathbb{R}^2 \to \mathrm{Isom}^+(\mathbb{E}^2) \to S^1 \to 0$$

The homomorphism $r : \mathrm{Isom}^+(\mathbb{E}^2) \to S^1$ may be interpreted as giving the *rotational part* of an isometry i. Here's how to calculate $r(i)$ for a given i. If l is any straight line in $\mathbb{E}^2$, then either l and $i(l)$ are parallel, or they intersect in a single point, at which the (signed) angle between l and $i(l)$ can be calculated as some number $a \in \mathbb{R}/2\pi = S^1$. If l and $i(l)$ are parallel, then $r(i)$ is either 0 or π depending on whether the orientation of l is preserved or reversed. Otherwise, $r(i) = a$.

To see that this makes sense, we need to check two things:

- The result should be independent of the choice of straight line l
- If i, j are two isometries, then

$$r(ij) = r(i) + r(j)$$

We leave the verification of these two routine facts to the reader.

3.1.2 The geometry of $\mathbb{E}^3$

Like $\mathbb{E}^2$, the space $\mathbb{E}^3$ is isotropic, and the stabilizer of any point p is isomorphic to the 3–dimensional orthogonal group

$$\text{stab}(p) \cong O(3, \mathbb{R})$$

Here $O(3, \mathbb{R})$ is the group of 3×3 matrices with real entries satisfying the relation $AA^t = \text{Id}$, where A^t denotes the transpose of the matrix A. Notice that this relation implies that elements of $O(3, \mathbb{R})$ are invertible, and have determinant equal to ± 1.

The subgroup $SO(3, \mathbb{R}) = \text{stab}(p) \cap \text{Isom}^+(\mathbb{E}^3)$ of index 2 consists of the matrices in $O(3, \mathbb{R})$ with determinant 1

As in dimension 2, an algebraic feature of the group of isometries of $\mathbb{E}^3$ which distinguishes Euclidean geometry from spherical or hyperbolic geometry is the existence of a homomorphism

$$\rho : Isom(\mathbb{E}^3) \to O(3, \mathbb{R})$$

where $\rho(\alpha)$ can be thought of as the *rotational part* of α. The homomorphism ρ is defined as follows: let (e_1, e_2, e_3) be an orthonormal frame at $T_p\mathbb{E}^3$ for some point $p \in \mathbb{E}^3$. Let t_α be the unique translation of $\mathbb{E}^3$ taking $\alpha(p)$ to p. Then $t_\alpha \alpha \in \text{stab}(p)$. One can check that if α, β are arbitrary, then $t_{\alpha\beta}\alpha\beta = t_\alpha \alpha t_\beta \beta$. The kernel of ρ is exactly the subgroup T of translations. So there is a short exact sequence

$$0 \to T \to \text{Isom}(\mathbb{E}^3) \to O(3, \mathbb{R}) \to 0$$

This short exact sequence is *split*. That is, there is a section $s : O(3, \mathbb{R}) \to \text{Isom}(\mathbb{E}^3)$ with the property that $s\rho : O(3, \mathbb{R}) \to O(3, \mathbb{R})$ is the identity.

Another way of saying this is that $\text{Isom}(\mathbb{E}^3)$ is a semi–direct product

$$\text{Isom}(\mathbb{E}^3) \cong T \rtimes O(3, \mathbb{R})$$

That is, the group consists of the pairs (t, r) with $t \in T, r \in O(3, \mathbb{R})$ and multiplication rule

$$(t_1, r_1)(t_2, r_2) = (t_1 r_1 t_2 r_1^{-1}, r_1 r_2)$$

Lemma 3.1.10. *Let $\alpha \in$ Isom$^+(\mathbb{E}^3)$ with $\rho(\alpha) \neq e$. Then either α fixes some point and is a rotation centered at that point, or it is the composition of a rotation r and a translation t such that the axis of r is parallel to the axis of t.*

Proof. We know from the structure of $\text{Isom}^+(\mathbb{E}^3)$ that $\alpha = tr$ for some translation t and rotation r. We show that we can choose t, r with the required properties.

If $t = \text{id}$ there is nothing to prove.

Otherwise, the axis l of r is well–defined. Let π be a plane perpendicular to l. Then r fixes π and acts as a rotation of it. Let $c \in \pi$ be the center of this rotation. Then we can write $t = t_1 t_2$ where t_1 is a translation in the direction of l and t_2 fixes π. So $\alpha = tr = t_1 t_2 r$. The composition $t_2 r$ fixes π. It has a nontrivial rotational part, so it is a rotation r' of π, fixing c' say. Since r' fixes π, its axis is perpendicular to π, so is parallel to l. Then the expression $\alpha = t_1 r'$ has the required properties. $\square$

A transformation α with $\rho(\alpha)$ nontrivial which is not itself a rotation is called a *screw motion*.

3.2 Spherical geometry

3.2.1 The 2–sphere

Elliptic geometry

Notation 3.2.1. The 2–sphere will be denoted by $\mathbb{S}^2$.

A very interesting "re–interpretation" of Euclid's first 4 axioms gives us elliptic geometry. A *point* in elliptic geometry consists of *two antipodal points* in $\mathbb{S}^2$. A *line* in elliptic geometry consists of a *great circle* in $\mathbb{S}^2$. The *antipodal* map $i : \mathbb{S}^2 \to \mathbb{S}^2$ is the map which takes any point to its antipodal point. A "line" or "point" with the interpretation above is invariant (as a set) under i, so we may think of the action as all taking place in the "quotient space" $\mathbb{S}^2/i$. An object in this quotient space is just an object in $\mathbb{S}^2$ which is invariant as a set by i. Any two great circles intersect in a pair of antipodal points, which is a single "point" in $\mathbb{S}^2/i$. If we think of $\mathbb{S}^2$ as a subset of $\mathbb{E}^3$, a great circle is the intersection of the sphere with a plane in $\mathbb{E}^3$ through the origin. A pair of antipodal points is the intersection of the sphere with a line in $\mathbb{E}^3$ through the origin. Thus, the geometry of $\mathbb{S}^2/i$ is equivalent to the geometry of planes and lines in $\mathbb{E}^3$. A plane in $\mathbb{E}^3$ through the origin is perpendicular to a unique line in $\mathbb{E}^3$ through the origin, and vice–versa. This defines a "duality" between lines and points in $\mathbb{S}^2/i$; so for any theorem one proves about lines and points in elliptic geometry, there is an analogous "dual" theorem with the idea of "line" and "point" interchanged. Let d denote the transformation which takes points to lines and vice versa.

Circles and angles make sense on a sphere, and one sees that the first 4 axioms of Euclid are satisfied in this model.

As distinct from Euclidean geometry where there are symmetries which change lengths, there is a natural length scale on the sphere. We set the diameter equal to 2π.

Spherical trigonometry

An example of this duality (and a justification of the choice of length scale) is given by the following

Lemma 3.2.2 (Spherical law of sines). *If T is a spherical triangle with side–lengths A, B, C and opposite angles α, β, γ, then*

$$\frac{\sin(A)}{\sin(\alpha)} = \frac{\sin(B)}{\sin(\beta)} = \frac{\sin(C)}{\sin(\gamma)}$$

Proof. Think of the sides of T as the intersection of $\mathbb{S}^2$ with planes π_i through the origin in $\mathbb{E}^3$, intersecting in lines l_i in $\mathbb{E}^3$. Then the lengths A, B, C are the angles between the l_i and the angles α, β, γ are the angles between the planes π_i.

Let v_1, v_2, v_3 be unit vectors with endpoints on the vertices of T. Then

$$\sin(A) = \|v_2 \times v_3\|, \ \sin(B) = \|v_1 \times v_3\|, \ \sin(C) = \|v_1 \times v_2\|$$

where $\times$ denotes cross product of vectors, and $\|\cdot\|$ denotes vector norm.

On the other hand, we know that $(v_1 \times v_3)/\|v_1 \times v_3|$ and $(v_2 \times v_3)/\|v_2 \times v_3|$ are unit vectors perpendicular to the planes π_2 and π_1, so that

$$\sin(\gamma) = \frac{\|(v_1 \times v_3) \times (v_2 \times v_3)\|}{\|v_1 \times v_3\| \, \|v_2 \times v_3\|}$$

Now, we have the identity

$$(v_1 \times v_3) \times (v_2 \times v_3) = v_3(v_1 \cdot (v_2 \times v_3))$$

and

$$(v_1 \cdot (v_2 \times v_3)) = \det(v_1, v_2, v_3)$$

where $\det$ denotes the determinant of the matrix with columns equal to the v_i. It follows that we have the formula

$$\frac{\sin(A)}{\sin(\alpha)} = \frac{\|v_1 \times v_2\| \, \|v_1 \times v_3\| \, \|v_2 \times v_3\|}{\det(v_1, v_2, v_3)}$$

which is obviously symmetric in the three v_i. □

Notice that the triangle $d(T)$ has side lengths $(\pi - \alpha), (\pi - \beta), (\pi - \gamma)$ and angles $(\pi - A), (\pi - B), (\pi - C)$. Notice too that $\sin(t) \approx t$ for small t, so that if T is a very small triangle, this formula approximates the sine rule for Euclidean space. Let $\mathbb{S}^2_t$ denote the sphere scaled to have diameter $2\pi t$; then the term $\frac{\sin(A)}{\sin(\alpha)}$ in the spherical sine rule should be replaced with $\frac{t\sin(t^{-1}A)}{\sin(\alpha)}$. In this way we may think of $\mathbb{E}^2$ as the "limit" as $t \to \infty$ of $\mathbb{S}^2$.

The area of a spherical triangle

Think of $\mathbb{S}^2$ as the unit sphere in $\mathbb{R}^3$, and let $\pi : \mathbb{S}^2 \to \mathbb{R}$ be the orthogonal projection onto the x–axis.

Theorem 3.2.3 (Archimedes). *The area of the preimage in $\mathbb{S}^2$ of a segment I in $[-1, 1]$ is proportional to its length. Explicitly,*

$$\textit{area } \pi^{-1}(I) = 2\pi \textit{ length}(I)$$

Proof. For $t \in [-1, 1]$ let C_t denote the spherical cap $C_t = \pi^{-1}([t, 1])$. Then if $t_1 - t_2 = s$ is very small, the difference $C_{t_1} - C_{t_2}$ is approximately an annulus with diameter $2\pi \sin(\alpha)$ and thickness $\frac{s}{\sin(\alpha)}$ where $\cos(\alpha) = t_1$. Thus this difference has area approximately equal to $2\pi s$ which is *independent* of the *values* t_1, t_2 and depends only on their *difference*. Passing to a limit as $s \to 0$ we see that the cap C_t has area $\int_t^1 2\pi dt = 2\pi(1 - t)$ and the result follows. □

In particular, the area of $\mathbb{S}^2$ is 4π. A *lune* is a spherical digon — that is, the area bounded between a pair of geodesic segments γ_1, γ_2 which join antipodal points $p, i(p)$. If $p, i(p)$ are thought of as the north and south pole respectively, the segments γ_i are longitudes. They are parameterized by angles $\alpha \in S^1$, and the difference in angles between two longitudes is the angle that the two geodesic segments make at p or at $i(p)$.

By rotational symmetry it follows that if L is a lune of $\mathbb{S}^2$ between the longitude 0 and the longitude α, then the area of L is $\frac{\alpha}{2\pi}\text{area}(\mathbb{S}^2) = 2\alpha$.

Now, let T be an arbitrary spherical triangle. If T is bounded by sides l_i which meet at vertices v_i then we can extend the sides l_i to great circles which cut up $\mathbb{S}^2$ into eight regions. Each pair of lines bound two lunes, and the six lunes so produced fall into two sets of three, which intersect exactly along the triangle T and the antipodal triangle $i(T)$. Every region in $\mathbb{S}^2$ is covered by at least one of the six lunes, and only the triangles T and $i(T)$ are covered by more than one lune — they are each covered by three lunes. See figure 3.1

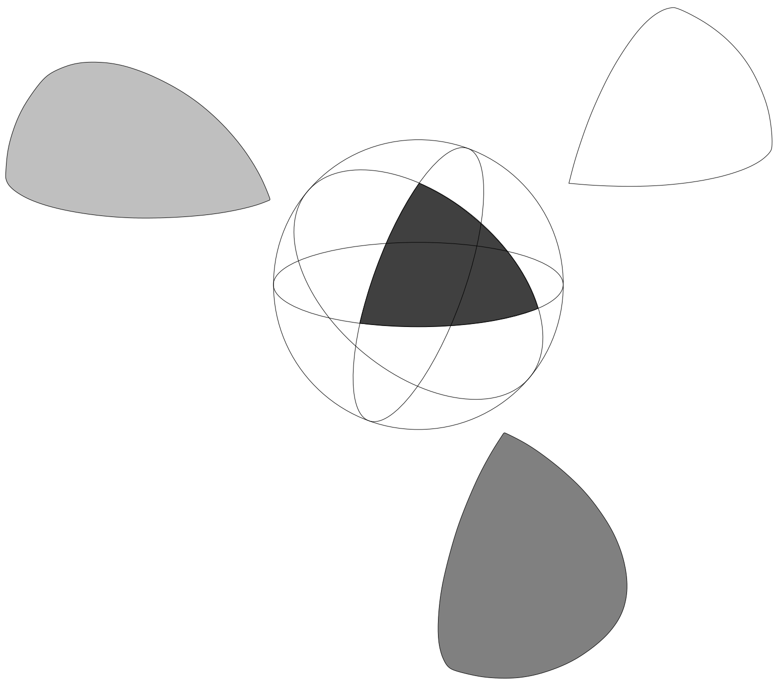

Figure 3.1: Three lunes cover half of $\mathbb{S}^2$ and overlap exactly on the triangle T.

It follows that we can calculate the area of T as follows

$$4\pi = \text{area}(\mathbb{S}^2) = \sum \text{area(lunes)} - 4\ \text{area}(T) = 4(\alpha + \beta + \gamma) - 4\ \text{area}(T)$$

In particular, we have the beautiful formula, which is a special case of the Gauss–Bonnet theorem:

Theorem 3.2.4. *Let T be a spherical triangle with angles α, β, γ. Then*

$$\textit{area}(T) = \alpha + \beta + \gamma - \pi$$

Notice that as T gets very small and the area $\to 0$, the sum of the angles of T approach π. Thus in the limit, we have Euclidean geometry in which the sum of the angles of a triangle are π. The angle formula for Euclidean triangles is equivalent to the parallel postulate.

Exercise 3.2.5. *Derive a formula for the area of a spherical polygon with n vertices in terms of the angles.*

Exercise 3.2.6. *Using the spherical law of sines and the area formula, calculate the area of a regular spherical n–gon with sides of length t.*

Kissing numbers — the Newton–Gregory problem

How many balls of radius 1 can be arranged in $\mathbb{E}^3$ so that they all touch a fixed ball of radius 1? It is understood that the balls are non–overlapping, but they may touch each other at a single point; figuratively, one says that the balls are "kissing" or "osculating" (from the Latin word for kiss), and that one wants to know the *kissing number* in 3–dimensions.

Exercise 3.2.7. *What is the kissing number in 2–dimensions? That is, how many disks of radius 1 can be arranged in $\mathbb{E}^2$ so that they all touch a fixed disk of radius 1?*

This question first arose in a conversation between Isaac Newton and David Gregory in 1694. Newton thought 12 balls was the maximum; Gregory thought 13 might be possible. It is quite easy to arrange 12 balls which all touch a fixed ball — arrange the centers at the vertices of a regular icosahedron. If the distance from the center of the icosahedron to the vertices is 2, it turns out the distance between adjacent vertices is ≈ 2.103, so this configuration can be physically realized (i.e. there is no overlapping). The problem is that there is some slack in this configuration — the balls roll around, and it is unclear whether by packing them more tightly there would be room for another ball.

Suppose we have a configuration of non–overlapping spheres S_i all touching the central sphere S. Let v_i be the points on S where they all touch. The non–overlapping condition is exactly equivalent to the condition that no two of the v_i are a distance of less than $\frac{\pi}{3}$ apart. If some of the S_i are loose, roll them around on the surface until they come into contact with other S_j; it's clear that we can roll "loose" S_i around until every S_i touches *at least* two other S_j, S_k. If S_i touches $S_{j_1}, \dots, S_{j_n}$ then join v_i to $v_{j_1}, \dots, v_{j_n}$ by segments of great circles on S. This gives a decomposition of S into spherical polygons, every edge of which has length $\frac{\pi}{3}$. It's easy to see that no polygon has 6 or more sides (why?).

Let f_n be the number of faces with n sides. Then there are $\frac{3}{2}f_3 + 2f_4 + \frac{5}{2}f_5$ edges, since every edge is contained in two faces. Recall Euler's formula for a polygonal decomposition of a sphere

$$\text{faces} \ - \ \text{edges} \ + \ \text{vertices} = 2$$

so the number of vertices is $2 + \frac{1}{2}f_3 + f_4 + \frac{3}{2}f_5$

Exercise 3.2.8. *Show that the largest spherical quadrilateral or pentagon with side lengths $\frac{\pi}{3}$ is the regular one. Use your formula for the area of such a polygon and the fact above to show that the kissing number is* 12 *in* 3*–dimensions. This was first shown in the 19th century.*

Exercise 3.2.9. *Show what we have implicitly assumed: namely that a connected nonempty graph in S^2 with embedded edges, and no vertices of valence* 1*, has polygonal complementary regions.*

Remark 3.2.10. In 1951 Schutte and van der Waerden ([8]) found an arrangement of 13 unit spheres which touches a central sphere of radius $r \approx 1.04556$ where r is a root of the polynomial

$$4096x^{16} - 18432x^{12} + 24576x^{10} - 13952x^8 + 4096x^6 - 608x^4 + 32x^2 + 1$$

This r is thought to be optimal.

Reflections, rotations, involutions; $SO(3)$

By thinking of $\mathbb{S}^2$ as the unit sphere in $\mathbb{E}^3$, and by thinking of points and lines in $\mathbb{S}^2$ as the intersection of the sphere with lines and planes in $\mathbb{E}^3$ we see that symmetries of $\mathbb{S}^2$ extend to linear maps of $\mathbb{E}^3$ to itself which fix the origin. These are expressed as 3×3 matrices. The condition that a matrix M induce a symmetry of $\mathbb{S}^2$ is exactly that it preserves distances on $\mathbb{S}^2$; equivalently, it preserves the angles between lines through the origin in $\mathbb{E}^3$. Consequently, it takes orthonormal frames to orthonormal frames. (A *frame* is another word for a *basis*.)

Any frame can be expressed as a 3×3 matrix F, where the columns give each of the vectors. F is orthonormal if $F^tF = \text{id}$. If M preserves orthonormality, then $F^tM^tMF = \text{id}$ for every orthonormal F; in particular, $M^tM = FF^t = \text{id}$. Observe that each of these transformations actually induces a symmetry of $\mathbb{S}^2$; in particular, we can identify the set of symmetries of $\mathbb{S}^2$ with the set of orthonormal frames in $\mathbb{E}^3$, which can be identified with the set of 3×3 matrices M satisfying $M^tM = \text{id}$. It is easy to see that such matrices form a *group*, known as the *orthogonal group* and denoted $O(3)$. The *subgroup* of orientation–preserving matrices (those with determinant 1) are denoted $SO(3)$ and called the *special orthogonal group*.

Exercise 3.2.11. *Show that every element of $O(3)$ has an eigenvector with eigenvalue* 1 *or* -1*. Deduce that a symmetry of $\mathbb{S}^2$ is either a rotation, a reflection, or a product $s \circ r$ where r is reflection in some great circle l and s is a rotation which fixes that circle. (How is this like a "glide reflection"?) In particular, every symmetry of $\mathbb{S}^2$ is a product of at most three reflections. Compare with the Euclidean case.*

Holonomy transport, $SO(3, \mathbb{R})$ is simple

A spherical triangle T with vertices a, b, c determines three rotations in $SO(3, \mathbb{R})$: the rotation γ_{ab} which fixes the great circle containing a and b and takes a to b, and similar rotations γ_{bc}, γ_{ca}. The net effect of the composition of these three rotations

$$\gamma_T = \gamma_{ca}\gamma_{bc}\gamma_{ab}$$

fixes the point a, and is therefore a rotation with axis the antipodal pair $a, -a$. What rotation is it?

If v is the vector at a tangent to ab, then γ_{ab} takes v to $\gamma_{ab}(v)$, a vector at b tangent to ab but in the opposite direction. $\gamma_{ab}(v)$ makes an angle of $\beta - \pi$ with the geodesic bc. Similarly, $\gamma_{bc}\gamma_{ab}(v)$ makes an angle of $\beta+\gamma-2\pi$ with ca. Finally, the composition $\gamma_T(v)$ makes an angle of $\alpha + \beta + \gamma - 3\pi$ with ab — that is to say, with v. But angles are only defined mod 2π, so the net result of γ_T is a rotation centered at $a, -a$ through the angle $\alpha + \beta + \gamma - \pi = \text{area}(T)$.

If we go around the triangle in the opposite direction, a similar argument shows that we get the rotation γ_T^{-1} centered at $a, -a$ through the same angle in the opposite direction.

More generally, if P is any spherical polygon, we can cut P up into triangles. Inductively, it is easy to see that the effect of composing a sequence of translations along the edges of P is a rotation γ_P, centered at the initial vertex and its antipode, through an angle equal in size to the area of P and with sign determined by the direction in which one goes around P. If Q is a more general embedded loop in $\mathbb{S}^2$, we can approximate Q by a sequence of polygons P_i and define $\gamma_Q = \lim_{i\to\infty} \gamma_{P_i}$ as the effect of "translating a vector" around Q, and see that γ_Q is precisely the rotation, centered at the initial point of the loop, through an angle equal to the area enclosed by Q. This (limiting) rotation γ_Q is called the *effect of holonomy transport around* Q, and $\gamma_Q(v)$ is called the *holonomy transport of* v *around* Q. Note that if I is merely a *path* and not a loop, we can define holonomy transport along I as a map which takes the tangent space at $I(0)$ to the tangent space at $I(1)$.

We can use this relationship between holonomy and area to give a proof that the group $SO(3,\mathbb{R})$ is simple — that is, that it contains no nontrivial proper normal subgroups.

Theorem 3.2.12. *$SO(3,\mathbb{R})$ is simple.*

Proof. Let G be a hypothetical nontrivial normal subgroup. If G is nontrivial, it contains some rotation r through an angle α. Since r is determined by the pair of antipodal points it fixes, and since $SO(3,\mathbb{R})$ acts transitively on pairs of antipodal points, it follows that G contains *every* rotation through angle α.

Note that rotation through angle α centered at points $p, -p$ is the same thing as translation along the great circle perpendicular to $p, -p$. Assume first that α is equal to π. Then if L is a lune with angle β, T_L is a rotation through 2β. Since β was arbitrary, $G = SO(3,\mathbb{R})$ as required. So assume $\alpha \neq \pi$. If Q is a spherical quadrilateral with side lengths all equal to α, the translations along the edges are all contained in G. So T_Q is contained in G, which is a rotation through area(Q) But a rhombic quadrilateral is flexible — we can pivot the angles to produce a 1–parameter family of quadrilaterals Q_t where Q_0 is degenerate, but Q_1 encloses a non–zero area. It follows that G contains all rotations through sufficiently small angles. But for any angle β, we can find a spherical polygon with small sides which encloses an area exactly equal to β, and therefore $G = SO(3,\mathbb{R})$ in this case too, and the theorem is proved. ∎

The simplicity of $\text{Isom}^+(\mathbb{S}^2)$, in contrast to the fact that $\text{Isom}^+(\mathbb{E}^2)$ is a semi–direct product, demonstrates a profound difference between spherical and Euclidean geometry.

Quaternions and the group $\mathbb{S}^3$

Recall that the *quaternions* are elements of the 4–dimensional real vector space spanned by $1, i, j, k$ with multiplication which is linear in each factor, and on the basis elements is given by

$$ij = k, jk = i, ki = j, i^2 = j^2 = k^2 = -1$$

This multiplication is associative. The *norm* of a quaternion, denoted by

$$\|a_1 + a_2 i + a_3 j + a_4 k\| = (a_1^2 + a_2^2 + a_3^2 + a_4^2)^{1/2}$$

is equal to the length of the corresponding vector (a_1, a_2, a_3, a_4) in $\mathbb{R}^4$. Norms are multiplicative. That is, $\|\alpha\beta\| = \|\alpha\|\|\beta\|$. The non–zero quaternions form a *group* under multiplication; the unit quaternions, which correspond exactly the to the unit length vectors in $\mathbb{R}^4$, are a subgroup which is denoted $\mathbb{S}^3$. Let π be the set of quaternions of the form $1 + ai + bj + ck$. Then π is a copy of $\mathbb{R}^3$, and is the *tangent space* to the sphere $\mathbb{S}^3$ of unit norm quaternions at 1. The group $\mathbb{S}$ acts on π by

$$\alpha \cdot z = \alpha^{-1} z \alpha$$

for $z \in \pi$. Since it preserves lengths, the image is isomorphic to a subgroup of $SO(3, \mathbb{R})$, which can be thought of as the group of orthogonal transformations of π. In fact, this homomorphism is *surjective*. Moreover, the kernel is exactly the center of $\mathbb{S}^3$, which is ± 1. That is, we have the isomorphism $\mathbb{S}^3 / \pm 1 \cong SO(3, \mathbb{R})$.

Exercise 3.2.13. *Write down the formula for an explicit homomorphism, in terms of standard quaternionic co–ordinates for $\mathbb{S}^3$ and matrix co–ordinates for $SO(3, \mathbb{R})$.*

In general, the conjugation action of a Lie group on its tangent space at the identity is called the *adjoint action* of the group. Since the group of linear transformations of this vector space is a matrix group, this gives a homomorphism of the Lie group to a matrix group.

3.2.2 The geometry of $\mathbb{S}^3$

As we have just seen in the last section, the 3–sphere is itself a group. Any compact group admits a Riemannian metric which is invariant under multiplication on both the left and the right. Such a metric will necessarily be homogeneous. Since the conjugation action of $\mathbb{S}^3$ on its tangent space at the identity factors through the usual action of $SO(3, \mathbb{R}) = \mathbb{S}^3 / \pm 1$ on $\mathbb{R}^3$, this homogeneous geometry is also isotropic. It follows that this homogeneous metric is unique up to scale, and is equal to the standard metric on $\mathbb{S}^3$ which it inherits by thinking of it as the unit sphere in $\mathbb{E}^4$.

In particular, this gives a natural inclusion $\mathbb{S}^3 \to \text{Isom}(\mathbb{S}^3)$ where the action is by left translation. Moreover, the conjugation action of $\mathbb{S}^3$ on itself generates the full group of (orientation preserving) isometries of the tangent space at the identity, and is therefore equal to the stabilizer of a point. These two actions of $\mathbb{S}^3$ together generate an action of $\mathbb{S}^3 \times \mathbb{S}^3$ on $\mathbb{S}^3$, where the multiplication of the first factor is on the left, and the multiplication of the second factor is on the right. Since left and right multiplication commute in a group, this defines a homomorphism

$$M : \mathbb{S}^3 \times \mathbb{S}^3 \to SO(4, \mathbb{R})$$

with kernel exactly equal to $(1, 1) \cup (-1, -1) \cong \mathbb{Z}/2\mathbb{Z}$ Notice that this description identifies $\mathbb{S}^3 \times \mathbb{S}^3$ with the universal cover of the group $SO(4, \mathbb{R})$, and shows that $SO(4, \mathbb{R})$, uniquely amongst orthogonal groups, is *semisimple* but not *simple*.

The Hopf fibration

We can think of $\mathbb{S}^3$ as the unit sphere in $\mathbb{R}^4$. But it is useful to think of $\mathbb{R}^4$ itself as $\mathbb{C}^2$, and $\mathbb{S}^3$ as the set of pairs of complex numbers (z, w) such that

$$z\bar{z} + w\bar{w} = 1$$

There is a natural diagonal action of $\mathbb{C}^*$ on $\mathbb{C}^2$ by multiplication on both factors. If $S^1 \subset \mathbb{C}^*$ denotes the subgroup of complex numbers of norm 1, then the action of S^1 on $\mathbb{C}^2$ preserves $\mathbb{S}^3$. Moreover, it is easy to see that this action has no fixed points, and therefore gives a decomposition of $\mathbb{S}^3$ into orbits of S^1, which is to say, into circles. The quotient of $\mathbb{S}^3$ by S^1 is the set of pairs of complex numbers (z, w) whose norms sum to 1, up to multiplication by a complex number of norm 1. Two such pairs (z_1, w_1) and (z_2, w_2) are in the same equivalence class exactly when their ratios are equal:

$$\frac{z_1}{w_1} = \frac{z_2}{w_2}$$

This gives a natural identification of $\mathbb{S}^3/S^1$ with complex projective space $\mathbb{CP}^1$ which can be thought of as the space of ratios $\mathbb{C}$ together with a point at infinity. So $\mathbb{CP}^1 = \mathbb{C} \cup \infty$ which is topologically the 2–sphere $\mathbb{S}^2$, and the projection

$$H : \mathbb{S}^3 \to \mathbb{S}^2$$

is known as the *Hopf fibration*. The *fibers* of the fibration are the circles in $\mathbb{S}^3$ which are the preimages of points in $\mathbb{S}^2$ under H.

It is easy to see that the circles of the Hopf fibration are all precisely equal to the intersections $\mathbb{S}^3 \cap \pi$ where π is a complex line in $\mathbb{C}^2$ through the origin. In particular, the circles of the Hopf fibration are all great circles, and are geodesic with respect to the Riemannian structure on $\mathbb{S}^3$. If γ is a fiber of the Hopf fibration, every nearby fiber γ' winds once positively around γ, like the grooves in a corkscrew.

Thinking of $\mathbb{S}^3$ as the unit quaternions, the subgroup S^1 as above consists of the unit quaternions with j, k part equal to 0. Any 1–parameter subgroup of $\mathbb{S}^3$ is conjugate to this S^1. In particular, the orbits of points under the action of any such 1–parameter subgroup of $\mathbb{S}^3$ acting on the left give a foliation of $\mathbb{S}^3$ by (positively winding) round circles. Similarly, a 1–parameter subgroup of $\mathbb{S}^3$ acting on $\mathbb{S}^3$ *on the right* gives a foliation of $\mathbb{S}^3$ by circles which wind about each other *negatively*.

The group $SU(2)$

The group $U(2)$ is defined as the set of 2×2 matrices A satisfying the identity

$$AA^* = \mathrm{Id}$$

where A^* is the *conjugate transpose* of A — that is, the matrix obtained from A^t by complex conjugation on each of its entries. It is clear from the definition that the determinant of any $A \in U(2)$ is a complex number of absolute value 1. The subgroup $SU(2) \subset U(2)$ consists of the matrices with determinant 1. Let ω denote the standard Hermitian form

$$\omega((z_1, z_2), (w_1, w_2)) = z_1\overline{w}_1 + z_2\overline{w}_2$$

Then $U(2)$ is the group of matrices which preserve ω. To see this, write (z_1, z_2) and (w_1, w_2) as column vectors, and denote them by z, w. Then observe that $\omega(z, w) = w^*z$. So

$$\omega(Az, Aw) = w^*A^*Az = w^*z = \omega(z, w)$$

In particular, $U(2)$ preserves $\mathbb{S}^3$. The action of $\mathbb{S}^3$ on itself gives precisely the subgroup $SU(2)$, where the identification is given by

$$(z_1, z_2) \to \begin{pmatrix} z_1 & -\bar{z}_2 \\ z_2 & \bar{z}_1 \end{pmatrix}$$

1–parameter subgroups of $SO(4, \mathbb{R})$

A 1–parameter subgroup θ_t of $SO(4, \mathbb{R})$ is generated by the simultaneous action of two 1–parameter subgroups θ_t^L, θ_t^R of $\mathbb{S}^3$ which act on the left and the right respectively. The subgroup θ_t^L keeps invariant a left–handed Hopf fibration H_L, and the subgroup θ_t^R keeps invariant a right–handed Hopf fibration H_R. For each circle γ of H_L, the union of the circles of H_R which intersect γ is either γ itself, or a torus T_γ. Such tori and circles fill up $\mathbb{S}^3$ to give a (singular) foliation. If $H : \mathbb{S}^3 \to \mathbb{S}^2$ denotes the Hopf fibration H_L, the tori T_γ project to round circles in $\mathbb{S}^2$ and give a singular foliation which fills up all of the sphere except for two points. That is, there are exactly two circles γ_1, γ_2 in H_R which are also in H_L, and the rest of $\mathbb{S}^3$ is filled up with tori T_γ. Each torus admits a transitive family of isometries, and is therefore intrinsically Euclidean.

Since the actions of θ_t^L and θ_t^R commute, the action of either subgroup preserves the foliation of $\mathbb{S}^3$ into tori and circles, and therefore the action has the following form: it acts on the circles γ_1, γ_2 as rotations at some speed, and acts on each torus T_γ by some (possibly irrational) rotation. Notice though 1–parameter subgroups of $\mathbb{S}^3$ (and therefore of $SO(3, \mathbb{R})$) are always compact circles, 1–parameter subgroups of $SO(4, \mathbb{R})$ may be noncompact.

3.3 Hyperbolic geometry

3.3.1 Inversions

Recall that a *conformal structure* is an equivalence class of Riemannian metrics, where two metrics are equivalent if they are proportional on every T_pM. Isometries of $\mathbb{E}^n$ are conformal, and so are dilations. Another important class of conformal transformations, defined now only on complements of points in $\mathbb{E}^n$, are *inversions*.

Definition 3.3.1. If S is a sphere of radius r centered at $p \in \mathbb{E}^n$ the inversion in S, denoted i_S, is a conformal transformation

$$i_S : \mathbb{E}^n - p \to \mathbb{E}^n - p$$

defined as follows. For $x \in \mathbb{E}^n - p$ there is a unique ray r_x from p which passes through x. Then $i_S(x)$ is the unique point on r_x with the property that $d(x,p)d(i_S(x),p) = r^2$.

It follows from the definition that i_S fixes S pointwise and interchanges the interior and exterior. It is easy to see that i_S takes spheres and planes to spheres and planes. In particular, if π is a plane, $i_S(\pi)$ is a sphere passing through p, and *vice versa*. Notice that S is not the only sphere taken to itself by i_S: if T is a sphere perpendicular to S, then the sphere $i_S(T)$ is also perpendicular to S (because i_S is conformal) and intersects S in the same circle (because i_S fixes every point in S). So $i_S(T) = T$.

In two dimensions, conformal structures are very flexible. If U, V are any two open domains in $\mathbb{C}$ bounded by topological circles, there is a conformal map from U to V. However in higher dimensions, conformal structures are much more rigid, even locally.

3.3.2 The hyperbolic plane

The problem of models

Notation 3.3.2. The hyperbolic plane will be denoted by $\mathbb{H}^2$.

The sphere is relatively easy to understand and visualize because there is a very nice model of it in Euclidean space: the unit sphere in $\mathbb{E}^3$. Symmetries of the sphere extend to symmetries of the ambient space, and distances and angles in the sphere are what one expects from the ambient embedding. No such model exists of the hyperbolic plane. Bits of the hyperbolic plane can be isometrically (i.e. in a distance–preserving way) embedded, but not in such a way that the natural symmetries of the plane can be realized as symmetries of the embedding. However, if we are willing to look at embeddings which distort distances, there are some very nice models of the hyperbolic plane which one can play with and get a good feel for.

The Poincaré Model:

Suppose we imagine the world as being circumscribed by the unit circle in the plane. In order to prevent people from falling off the edge, we make the edges very cold. As everyone knows, objects shrink when they get cold, so people wandering around on the disk would get smaller and smaller as they approached the edge, so that its apparent distance (to them) would get larger and larger and they could never reach it. Technically, the "length elements" at the point (x, y) are

$$\left(\frac{2dx}{(1-x^2-y^2)}, \frac{2dy}{(1-x^2-y^2)}\right)$$

or in polar co–ordinates, the "length elements" at the point r, θ are

$$\left(\frac{2dr}{(1-r^2)}, \frac{2rd\theta}{(1-r^2)}\right)$$

This is called the *Poincaré metric* on the unit disk, and the disk with this metric is called the *Poincaré model of the hyperbolic plane*. With this choice of metric, the length of a radial line from the origin to the point $(r_1, 0)$ is

$$\int_0^{r_1} \frac{2dr}{(1-r^2)} = \log\left(\frac{1+r}{1-r}\right)\Bigg|_0^{r_1} = \log\left(\frac{1+r_1}{1-r_1}\right)$$

If γ is any other path from the origin to $(r_1, 0)$ whose Euclidean length is l, then its length in the hyperbolic metric is

$$\text{hyperbolic length of } \gamma = \int_0^l \frac{2dt}{(1-r^2(\gamma(t)))}$$

where $r(\gamma(t))$ is the distance from the point $\gamma(t)$ to the origin. Obviously, $l \geq r_1$ and $r(\gamma(t)) \leq t$ with equality if and only if γ is a Euclidean straight line. This implies that the shortest curve *in the hyperbolic metric* from the origin to a point in the disk is the Euclidean straight line.

Notice that for a point p at Euclidean distance ϵ from the boundary circle, the ratio of the hyperbolic to the Euclidean metric is

$$\frac{2}{1-(1-\epsilon)^2} \approx \frac{1}{\epsilon}$$

for sufficiently small ϵ.

Exercise 3.3.3. *(Hard). Let E be a simply–connected (i.e. without holes) domain in $\mathbb{R}^2$ bounded by a smooth curve γ. Define a "metric" on E as follows. Let f be a smooth, nowhere zero function on E which is equal to $\frac{1}{dist(p,\gamma)}$ for all p sufficiently close to γ. Let the length elements on E be given by (dxf, dyf) where (dx, dy) are the usual Euclidean length elements. Show that there is a continuous, 1–1 map $\phi : E \to D$ which distorts the lengths of curves by a bounded amount. That is, there is a constant $K > 0$ such that for any curve α in E,*

$$\frac{1}{K} length_D(\phi(\alpha)) \leq length_E(\alpha) \leq K length_D(\phi(\alpha))$$

Definition 3.3.4. The circle ∂D is called the *circle at infinity* of D, and is denoted S^1_∞. A point in S^1_∞ is called an *ideal point*.

Now think of the unit disk D as the set of complex numbers of norm ≤ 1. Let α, β be two complex numbers with $|\alpha|^2 - |\beta^2| = 1$. The set of matrices of the form

$$M = \begin{bmatrix} \alpha & \bar{\beta} \\ \beta & \bar{\alpha} \end{bmatrix}$$

form a group, called the *special unitary group* $SU(1,1)$. This is exactly the group of complex linear transformations of $\mathbb{C}^2$ which preserves the function $v(z,w) = |z|^2 - |w|^2$ and have determinant 1. These are the matrices M of determinant 1 satisfying $\overline{M}^t JM = J$ where $J = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$.

Exercise 3.3.5. *Define $U(1,1)$ to be the group of 2×2 complex matrices M with $\overline{M}^t JM = J$ with J as above, and no condition on the determinant. Find the most general form of a matrix in $U(1,1)$. Show that these are exactly the matrices whose column vectors are an orthonormal basis for the "norm" defined by v.*

Now, there is a natural action of $SU(1,1)$ on D by

$$\begin{bmatrix} \alpha & \bar{\beta} \\ \beta & \bar{\alpha} \end{bmatrix} \cdot z \to \frac{\alpha z + \bar{\beta}}{\beta z + \bar{\alpha}}$$

Observe that two matrices which differ by ± 1 act on D in the same way. So the action descends to the quotient group $SU(1,1)/\pm 1$ which is denoted by $PSU(1,1)$ for the *projective special unitary group*.

Observe that this transformation preserves the boundary circle. Furthermore it takes lines and circles to lines and circles, and preserves angles of intersection between them; in particular, it permutes segments of lines and circles perpendicular to ∂D.

Exercise 3.3.6. *Show that there is a transformation in $PSU(1,1)$ taking any point in the interior of D to any other point.*

Exercise 3.3.7. *Show that the subgroup of $PSU(1,1)$ fixing any point is isomorphic to a circle. Deduce that we can identify D with the coset space $PSU(1,1)/S^1$; i.e. D is a* homogeneous space for $PSU(1,1)$.

Exercise 3.3.8. *Show that the action of $PSU(1,1)$ preserves the Poincaré metric in D^2. Deduce that the shortest hyperbolic path between any two points is through an arc of a circle orthogonal to ∂D or, if the points are on the same diameter, by a segment of this diameter.*

The upper half–space model

The *upper half–space*, denoted $\mathbb{H}$, is the set of points $x, y \in \mathbb{R}^2$ with $y > 0$. Suppose now that the real line is chilled, so that distances in this model are scaled in proportion to the distance to the boundary. That is, in (x, y) co–ordinates the "length elements" of the metric are

$$\left(\frac{dx}{y}, \frac{dy}{y}\right)$$

Observe that translations parallel to the x–axis preserve the metric, and are therefore isometries. Also, dilations centered at points on the x–axis preserve the metric too. The length of a vertical line segment from (x, y_1) to (x, y_2) is

$$\int_{y_1}^{y_2} \frac{dy}{y} = \log \frac{y_2}{y_1}$$

A similar argument to before shows that this is the shortest path between these two points.

The group of 2×2 matrices with real coefficients and determinant 1 is called the *special linear group*, and denoted $SL(2, \mathbb{R})$ or $SL(2)$ if the coefficients are understood. These matrices act on the upper half–plane, thought of as a domain in $\mathbb{C}$, by

$$\begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix} \cdot z \to \frac{\alpha z + \beta}{\gamma z + \delta}$$

Again, two matrices which differ by a constant multiple act on $\mathbb{H}$ in the same way, so the action descends to the quotient group $SL(2, \mathbb{R})/ \pm 1$ which is denoted by $PSL(2, \mathbb{R})$ for the *projective special linear group*.

As before, these transformations take lines and circles to lines and circles, and preserve the real line.

Exercise 3.3.9. *Show that the action of $PSL(2, \mathbb{R})$ preserves the metric on $\mathbb{H}$. Deduce that the shortest hyperbolic path between any two points in the upper half–space is through an arc of a circle orthogonal to $\mathbb{R}$ or, if the points are on the same vertical line, through a segment of this line.*

Exercise 3.3.10. *Find a transformation from D to $\mathbb{H}$ which takes the Poincaré metric on the disk to the hyperbolic metric on $\mathbb{H}$. Deduce that these models describe "the same" geometry. Find an explicit isomorphism $PSL(2, \mathbb{R}) \cong PSU(1, 1)$.*

The hyperboloid model

In $\mathbb{R}^3$ let H denote the sheet of the hyperboloid $x^2 + y^2 - z^2 = -1$ with z positive. Let $O(2, 1)$ denote the set of 3×3 matrices with real entries which preserve the function $v(x, y, z) = x^2 + y^2 - z^2$, and $SO(2, 1)$ the subgroup with determinant 1. Equivalently, $O(2, 1)$ is the group of real matrices M such that $M^t J M = J$ where

$$J = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix}$$

Then $SO(2, 1)$ preserves the sheet H.

Definition 3.3.11. A vector v in $\mathbb{R}^3$ is *timelike* if $v^tJv < 0$, *spacelike* if $v^tJv > 0$ and *lightlike* if $v^tJv = 0$. The *Lorentz length* of a vector is $(v^tJv)^{1/2}$, denoted $\|v\|$, and can be positive, zero, or imaginary. The *timelike angle* between two timelike vectors v, w is

$$\eta(v,w) = \cosh^{-1}\left(\frac{v^tJw}{\|v\|\,\|w\|}\right)$$

Compare this with the usual angle between two vectors in $\mathbb{R}^3$:

$$\nu(v,w) = \cos^{-1}\left(\frac{v^tw}{\|v\|\,\|w\|}\right)$$

where in this equation $\|\cdot\|$ denotes the usual length of a vector. Notice that H is exactly the set of vectors of Lorentz length i, just as $\mathbb{S}^2$ is the set of vectors of usual length 1 (this has led some people to comment that the hyperbolic plane should be thought of as a "sphere of imaginary radius"). Since distances between points in $\mathbb{S}^2$ are defined as the angle between the vectors, it makes sense to define distances in H as the timelike angle between vectors. For two vectors $v, w \in H$ the formula above simplifies to

$$\eta(v,w) = \cosh^{-1}(-v^tJw)$$

Exercise 3.3.12. *Let K be the group of matrices of the form*

$$\begin{bmatrix} \cos(\alpha) & \sin(\alpha) & 0 \\ -\sin(\alpha) & \cos(\alpha) & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

and A the group of matrices of the form

$$\begin{bmatrix} \cosh(\gamma) & 0 & \sinh(\gamma) \\ 0 & 1 & 0 \\ \sinh(\gamma) & 0 & \cosh(\gamma) \end{bmatrix}$$

Show that every element of $SO(2,1)$ can be expressed as k_1ak_2 for some $k_1, k_2 \in K$ and $a \in A$; that is, we can write $SO(2,1) = KAK$. How unique is such an expression?

Notice the group K above is precisely the stabilizer of the point $(0,0,1) \in H$. Thus we can identify H with the homogeneous space $SO(2,1)/K$.

Exercise 3.3.13. *Let K' be the subgroup of $PSL(2,\mathbb{R})$ consisting of matrices of the form $\begin{bmatrix} \cos(\alpha) & \sin(\alpha) \\ -\sin(\alpha) & \cos(\alpha) \end{bmatrix}$ and A' the subgroup of matrices of the form $\begin{bmatrix} s & 0 \\ 0 & s^{-1} \end{bmatrix}$. Find an isomorphism from $SO(2,1)$ to $PSL(2,\mathbb{R})$ taking K to K' and A to A'. (Careful! The isomorphism $K \to K'$ might not be the one you first think of...)*

Remark 3.3.14. The isomorphisms $PSU(1,1) \cong PSL(2,\mathbb{R}) \cong SO(2,1)$ are known as *exceptional isomorphisms*, and one should not assume that the matter is so simple in higher dimensions. Such exceptional isomorphisms are rare and are a very powerful tool, since difficult problems about one of the groups can become simpler when translated into a problem about another of the groups.

After identifying $PSL(2,\mathbb{R})$ with $SO(2,1)$ we can identify their homogeneous spaces $PSL(2,\mathbb{R})/K'$ and $SO(2,1)/K$. This identification of H with $\mathbb{H}$ shows that H is an equivalent model of the hyperbolic plane. The straight lines in H are the intersection of planes in $\mathbb{R}^3$ through the origin with H. In many ways, the hyperboloid model of the hyperbolic plane is the closest to the model of $\mathbb{S}^2$ as the unit sphere in $\mathbb{R}^3$.

Exercise 3.3.15. *Show that the identification of H with $\mathbb{H}$ preserves metrics.*

Two vectors $v, w \in \mathbb{R}^3$ are *Lorentz orthogonal* if $v^t J w = 0$. It is easy to see that if v is timelike, any orthogonal vector w is spacelike. If $v \in H$ and w is a tangent vector to H at v, then $v^t J w = 0$, since the derivative $\frac{d}{dt}\|v + tw\|$ should be equal to 0 at $t = 0$ (by the definition of a tangent vector). For two spacelike vectors v, w which span a spacelike vector space, the value of $\frac{v^t J w}{\|v\|\,\|w\|} < 1$, so $\eta(v,w)$ is an imaginary number. The *spacelike angle* between v and w is defined to be $-i\eta(v,w)$. It is a fact that the hyperbolic angle between two tangent vectors at a point in H is exactly equal to their spacelike angle. The "proof" of this fact is just that the symmetries of the space H preserve this spacelike angle, and the total spacelike angle of a circle is 2π. Since these two properties uniquely characterize hyperbolic angles, the two notions of angle agree.

The fact that hyperbolic lengths and angles can be expressed so easily in terms of trigonometric functions and linear algebra makes the hyperboloid model the model of choice for doing hyperbolic trigonometry.

The Klein (projective) model

Let D_1 be the disk consisting of points in $\mathbb{R}^3$ with $x^2 + y^2 \leq 1$ and $z = 1$. Then we can project H to D_1 along rays in $\mathbb{R}^3$ which pass through the origin.

Exercise 3.3.16. *Verify that this stereographic projection is* 1–1 *and onto.*

This projection takes the straight lines in H to (Euclidean) straight lines in D_1. This gives us a new model of the hyperbolic plane as the unit disk, whose points are usual points, and whose lines are exactly the segments of Euclidean lines which intersect D_1. One should be wary that Euclidean angles in this model do not accurately depict the true hyperbolic angles. In this model, two lines l_1, l_2 are perpendicular under the following circumstances:

- If l_1 passes through the origin, l_2 is perpendicular to l_1 if any only if it is perpendicular in the Euclidean sense.
- Otherwise, let m_1, m_2 be the two tangent lines to ∂D_1 which pass through the endpoints of l_1. Then l_1 and l_2 are perpendicular if and only if l_2, m_1, m_2 intersect in a point.

Exercise 3.3.17. *Verify that l_1 is perpendicular to l_2 if and only if l_2 is perpendicular to l_1.*

It is easy to verify in this model that all Euclid's axioms but the fifth are satisfied.

The relationship between the Klein model and the Poincaré model is as follows: we can map the Poincaré disk to the northern hemisphere of the unit sphere by stereographic projection. This preserves angles and takes lines and circles to lines and circles. In this model, the straight lines are exactly the arcs of circles perpendicular to the equator. Then the Klein model and this (curvy) Poincaré model are related by placing thinking of the Klein disk as the flat Euclidean disk spanning the equator, and mapping D_1 to the upper hemisphere by projecting points along lines parallel to the z–axis. This takes lines in D_1 to the intersection of the upper hemisphere with vertical planes. These intersections are the circular arcs which are perpendicular to the equator, so this map takes straight lines to straight lines as it should.

Exercise 3.3.18. *Write down the metric in D_1 for the Klein model. Using this formula, calculate the hyperbolic distance between the center and a point at radius r_1 in D_1.*

Exercise 3.3.19. *Show that* Pappus' theorem *is true in the hyperbolic plane; this theorem says that if a_1, a_2, a_3 and b_1, b_2, b_3 are points in two lines l_1 and l_2, then the six line segments joining the a_i to the b_j for $i \neq j$ intersect in three points which are collinear.*

Hyperbolic trigonometry

Hyperbolic and spherical geometry are two sides of the same coin. For many theorems in spherical geometry, there is an analogous theorem in hyperbolic geometry. For instance, we have the

Lemma 3.3.20 (Hyperbolic law of sines). *If T is a hyperbolic triangle with sides of length A, B, C opposite angles α, β, γ then*

$$\frac{\sinh(A)}{\sin(\alpha)} = \frac{\sinh(B)}{\sin(\beta)} = \frac{\sinh(C)}{\sin(\gamma)}$$

Exercise 3.3.21. *Prove the hyperbolic law of sines by using the hyperboloid model and trying to imitate the vector proof of the spherical law of sines.*

The area of a hyperbolic triangle

The parallels between spherical and hyperbolic geometry are carried further by the theorem for the area of a hyperbolic triangle. We relax slightly the notion of a triangle: we allow some or all of the vertices of our triangle to be ideal points. If all three vertices are ideal, we say that we have an *ideal triangle*. Notice that since every hyperbolic straight line is perpendicular to S^1_∞, the angle of a triangle at an ideal point is 0.

Theorem 3.3.22. *Let T be a hyperbolic triangle with angles α, β, γ. Then*

$$\textit{area}(T) = \pi - \alpha - \beta - \gamma$$

Proof. In the upper half–space model, let T be the triangle with one ideal point at ∞ and two ordinary points at $(\cos(\alpha), \sin(\alpha))$ and $(\cos(\pi - \beta), \sin(\pi - \beta))$ in Euclidean co–ordinates, where α, β are both $\leq \pi/2$. Such a triangle has angles $0, \alpha, \beta$. The hyperbolic area is

$$\int_{x=\cos(\pi-\beta)}^{\cos(\alpha)} \left(\int_{y=1-x^2}^{\infty} \frac{1}{y^2} dy \right) dx$$

$$= \int_{x=\cos(\pi-\beta)}^{\cos(\alpha)} \frac{1}{1-x^2} dx$$

$$= -cos^{-1}(x) \Big|_{\cos(\pi-\beta)}^{\cos(\alpha)} = \pi - \alpha - \beta$$

In particular, a triangle with one or two ideal points satisfies the formula. Now, for an arbitrary triangle T with angles α, β, γ we can dissect an ideal triangles with all angles 0 into T and three triangles, each of which has two ideal points, and whose third angle is one of $\pi - \alpha, \pi - \beta, \pi - \gamma$. That is,

$$\text{area}(T) = \pi - (\pi - (\pi - \alpha)) - (\pi - (\pi - \beta)) - (\pi - (\pi - \gamma)) = \pi - \alpha - \beta - \gamma$$

$\square$

Projective geometry

The group $PSL(2, \mathbb{R})$ acts in a natural way on another space called the *projective line*, denoted $\mathbb{RP}^1$. This is the space whose points are the lines through the origin in $\mathbb{R}^2$. Equivalently, this is is the quotient of the space $\mathbb{R}^2 - 0$ by the equivalence relation that $(x, y) \sim (\lambda x, \lambda y)$ for any $\lambda \in \mathbb{R}^*$. The unit circle maps 2–1 to $\mathbb{RP}^1$, so one sees that $\mathbb{RP}^1$ is itself a circle. The natural action of $PSL(2, \mathbb{R})$ on $\mathbb{RP}^1$ is the *projectivization* of the natural action of $SL(2, \mathbb{R})$ on $\mathbb{R}^2$. That is,

$$\begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \alpha x + \beta y \\ \gamma x + \delta y \end{bmatrix}$$

We can write an equivalence class (x, y) unambiguously as x/y, where we write ∞ when $y = 0$; in this way, we can naturally identify $\mathbb{RP}^1$ with $\mathbb{R} \cup \infty$. One sees that in this formulation, this is exactly the action of $PSL(2, \mathbb{R})$ on the ideal boundary of $\mathbb{H}^2$ in the upper half–space model. That is, *the geometry of* $\mathbb{RP}^1$ *is hyperbolic geometry at infinity*. Observe that for any two triples of points a_1, a_2, a_3 and b_1, b_2, b_3 in $\mathbb{RP}^1$ which are circularly ordered, there is a *unique* element of $PSL(2, \mathbb{R})$ taking a_i to b_i. For a *four–tuple* of points a_1, a_2, a_3, a_4 let γ be the transformation taking a_1, a_2, a_3 to $0, 1, \infty$. Then $\gamma(a_4)$ is an invariant of the 4–tuple, called the *cross–ratio* of the four points, denoted $[a_1, a_2, a_3, a_4]$. Explicitly,

$$[a_1, a_2, a_3, a_4] = \frac{(a_1 - a_3)(a_2 - a_4)}{(a_1 - a_2)(a_3 - a_4)}$$

The subgroup of $PSL(2, \mathbb{R})$ which fixes a point in $\mathbb{R} \cup \infty$ is isomorphic to the group of *orientation preserving similarities of* $\mathbb{R}$, which we could denote by $\text{Aut}^+(\mathbb{R})$. This group is isomorphic to $\mathbb{R}^+ \ltimes \mathbb{R}$. where $\mathbb{R}^+$ acts on $\mathbb{R}$ by multiplication. We can think of $\mathbb{RP}^1$ as the homogeneous space $PSL(2, \mathbb{R})/\mathbb{R}^+ \ltimes \mathbb{R}$.

Projective geometry is the geometry of *perspective*. Imagine that we have a transparent glass pane, and we are trying to capture a landscape by setting up the pane and painting the scenery on the pane as it appears to us. We could move the pane to the right or left; this would translate the scene left or right respectively. We could move the pane closer or further away; this would shrink or magnify the image. Or we could rotate the pane and ourselves so that the sun doesn't get in our eyes. The horizon in our picture is $\mathbb{RP}^1$, and the transformations we can perform on the image is precisely the projective group $PSL(2, \mathbb{R})$.

Elliptic, parabolic, hyperbolic isometries

There are three different kinds of transformations in $PSL(2, \mathbb{R})$ which can be distinguished by their action on S^1_∞.

Definition 3.3.23. A non–trivial element $\alpha \in PSL(2, \mathbb{R})$ is *elliptic, parabolic* or *hyperbolic* if it has respectively $0, 1$ or 2 fixed points in S^1_∞. These cases can be distinguished by the property that $|tr(\gamma)|$ is $<, =$ or $>$ 2, where tr denotes the trace of a matrix representative of γ.

An elliptic transformation has a unique fixed point in $\mathbb{H}^2$ and acts as a rotation about that point. A hyperbolic transformation fixes the geodesic running between its two ideal fixed points and acts as a translation along this geodesic. Furthermore, the points in $\mathbb{H}^2$ moved the shortest distance by the transformation are exactly the points on this geodesic.

A parabolic transformation has no analogue in Euclidean or Spherical geometry. It has no fixed point, but moves points an arbitrarily short amount. In some sense, it is like a "rotation about an ideal point". Two elliptic elements are conjugate iff they rotate about their respective fixed points by the same amount. This angle of rotation is equal to $\cos^{-1}(|tr(\gamma)/2|)$. Two hyperbolic elements are conjugate iff they translate along their geodesic by the same amount. This translation length is equal to $\cosh^{-1}(|tr(\gamma)/2|)$. If a, b are any two parabolic elements then either a and b are conjugate or a^{-1} and b are conjugate.

Exercise 3.3.24. *Prove the claims made in the previous paragraph.*

Exercise 3.3.25. *Recall the subgroups K' and A' defined in the prequel. Let N denote the group of matrices of the form $\begin{bmatrix} 1 & t \\ 0 & 1 \end{bmatrix}$. Show that every element of $PSL(2, \mathbb{R})$ can be expressed as kan for some $k \in K'$, $a \in A'$ and $n \in N$. How unique is this expression? This is an example of what is known as the KAN or* Iwasawa decomposition.

Exercise 3.3.26. *Consider the group $SL(n, \mathbb{R})$ of $n \times n$ matrices with real entries and determinant 1. Let K be the subgroup $SO(n, \mathbb{R})$ of real $n \times n$ matrices M satisfying $M^t M = id$. Let A be the subgroup of diagonal matrices. Let N be the subgroup of matrices with 1's on the diagonal and 0's below the diagonal. Show A is abelian and N is nilpotent. Further, K is* compact *(see appendix), thought of as a topological subspace of the space $\mathbb{R}^{n^2}$ of $n \times n$ matrices. Show that there is a KAN decomposition for $SL(n, \mathbb{R})$.*

Horocircular geometry

In the Poincaré disk model, the Euclidean circles in D which are tangent to S^1_∞ are special; they are called *horocircles* and one can think of them as *circles of infinite radius*. If the point of tangency is taken to be ∞ in the upper half–space model, these circles correspond to the horizontal lines in the upper half–plane. Observe that a parabolic element fixes the family of horocircles tangent to its fixed ideal point, and acts on each of them by translation.

3.3.3 The geometry of $\mathbb{H}^3$

One can define 3–dimensional hyperbolic geometry implicitly as $(\mathbb{D}^3, G)$ where G is the group of conformal transformations of the unit ball $\mathbb{D}^3$ to itself. By Liouville's theorem, one can show that G consists of transformations of the form Ai_S where $A \in O(3,\mathbb{R})$ and i_S is either the identity or inversion in a sphere perpendicular to the unit sphere.

First, observe that G is transitive. We investigate the point stabilizers.

Lemma 3.3.27. *Let $p \in \mathbb{D}^3$ be arbitrary. Then stab$(p) \subset G$ is isomorphic to $O(3,\mathbb{R})$.*

Proof. Since G is transitive, we can assume without loss of generality that $p = 0$. Now, if $\phi \in \text{stab}(p)$ we write $\phi(x) = \lambda Ai(x) + t$. Since ϕ takes the center of S to the center of S, it follows that i is the identity. For, if S, T are arbitrary spheres with distinct centers, the inversion in T takes S to a new sphere $i_T(S)$ and if c is the center of S, $i_T(c)$ is *not* the center of $i_T(S)$. Let d denote the center of T. One way to see this is to look at the effect of i_T on the ray r_c from d through c. Then the ray r_c intersects S at points p, q with $d(p,c) = d(c,q)$. However, we have

$$d(d, i_T(p)) = \frac{r^2}{d(d,p)}, d(d, i_T(q)) = \frac{r^2}{d(d,q)}, d(d, i_T(c)) = \frac{r^2}{d(d,c)}$$

so in particular $d(i_T(p), i_T(c)) \neq d(i_T(q), i_T(c))$. By definition, ϕ preserves the unit sphere and its center, so i is the identity.

Thus $\phi(t) = \lambda A(x) + t$. Since it preserves the unit sphere, $\lambda = 1$ and $t = 0$. Thus $\phi \in O(3,\mathbb{R})$. The lemma follows. $\square$

We define the hyperbolic metric at the origin by declaring it to be equal to 2 times the Euclidean metric there. That is,

$$\langle u, v\rangle^{\text{hyp}}_{T_0\mathbb{D}^3} = 4\langle u, v\rangle^{\text{Euc}}_{T_0\mathbb{D}^3}$$

Now, for any $p \in \mathbb{D}^3$ we define the hyperbolic inner product on $T_p\mathbb{D}^3$ by

$$\langle u, v\rangle^{\text{hyp}}_{T_p\mathbb{D}^3} = \langle d\alpha(u), d\alpha(v)\rangle^{\text{hyp}}_{T_0\mathbb{D}^3}$$

where $\alpha \in G$ takes p to 0.

To check that this is well–defined, suppose β is some other element with $\beta(p) = 0$. Then $\alpha\beta^{-1} \in \text{stab}(0)$ and therefore the action of $d(\alpha\beta^{-1}) = d\alpha d(\beta^{-1})$ on $T_0\mathbb{D}^3$

preserves lengths of vectors in the Euclidean metric there, and therefore also in the hyperbolic metric. So

$$\langle d\alpha(u), d\alpha(v)\rangle^{\text{hyp}}_{T_0\mathbb{D}^3} = \langle d\beta d(\alpha^{-1})d\alpha(u), d\beta d(\alpha^{-1})d\alpha(v)\rangle^{\text{hyp}}_{T_0\mathbb{D}^3} = \langle d\beta(u), d\beta(v)\rangle^{\text{hyp}}_{T_0\mathbb{D}^3}$$

and the inner product at $T_p\mathbb{D}^3$ is well–defined. In the sequel we will drop the subscripts and superscripts from our inner products if it is clear where and how they are defined.

A similar calculation shows that this metric is invariant under the action of G. That is, if $u, v \in T_p\mathbb{D}^3$ and $\alpha \in G$ is arbitrary,

$$\langle u, v\rangle = \langle d\alpha(u), d\alpha(v)\rangle$$

Exercise 3.3.28. *Verify the claim made in the last paragraph.*

Since stab$(0) = O(3, \mathbb{R})$ preserves the Euclidean length of vectors, and acts transitively on the set of points at Euclidean distance r from 0, there are the following consequences:

1. The hyperbolic distance from 0 to p only depends on the Euclidean distance r from 0 to p.
2. The ratio of the Euclidean to the hyperbolic metric on $T_p\mathbb{D}^3$ only depends on the Euclidean distance r from 0 to p.

Exercise 3.3.29. *Show that*

$$\langle u, v\rangle^{hyp} = \frac{4}{(1-r^2)^2}\langle u, v\rangle^{Euc}$$

for $u, v \in T_p\mathbb{D}^3$ where the Euclidean distance from 0 to p is r. In particular,

$$|u|^{hyp} = \frac{2}{1-r^2}|u|^{Euc}$$

Geodesics and totally geodesic planes

Definition 3.3.30. In a Riemannian manifold, a curve $\alpha : I \to M$ is a *geodesic* if it is *locally distance minimizing*. That is, there is a fine mesh $p_i \in I$ so that

$$d_M(\alpha(p_i), \alpha(p_{i+1})) = \text{length}(\alpha([p_i, p_{i+1}]))$$

Note that any (globally) distance–minimizing curve, if it exists, is a geodesic.

We have the following basic theorem of Riemannian geometry:

Theorem 3.3.31. *Let M be a Riemannian manifold. For any $p \in M$ there exists a neighborhood U of p such that for all $q \in U$, there is a unique distance–minimizing geodesic from p to q.*

Theorem 3.3.32. *Let $v \in T_pM$ be arbitrary. Then there is a unique geodesic emanating from p which is tangent to v at p. If M is* complete *as a metric space, any geodesic can be continued indefinitely in either direction.*

For a proof, see [11].

Theorem 3.3.33. *Suppose M is a Riemannian manifold and $i : M \to M$ is an isometry. Then the fixed point set of i consists of totally geodesic components.*

Proof. If some fixed points of i are isolated, they are totally geodesic 0–dimensional components. Suppose α is a 1–dimensional component, and p, q are two close points on α. Then if γ is the unique distance minimizing geodesic between p and q, the image of γ under i is another curve of the same distance between p and q, so it must necessarily be γ. Moreover, if $r \in \gamma$ is such that $d(p, r) = t$, the image $i(r) \in \gamma$ must also satisfy $d(p, i(r)) = t$ since $d(p, i(r)) = d(i(p), i(r)) = d(p, r)$ since i is an isometry. Hence $r = i(r)$ and γ is fixed by i. Since the fixed set is 1–dimensional near p, γ is a segment of α. So α is a geodesic. A similar argument works in higher dimensions. $\square$

Let S be a sphere perpendicular to the unit sphere in $\mathbb{E}^3$. Then i_S is a conformal transformation of the unit ball to itself, and therefore an isometry of the hyperbolic metric. The fixed point set of i_S is exactly the intersection $S \cap \mathbb{D}^3$. If S_1, S_2 are two spheres perpendicular to the unit sphere and transverse to each other, their intersection is a circle or straight line perpendicular to the unit sphere. It follows that the arcs of circles and straight line segments in $\mathbb{D}^3$ perpendicular to the boundary are geodesics. By the uniqueness of geodesics passing through a point with a given tangent vector, one sees that these are exactly the set of all geodesics. That is

Theorem 3.3.34. *The geodesics in $\mathbb{D}^3$ with the hyperbolic metric are the arcs of circle and straight line segments perpendicular to the boundary.*

3.3.4 Matrix models of Isom$^+(\mathbb{H}^3)$

The conformal models

In the conformal upper half–space model of $\mathbb{H}^3$, the sphere at infinity S^2_∞ can be identified with the Riemann sphere $\widehat{\mathbb{C}} = \mathbb{C} \cup \infty$. The isometries of $\mathbb{H}^3$ are the restriction of conformal transformations of $\mathbb{R}^3 \cup \infty$ which fix the upper half–space, and therefore they restrict to conformal transformations of $\widehat{\mathbb{C}}$. A conformal transformation ϕ from $\widehat{\mathbb{C}}$ to itself is necesarily analytic. One can easily check that the group of analytic isomorphisms from $\widehat{\mathbb{C}}$ to itself is exactly the group $PSL(2, \mathbb{C})$ consisting of 2×2 complex matrices of determinant 1 modulo the subgroup $\pm$Id, and acting on $\widehat{\mathbb{C}}$ by transformations

$$\begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix} : z \to \frac{\alpha z + \beta}{\gamma z + \delta}$$

Since every complex number has a complex square root, there is an isomorphism between the groups $PSL(2, \mathbb{C}) = PGL(2, \mathbb{C})$ where $PGL(2, \mathbb{C})$ denotes the group of 2×2 complex matrices of nonzero determinant modulo the subgroup of matrices which are scalar multiples of the identity; that is, matrices of the form $\begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}$.

With this notation, we can think of $\widehat{\mathbb{C}}$ as the *complex projective line* $\mathbb{CP}^1$. That is, we can think of it as the equivalence space $(\mathbb{C}^2 - 0)/\mathbb{C}^*$, the space of ratios of pairs of complex numbers which are not both zero. Geometrically, this is the space of complex lines in $\mathbb{C}^2$. Two points (z_1, w_1) and (z_2, w_2) both not the origin are on the same complex line iff there is some $\lambda \in \mathbb{C}^*$ with $z_2 = \lambda z_1$ and $w_2 = \lambda w_1$. So the two ratios $(z_1 : w_1)$ and $(z_2 : w_2)$ are equal. The action of $PGL(2, \mathbb{C})$ on $\mathbb{CP}^1$ is the projectivization of the usual action of $GL(2, \mathbb{C})$ on $\mathbb{C}^2$.

These transformations extend naturally to conformal transformations of the upper half–space, and we see therefore that this group is isomorphic to Isom$^+(\mathbb{H}^3)$. Thinking of the upper half space as the subspace

$$H = \{x + iy + jz \in \text{quaternions} \mid z > 0\}$$

we can extend the action of $A \in PSL(2, \mathbb{C})$ as follows:

$$\begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix} : q \to (\alpha q + \beta)(\gamma q + \delta)^{-1}$$

Note that the order of multiplication of the quaternionic factors is important.

Definition 3.3.35. Let A be a 2×2 complex matrix with determinant 1 representing an element of $PSL(2, \mathbb{C})$. We say

- A is *elliptic* if tr(A) is real and contained in $(-2, 2)$.
- A is *parabolic* if tr(A) is real and equal to ± 2.
- A is *hyperbolic* if tr(A) is real and has absolute value > 2.
- A is *loxodromic* if tr(A) is not real.

Theorem 3.3.36. *The notation above characterizes the nature of the action on $\mathbb{H}^3$ of an element of Isom$^+(\mathbb{H}^3)$. A transformation is* elliptic *if it is a rotation around some axis. It is* parabolic *if it fixes a unique point in S^2_∞ and no points in the interior of $\mathbb{H}^3$. It is* hyperbolic *if it acts as a translation along some geodesic axis, and* loxodromic *if it acts as a "screw motion" along some geodesic axis.*

Proof. The conditions on the trace are invariant under $A \to -A$ so the type of a transformation is well–defined independent of the representative in $SL(2, \mathbb{C})$. Moreover, trace is invariant under conjugation. Since the equation $\alpha z + \beta = \gamma z^2 + \delta z$ always has a solution in $\widehat{\mathbb{C}}$, we can assume after conjugation that ∞ is a fixed point — that is, that our matrix A is of the form $\begin{bmatrix} \alpha & \beta \\ 0 & \alpha^{-1} \end{bmatrix}$. Such a transformation acts on $\mathbb{C}$ by $z \to \alpha^2 z + \alpha\beta$, which is to say, by an arbitrary similarity of the standard Euclidean structure on $\mathbb{C}$. One sees that the transformation is parabolic iff $\alpha = \pm 1$, in which case it acts on $\mathbb{C}$ as a translation. It is elliptic iff $|\alpha| = 1$ and $\alpha \neq \pm 1$ in which case it acts on $\mathbb{C}$ as a rotation about some point p, and on $\mathbb{H}^3$ as a rotation about the axis joining p to ∞. It is hyperbolic iff α is real and not ± 1 in which case it acts on $\mathbb{C}$ as a dilation about some point p, and on $\mathbb{H}^3$ as a translation along the axis joining p to ∞. It is loxodromic iff it is a product of an elliptic and a hyperbolic element with the same axis, in which case it acts on $\mathbb{H}^3$ as a screw motion, translating along an axis and rotating about it simultaneously. $\square$

The other conformal model of $\mathbb{H}^3$ is the unit ball. Define

$$B = \{x + iy + jz \in \text{quaternions} \mid x^2 + y^2 + z^2 < 1\}$$

For a quaternion $q = q_0 + q_1 i + q_2 j + q_3 k$ denote $q' = q_0 - q_1 i - q_2 j + q_3 k$ and $\overline{q} = q_0 - q_1 i - q_2 j - q_3 k$. Then define $SB(2, H)$ to be the group of 2×2 matrices with quaternionic entries of the form $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ such that $d = a', b = c'$ and $a\overline{a} - c\overline{c} = 1$. The natural action of $SB(2, H)$ on B by fractional linear transformations as above has kernel exactly $\pm$Id and gives an isomorphism of $SB(2, H)/\pm$Id with the group of conformal transformations of B — that is, with the group $Isom^+(\mathbb{H}^3)$ for the conformal unit ball model.

The hyperboloid model

Let $\langle \cdot, \cdot \rangle$ denote the symmetric bilinear form on $\mathbb{R}^{n+1}$ which in terms of a basis $e_1 \dots e_{n+1}$ is given by

$$\left\langle \sum_{i=1}^{n+1} r_i e_i, \sum_{i=1}^{n+1} s_i e_i \right\rangle = \sum_{i=1}^{n} r_i s_i - r_{n+1} s_{n+1}$$

For a vector v, define the *Lorentz norm* $\|v\| = \langle v, v \rangle^{\frac{1}{2}}$.

Let J denote the matrix

$$J = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & -1 \end{bmatrix}$$

The group of linear transformations of $\mathbb{R}^{n+1}$ which preserve the Lorentz norm is exactly the matrices A such that $A^t J A = J$, and is denoted $O(n, 1)$. The subgroup of determinant 1 is $SO(n, 1)$.

The set of vectors of norm i is a two–sheeted hyperboloid. Denote the positive sheet (i.e. for which the last co–ordinate is positive) by H^+, and the subgroup preserving this sheet as $SO^+(n, 1)$. Then $(H^+, SO^+(n, 1))$ is a model for hyperbolic n space $\mathbb{H}^n$.

If v is a vector, $v^\perp$ denotes the subspace of vectors w such that $\langle v, w \rangle = 0$. If v has imaginary norm, every vector in w has real (positive) norm, and the restriction of $\langle \cdot, \cdot \rangle$ to $v^\perp$ gives a standard Euclidean inner product. For $v \in H^+$ the tangent space $T_v H^+$ is a translate of $v^\perp$. To see this, observe that w is tangent to H^+ at v iff

$$\frac{d}{dt} \langle v + tw, v + tw \rangle = 0$$

Giving $T_v H^+$ the inner product $\langle \cdot, \cdot \rangle$ gives H^+ the structure of a Riemannian manifold. Since everything is invariant under the action of $SO^+(n, 1)$, this must be (a scaled multiple of) the standard hyperbolic metric on H^+. It is easy to check by a direct computation that this metric on the space is given by

$$\cosh(d(v, w)) = -\langle v, w \rangle$$

for vectors v, w of norm 1.

Example 3.3.37. In $\mathbb{R}^{25}$ consider the "standard" form g of signature $(24, 1)$. The vector

$$\lambda = (3, 5, 7, \ldots, 47, 51, 145)$$

satisfies $|\lambda|^2 = -1$, so the restriction of the form g to $L = \lambda^\perp$ is positive definite. The intersection of the cubical lattice I with L is a lattice in L called the *Leech lattice*. See [12] for more details, and many more constructions of the Leech and other lattices.

3.3.5 Horospherical geometry

In one of the conformal models of hyperbolic geometry, a Euclidean sphere whose interior is contained in $\mathbb{H}^3$ and is tangent to S^2_∞ at some point is called a *horosphere*. In the upper half–space model, a sphere tangent to ∞ is a horizontal plane. The stabilizer of a horosphere can be seen to be isomorphic to the group of isometries of $\mathbb{E}^2$, that is, to $\mathbb{R}^2 \rtimes O(2)$. The orientation–preserving subgroup preserving ∞ in the upper half–space model is exactly the group of matrices of the form $\begin{bmatrix} \theta & \alpha \\ 0 & \theta^{-1} \end{bmatrix}$, where θ is a complex number of norm 1, and $\alpha \in \mathbb{C}$ is arbitrary. The intrinsic geometry $(H, \text{stab}(H))$ is thus identical to $(\mathbb{E}^2, \text{Isom}(\mathbb{E}^2))$. That is, the intrinsic geometry of a horosphere is Euclidean.

A horosphere is a limit of spheres which are contained entirely in the interior of $\mathbb{H}^3$. These spheres are isometric to Euclidean spheres of various sizes; a horosphere can be thought of as a "sphere of infinite radius", and exhibits Euclidean geometry as a limit of spherical geometry in the small.

Definition 3.3.38. A *Busemann function* is a function $f : \mathbb{H}^3 \to \mathbb{R}$ defined as a limit

$$f(p) = \lim_{i \to \infty} \text{dist}(p, p_i) - \text{dist}(q, p_i)$$

for some sequence $p_i \to p_\infty \in S^2_\infty$ and fixed $q \in \mathbb{H}^3$.

A Busemann function measures the relative distance to a point $p_\infty \in S^2_\infty$. The level sets of a Busemann function are the horospheres tangent to p_∞. Thus p_∞ may be thought of as the "center" of these horospheres, and the horospheres may be thought of as the locus of points equidistant to an ideal point. One must be careful not to abuse this metaphor, but it is a useful and suggestive one.

3.3.6 Volumes of ideal hyperbolic simplices

Let Δ be an ideal tetrahedron — that is, a geodesic tetrahedron with all four vertices contained in S^2_∞. After translating Δ by an element of $PSL(2, \mathbb{C})$, we can assume that the vertices are at $0, 1, z, \infty$ for some z. For different choices of vertices to be sent to $0, 1, \infty$ the parameter z is replaced by $\frac{1}{1-z}$ or $\frac{z-1}{z}$ for an even permutation, or $\frac{1}{z}, \frac{1-z}{1}, \frac{z}{z-1}$ for an odd permutation. We say an ordering of vertices is *positively oriented* if $\text{Im}(z) > 0$. The space of (possibly degenerate) oriented ideal tetrahedra is isomorphic to the quotient of $\mathbb{C} - \{0, 1\}$ by the group G generated by $z \to \frac{1}{1-z}$ and

$z \to \frac{1}{z}$. This group has order 6, and the quotient of the thrice punctured sphere by this group is the $(2, 3, \infty)$ orbifold, which is exactly the moduli space of flat structures on a torus. The correspondence is given as follows: if $T = \mathbb{C}/\Lambda$ is a flat torus, we can form the quotient $T/\langle i \rangle$ where i is the involution $z \to -z$. The quotient is an orbifold S which is a sphere with four orbifold points of order 2. The underlying sphere is conformally equivalent to $\mathbb{CP}^1$, and the four orbifold points pick out 4 distinct points in $\mathbb{CP}^1$.

Example 3.3.39. With the notation as above, what is the locus of flat structures on the torus which correspond to real simplex parameters?

If Δ is an ideal simplex with one vertex at infinity, the three faces incident to this vertex are subsets of vertical Euclidean planes, in the upper half–space model. A horosphere S centered at infinity is a horizontal plane, so the intersection of Δ with S is a Euclidean triangle. If γ is an edge of Δ incident to ∞, represented by a vertical line in this picture, a hyperbolic plane perpendicular to γ at some point $p = S \cap \gamma$ is tangent to S at that point. It follows that the hyperbolic dihedral angle between the planes incident at γ is equal to the Euclidean angle at p in the triangle $S \cap \Delta$. It follows that for three edges $\gamma_1, \gamma_2, \gamma_3$ incident to a vertex of an ideal simplex, the sum of the dihedral angles at those edges is π.

If the six dihedral angles of Δ are labelled $\alpha, \beta, \gamma, \alpha', \beta', \gamma'$ where α and α' are the angles on "opposite" edges, and so on for β, γ then we have identities

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

It follows that $\alpha = \alpha', \beta = \beta', \gamma = \gamma'$. That is, we have proved

Lemma 3.3.40. *Dihedral angles of opposite edges of an ideal hyperbolic simplex are equal.*

In fact, it is easy to see from this that a permutation of the vertices of Δ by $(12)(34)$ or its conjugates can be realized as a hyperbolic isometry. There are 3 such conjugates; these elements together with an isometry form a normal subgroup of S_4, the group of permutations of the vertices of a simplex, called the *Klein* 4*–group* K. The quotient group S_4/K is the group G discussed above.

Definition 3.3.41. The *Lobachevsky function* $\Lambda(\theta)$ is defined by the following integral formula

$$\Lambda(\theta) = -\int_0^\theta \log|2\sin t| dt$$

for $\theta \in \mathbb{R}$.

We establish some basic properties of the function Λ.

Theorem 3.3.42. *The function* $\Lambda(\theta)$ *satisfies the following properties:*

1. Λ *is continuous*

2. Λ *is odd*

3. *$\Lambda'(\theta)$ exists for all $\theta \neq k\pi$ for $k \in \mathbb{Z}$*

4. *Λ is periodic with period π*

5. *in the interval $[0,\pi]$ the function Λ has a unique maximum at $\frac{\pi}{6}$*

6. *for $m \in \mathbb{Z}$ we have*
$$\Lambda(m\theta) = m\sum_{k=0}^{m-1} \Lambda\Big(\theta + \frac{k\pi}{m}\Big)$$

Proof. To prove 1 it is obvious that Λ is continuous for θ not a multiple of π. For θ close to 0, $2\sin(\theta) \sim 2\theta$, and $\int_0^\theta \log t dt = t\log t - t|_0^\theta$ which is certainly finite, since t goes to 0 much faster than $\log t$ blows up.

2 is obvious, since the integrand $\log|2\sin t|$ is even.

3 is immediate from the existence of the integral formula.

4 will follow if we can show $\int_0^\pi \log|2\sin t|dt = 0$. We use the identity
$$\sin(2t) = 2\sin t\cos t = 2\sin t\sin(t+\pi/2)$$
which gives us
$$\frac{\Lambda(2\theta)}{2} = -\frac{1}{2}\int_0^{2\theta} \log|2\sin t|dt = -\int_0^\theta \log|2\sin(2t)|dt$$
$$= -\int_0^\theta \log|2\sin t|dt - \int_0^\theta \log|2\sin(t+\pi/2)|dt$$
$$= \Lambda(\theta) - \int_{\pi/2}^{\pi/2+\theta} \log|2\sin t|dt = \Lambda(\theta) + \Lambda(\theta+\pi/2) - \Lambda(\pi/2)$$
So if $\theta = \pi/2$ this formula shows $\Lambda(\pi) = 0$.

To prove 5, observe that $\log|2\sin t|$ vanishes when $2\sin t = 1$, so $t = \pi/6$ or $5\pi/6$. The derivative of $\Lambda(\theta)$ is negative in the interval $(\pi/6, 5\pi/6)$ so the unique maximum is at $\pi/6$.

Finally, to prove 6, we need the following identity for an integer m:
$$2\sin(mt) = \prod_{k=0}^{m-1} 2\sin(t + k\pi/m)$$
Denoting $e^{it} = z$ and $e^{ik\pi/m} = \omega$, the left hand side of this expression is equal to
$$-i\Big(z^m - \frac{1}{z^m}\Big)$$
and the right hand side is equal to
$$(-i)^m\Big(z - \frac{1}{z}\Big)\Big(z\omega - \frac{1}{z\omega}\Big)\Big(z\omega^2 - \frac{1}{z\omega^2}\Big)\ldots\Big(z\omega^{m-1} - \frac{1}{z\omega^{m-1}}\Big)$$

Evidently, the product of the $z\omega^k$ terms gives $-iz^m$. The product of the $\frac{1}{z\omega^k}$ terms gives iz^{-m}. Each cross term has $\sum_{j=0}^{m-1} \omega^j$ as a factor, but this expression is equal to zero, so there are no other terms, and the identity is proved.

It follows that

$$\frac{\Lambda(m\theta)}{m} = -\frac{1}{m}\int_0^{m\theta} \log|2\sin t|dt = -\int_0^{\theta} \log|2\sin(mt)|dt$$

$$= -\sum_{k=0}^{m-1}\int_0^{\theta} \log|2\sin(t + k\pi/m)|dt$$

$$= -\sum_{k=0}^{m-1}\left(\int_0^{\theta+k\pi/m} \log|2\sin t|dt - \int_0^{k\pi/m} \log|2\sin t|dt\right)$$

The second term doesn't involve θ, and therefore gives a constant $C(m)$ depending only on m.

$$= \sum_{k=0}^{m-1} \Lambda(\theta + k\pi/m) + C(m)$$

Since Λ is odd and π–periodic, $\int_0^{\pi} \frac{\Lambda(m\theta)}{m} d\theta = 0$ and $\int_0^{\pi} \sum_{k=0}^{m-1} \Lambda(\theta + k\pi/m)d\theta = 0$. It follows that $C(m) = 0$ and the theorem is proved. □

The volume of an ideal hyperbolic simplex has a very simple description in terms of the Lobachevsky function.

Theorem 3.3.43 (Lobachevsky). *If σ is an ideal simplex with dihedral angles α, β, γ,*

$$\textit{volume}(\sigma) = \Lambda(\alpha) + \Lambda(\beta) + \Lambda(\gamma)$$

Proof. Place one of the vertices at ∞. Then σ is bounded by four hyperbolic planes, three of which can be realized as vertical planes in the upper half–space model. These vertical planes intersect $\mathbb{C}$ in three lines cutting out a Euclidean triangle T with angles α, β, γ. By a hyperbolic isometry, we can arrange for the vertices of T to lie on the unit circle in $\mathbb{C}$. We cut T into six triangles, such that each triangle has vertices consisting of a vertex of T, a midpoint of an edge of T incident to that vertex, and 0. These triangles are right angled, with the angle incident to 0 one of α, β or γ. Up to isometry, there are two regions K_α, and two each of K_β, K_γ. We calculate the volume of the region in σ lying above one of these triangles.

$$\text{volume}(K_\alpha) = \int_{x=0}^{\cos\alpha} dx \int_{y=0}^{x\tan\alpha} dy \int_{z=\sqrt{1-x^2-y^2}}^{\infty} \frac{dz}{z^3}$$

$$= \frac{1}{2}\int_0^{\cos\alpha} dx \int_0^{x\tan\alpha} \frac{dy}{1-x^2-y^2}$$

$$= \frac{1}{4}\int_0^{\cos\alpha} \log\frac{\sqrt{1-x^2}\cos\alpha + x\sin\alpha}{\sqrt{1-x^2}\cos\alpha - x\sin\alpha}\frac{dx}{\sqrt{1-x^2}}$$

$$= -\frac{1}{4}\int_{\pi/2}^{\alpha} \log \frac{\sin t \cos\alpha + \cos t \sin\alpha}{\sin t \cos\alpha - \cos t \sin\alpha} dt$$

$$= \frac{1}{4}\int_{\alpha}^{\pi/2} \log \frac{\sin(t+\alpha)}{\sin(t-\alpha)} dt$$

$$= \frac{1}{4}\int_{\alpha}^{\pi/2} \log|2\sin(t+\alpha)|dt - \frac{1}{4}\int_{\alpha}^{\pi/2} \log|2\sin(t-\alpha)|dt$$

$$= \frac{1}{4}\int_{2\alpha}^{\alpha+\pi/2} \log|2\sin t|dt - \frac{1}{4}\int_{0}^{\pi/2-\alpha} \log|2\sin t|dt$$

$$= \frac{1}{4}(-\Lambda(\alpha+\pi/2) + \Lambda(2\alpha) + \Lambda(\pi/2-\alpha)) = \frac{\Lambda(\alpha)}{2}$$

and the theorem follows. □

The Lobachevsky function and the Dilogarithm

The Lobachevsky function is related to the *dilogarithm function*

Definition 3.3.44. The *dilogarithm function* is defined by the power series

$$\mathrm{Li}_2(z) = \sum_{n=1}^{\infty} \frac{z^n}{n^2}$$

Lemma 3.3.45. *The dilogarithm function is defined by the following integral formula*

$$Li_2(z) = \int_0^z \frac{-\log(1-x)}{x} dx$$

Proof. This follows from the following computation

$$-\log(1-x) = \int_0^x \frac{1}{1-s} ds = \int_0^x 1 + s + s^2 + \dots ds$$

$$= x + \frac{x^2}{2} + \frac{x^3}{3} + \dots$$

Dividing by x and integrating termwise, the identity follows. □

Lemma 3.3.46. *The dilogarithm and Lobachevsky functions are related by the identity*

$$Li_2(e^{2i\theta}) = \frac{\pi^2}{6} - \theta(\pi-\theta) + 2i\Lambda(\theta)$$

Proof. We can easily calculate

$$\frac{d}{d\theta}\mathrm{Li}_2(e^{2i\theta}) = 2\theta - \pi - 2i\log(2\sin\theta)$$

which proves the identity up to an additive constant. Evaluating either side at $z \to 0$ and applying Euler's formula $\sum \frac{1}{n^2} = \frac{\pi^2}{6}$ the formula follows. □

Taking imaginary parts of both sides, we get a Fourier series expansion for $\Lambda(\theta)$:

$$2\Lambda(\theta) = \sum \frac{\sin(2n\theta)}{n^2}$$

For our purposes, the most significant relation between the dilogarithm and the Lobachevsky function is the following theorem of Kummer:

Theorem 3.3.47 (Kummer).

$$\mathit{Im}(\mathit{Li}_2(z)) = \Lambda(\arg z) + \Lambda\Big(\arg \frac{z-1}{z}\Big) + \Lambda\Big(\arg \frac{1}{1-z}\Big) - \log|z| \arg(1-z)$$

Defining $D(z) = \mathrm{Im}(\mathrm{Li}_2(z)) + \log|z|\arg(1-z)$ we see that the volume of an ideal simplex with parameter z is $D(z)$. It follows that $D(z)$ satisfies the following identities:

$$D(z) = D(1 - \frac{1}{z}) = D(\frac{1}{1-z})$$

and

$$D(z) = -D(-z)$$

For $\omega = e^{i\pi/3}$, one can calculate

$$D(\omega) = 1.0149\ldots$$

the volume of a regular ideal tetrahedron.

Remark 3.3.48. The polylogarithms are defined as power series by the formulae

$$\mathrm{Li}_m(z) = \sum \frac{z^n}{n^m}$$

Volumes of $(2m-1)$–dimensional ideal hyperbolic simplices are related to values of Li_m.

Chapter 4

Discrete groups

4.1 A crash course in algebraic topology

4.1.1 The topology of surfaces

Gluing polygons

Certain computer games get around the constraint of a finite screen by means of a trick: when a spaceship comes to the left side of the screen, it disappears and "reappears" on the right side of the screen. Likewise, an asteroid which disappears beyond the top of the screen might reappear menacingly from the bottom. The screen can be represented by a square whose sides are *labelled in pairs*: the left and right sides get one label, the top and bottom sides get another label. These labels are instructions for obtaining an idealized topological space from the flat screen: the left and right sides can be glued together to make a cylinder, then the top and bottom sides can be glued together to make a torus (the surface of a donut). Actually, we have to be somewhat careful: there are *two* ways to glue two sides together; an unambiguous instruction must specify the orientation of each edge.

If we have a collection of polygons P_i to be glued together along pairs of edges, we can imagine a graph Γ whose vertices are the polygons, and whose edges are the pairs of edges in the collection. We can glue in any order. If we first glue along the edges corresponding to a *maximal tree* in Γ, the result of this first round of gluing will produce a connected polygon. Thus, without loss of generality, it suffices to consider gluings of the sides of a single polygon.

Definition 4.1.1. A *surface* is a 2–dimensional manifold. That is, a Hausdorff topological space with a countable basis, such that every point has a neighborhood homeomorphic to the open unit disk in $\mathbb{R}^2$ (see appendix for definitions). A *piecewise–linear surface* is a surface obtained from a countable collection of polygons by glueing together the edges in pairs, in such a way that only finitely many edges are incident to any vertex.

Exercise 4.1.2. *Why is the finiteness condition imposed on vertices?*

The following theorem was proved by T. Rado in 1924 (see [6]):

Theorem 4.1.3. *Any surface is homeomorphic to a piecewise–linear surface. Any compact surface is homeomorphic to a piecewise–linear surface made from only finitely many polygons.*

Definition 4.1.4. A surface is *oriented* if there is an unambiguous choice of "top" and "bottom" side of each polygon which is compatible with the glueing. i.e. an orientable surface is "two–sided".

The fundamental group

The definition of the fundamental group of a surface Σ requires a choice of a *basepoint* in Σ. Let $p \in \Sigma$ be such a point.

Definition 4.1.5. Define $\Omega_1(\Sigma, p)$ to be the space of continuous maps $c : S^1 \to \Sigma$ sending $0 \in S^1$ to $p \in \Sigma$ (here we think of S^1 as $I/0 \sim 1$). Two such maps c_1, c_2 are called *homotopic* if there is a map $C : S^1 \times I \to \Sigma$ such that

1. $C(\cdot, 0) = c_1(\cdot)$.
2. $C(\cdot, 1) = c_2(\cdot)$.
3. $C(0, \cdot) = p$.

Exercise 4.1.6. *Show that the relation of being homotopic is an* equivalence relation *on* $\Omega_1(\Sigma, p)$.

In fact, $\Omega_1(\Sigma, p)$ has the natural structure of a topological space; with respect to this topological structure, the equivalence classes determined by the homotopy relation are the path–connected components.

The importance of the relation of homotopy equivalence is that the equivalence classes form a *group*:

Definition 4.1.7. The *fundamental group* of Σ with basepoint p, denoted $\pi_1(\Sigma, p)$, has as elements the equivalence classes

$$\pi_1(\Sigma, p) = \Omega_1(\Sigma, p)/\text{homotopy equivalence}$$

with the group operation defined by

$$[c_1] \cdot [c_2] = [c_1 * c_2]$$

where $c_1 * c_2$ denotes the map $S^1 \to \Sigma$ defined by

$$c_1 * c_2(t) = \begin{cases} c_1(2t) & \text{for } t \leq 1/2 \\ c_2(2t-1) & \text{for } t \geq 1/2 \end{cases}$$

where the identity is given by the equivalence class $[e]$ of the constant map $e : S^1 \to p$, and inverse is defined by $[c]^{-1} = [i(c)]$, where $i(c)$ is the map defined by $i(c)(t) = c(1-t)$.

Exercise 4.1.8. *Check that* $[c][c]^{-1} = [e]$ *with the definitions given above, so that* $\pi_1(\Sigma, p)$ *really is a group.*

For Σ a *piecewise–linear surface* with basepoint v a vertex of Σ, define $O_1(\Sigma)$ to be the space of polygonal loops γ from v to v contained in the edges of Σ. Such a loop consists of a sequence of oriented edges

$$\gamma = e_1, e_2, \dots, e_n$$

where e_1 starts at v and e_n ends there, and e_i ends where e_{i+1} starts. Let $r(e)$ denote the same edge e with the opposite orientation. For a polygonal path γ (not necessarily starting and ending at v) let γ^{-1} denote the path obtained by reversing the order and the orientation of the edges in γ.

One can perform an *elementary move* on a polygonal loop γ, which is one of the following two operations:

- If there is a vertex w which is the endpoint of some e_i, and α is any polygonal path beginning at w, we can insert or delete $\gamma\gamma^{-1}$ between e_i and e_{i+1}. That is,

$$e_1, e_2, \dots, e_i, \gamma, \gamma^{-1}, e_{i+1}, \dots, e_n \longleftrightarrow e_1, e_2, \dots, e_n$$

- If γ is a loop which is the boundary of a polygonal region, and starts and ends at a vertex w which is the endpoint of some e_i, then we can insert or delete γ between e_i and e_{i+1}. That is,

$$e_1, e_2, \dots, e_i, \gamma, e_{i+1}, \dots, e_n \longleftrightarrow e_1, e_2, \dots, e_n$$

Definition 4.1.9. Define the *combinatorial fundamental group* of Σ, denoted $p_1(\Sigma, v)$, to be the group whose elements are the equivalence classes

$$O_1(\Sigma, v)/\text{elementary moves}$$

With the group operation defined by $[\alpha][\beta] = [\alpha\beta]$, where the identity is given by the "empty" polygonal loop 0 starting and ending at v, and with inverse given by $i(\gamma) = \gamma^{-1}$.

Exercise 4.1.10. *Check that the above makes sense, and that* $p_1(\Sigma, v)$ *is a group.*

Now suppose that Σ is obtained by glueing up the sides of a single polygon P in pairs. Suppose further that after the result of this glueing, all the vertices of this polygon are identified to a single vertex v. Let $e_1, \dots, e_n$ be the edges and

$$\gamma = e_{i_1}^{\pm 1} e_{i_2}^{\pm 1} \dots e_{i_m}^{\pm 1}$$

the oriented boundary of P. Then there is a natural isomorphism

$$p_1(\Sigma, v) = \langle e_1, \dots, e_n | \gamma \rangle$$

that is, p_1 can be thought of as the group *generated by* the edges e_i subject to the *relation* defined by γ. Notice that each of the e_i appears twice in γ, possibly with distinct signs. If Σ is orientable, each e_i appears with opposite signs.

Example 4.1.11. Let T denote the surface obtained by glueing opposite sides of a square by translation. Thus the edges of the square can be labelled (in the circular ordering) by a, b, a^{-1}, b^{-1} and a presentation for the group is

$$p_1(T, v) = \langle a, b | [a, b] \rangle$$

It is not too hard to see that this is isomorphic to the group $\mathbb{Z} \oplus \mathbb{Z}$.

Homotopy theory

The following definition generalizes the notation of homotopy equivalence of maps $S^1 \to \Sigma$:

Definition 4.1.12. Two continuous maps $f_1, f_2 : X \to Y$ are *homotopic* if there is a map $F : X \times I \to Y$ satisfying $F(\cdot, 0) = f_1$ and $F(\cdot, 1) = f_2$. If $M \subset X$ and $N \subset Y$ with $f_1|M = f_2|M$ and $f_i(M) \subset N$, then f_1, f_2 are *homotopic relative to* M if a map F can be chosen as above with $F(m, \cdot)$ constant for every $m \in M$.

The set of homotopy classes of maps from X to Y is usually denoted $[X, Y]$. If these space have basepoints x, y the set of maps taking x to y modulo the equivalence relation of homotopy relative to x is denoted $[X, Y]_0$.

We can define a very important category whose objects are topological spaces and whose morphisms are *homotopy equivalence classes of continuous maps*. A refinement of this category is the category whose objects are topological spaces with basepoints, and whose morphisms are *homotopy equivalence classes of continuous maps relative to base points*.

Definition 4.1.13. The *fundamental group* $\pi_1(X, x)$ of an arbitrary topological space with a basepoint x as the group whose elements are homotopy classes of maps $(S^1, 0) \to (X, x)$, where multiplication is defined by $[c_1][c_2] = [c_1 * c_2]$. In the notation above, the elements of $\pi_1(X, x)$ correspond to elements of $[S^1, X]_0$.

Lemma 4.1.14. *Let $f : X \to Y$ be a continuous map taking x to y. Then f induces a natural homomorphism $f_* : \pi_1(X, x) \to \pi_1(Y, y)$.*

Definition 4.1.15. A path–connected space X is *simply–connected* if $\pi_1(X, x)$ is the trivial group.

Observe that a path–connected space is simply–connected if and only if every loop in X can be shrunk to a point.

Simplicial approximation

The following is known as the *simplicial approximation theorem*:

Theorem 4.1.16. *Let K, L be two simplicial complexes. Then any continuous map $f : K \to L$ is homotopic to a* simplicial map *$f' : K' \to L$ where K' is obtained from K by subdividing simplices. Furthermore, if $C \subset K$ is a simplicial subset, and $f : C \to L$ is simplicial, then we can require f' to agree with f restricted to C.*

Exercise 4.1.17. *Using this theorem, show that by choosing $p = v$, every continuous map $c : S^1 \to \Sigma$ taking the base point to v is homotopic through basepoint–preserving maps to a simplicial loop $\gamma \subset \Sigma$ which begins and ends at v. Moreover, two such simplicial loops are homotopic if and only if they differ by a sequence of elementary moves. Thus there is a natural isomorphism $\pi_1(\Sigma, p) \cong p_1(\Sigma, v)$.*

This is actually a very powerful observation: the group $\pi_1(\Sigma, p)$ is *a priori* very difficult to compute, but manifestly doesn't depend on a piecewise linear structure on Σ. On the other hand, $p_1(\Sigma, v)$ is easy to compute (or at least find a presentation for), but it is *a priori* hard to see that this group, up to isomorphism, doesn't depend on the piecewise linear structure.

Exercise 4.1.18. *Suppose K is a simplicial complex. Let K^2 denote the union of the simplices of K of dimension at most 2. Use the simplicial approximation theorem to show that for any vertex v of K, $\pi_1(K, v) \cong \pi_1(K^2, v)$.*

Covering spaces

Definition 4.1.19. A space Y is a *covering space* for X if there is a map $f : Y \to X$ (called a *covering projection*) with the property that every point $x \in X$ has an open neighborhood $x \in U$ such that $f^{-1}(U)$ is a disjoint union of open sets $U_i \subset Y$, and f maps each U_i homeomorphically to U. The *universal cover* of a space X (if one exists) is a simply–connected space $\widetilde{X}$ which is a covering space of X.

An open neighborhood U of a point x of the kind provided in the definition is said to be *evenly covered* by its preimages $f^{-1}(U)$. If X is locally connected, we can assume that the open neighborhoods which are evenly covered are connected.

For a path $I \subset X$ we can find, for each point $p \in I$ an open neighborhood U_p of p which is evenly covered. Since I is *compact* (see appendix) only finitely many open neighborhoods are needed to cover I; call these $U_1, \ldots, U_n$. If we let V_1 denote some component of $f^{-1}(U_1)$ which maps homeomorphically to U_1. Then there is a unique map $g : I \cap U_1 \to V_1$ such that $fg = \mathrm{id}$. Moreover, there is a unique choice of V_2 from amongst the components of $f^{-1}(U_2)$ such that g can be extended to $g : I \cap (U_1 \cup U_2) \to V_1 \cup V_2$ with $fg = \mathrm{id}$. Continuing inductively, we see that the choice of $V_3, V_4, \ldots, V_n$ are all uniquely determined by the original choice V_1.

Exercise 4.1.20. *Modify the above argument to show that for every map $g : I \to X$ any every $p \in f^{-1}g(0)$ there is a unique* lift *$\widetilde{g} : I \to Y$ such that $\widetilde{g}(0) = p$ and $f\widetilde{g} = g$.*

Exercise 4.1.21. *Suppose $g_1, g_2 : I \to X$ with $g_1(0) = g_2(0)$ and $g_1(1) = g_2(1)$ are* homotopic *through homotopies which keep the endpoints of I fixed. Let $\widetilde{g_1}$ be some lift of g_1. Show that the lift $\widetilde{g_2}$ of g_2 with $\widetilde{g_1}(0) = \widetilde{g_2}(0)$ also satisfies $\widetilde{g_1}(1) = \widetilde{g_2}(1)$, and these two lifts are also homotopic rel. endpoints. (Hint: let $G : I \times I \to X$ be a homotopy between g_1 and g_2. Try to "lift" G to a map $\widetilde{G} : I \times I \to X$ satisfying appropriate conditions.)*

Let $p \in X$ be a basepoint, and let $\widetilde{p} \in f^{-1}(p)$. Then the projection $f : Y \to X$ induces a homomorphism $f_* : \pi_1(Y, \widetilde{p}) \to \pi_1(X, p)$. Let $K \subset \pi_1(X, p)$ denote the image of f_*. Then a loop $\alpha : S^1 \to X$ with $[\alpha] \in K$ can be lifted to a loop $\widetilde{\alpha} : S^1 \to Y$, by the argument of the previous exercise. Conversely, if two loops α, β represent the same element of K, then their lifts represent the same element of $\pi_1(Y, \widetilde{p})$. Thus we may identify $\pi_1(Y, \widetilde{p})$ with the subgroup K.

Exercise 4.1.22. *Show that for a space Z and a map $g : Z \to X$ there is a* lift *of g to $\widetilde{g} : Z \to Y$ with $f\widetilde{g} = g$ if and only if $g_*(\pi_1(Z)) \subset K$.*

Definition 4.1.23. A space is *semi–locally simply–connected* if every point p has a neighborhood U so that every loop in U can be shrunk to a point in a possibly larger open neighborhood V.

Informally, a space is semi–locally simply–connected if sufficiently small loops in the space are homotopically inessential.

Theorem 4.1.24. *Let X be connected, locally connected and semi–locally simply–connected. For any subgroup $G \subset \pi_1(X, x)$ there is a covering space X_G of X and a point $y \in f^{-1}(x)$ such that $f_*(\pi_1(X_G, y)) = G$.*

Proof. Let $\Omega(X)$ be the space of paths $\gamma : I \to X$ which start at x. We induce an equivalence relation on $\Omega(X)$ where we say two paths γ_1, γ_2 are equivalent if $[\gamma_2^{-1} * \gamma_1] \in G$. Set $X_G = \Omega(X)/\sim$. Since X is semi–locally simply connected, for sufficiently small paths γ_1, γ_2 between points $x_1, x_2 \in X$ the loop $\gamma_2^{-1} * \gamma_1$ is contractible. So any two paths which differ only by substituting γ_1 in one for γ_2 in the other will be equivalent in $\Omega(X)$, and this says that the equivalence classes of $\Omega(X)$ are parameterized locally by points in X; that is, X_G is a covered space of X. It can be verified that X_G is simply connected, and that it satisfies the conditions of the theorem. $\square$

Exercise 4.1.25. *Fill in the gaps in the sketch of the proof above.*

Exercise 4.1.26. *Show that the universal cover of a space X, it if exists, is unique, by using the lifting property.*

4.1.2 Topological 3–manifolds

A *manifold* is a Hausdorff, second countable topological space X with the property that every point $x \in X$ has a neighborhood (i.e. is contained in an open set) homeomorphic to $\mathbb{R}^n$ for some n. If X is connected, n is contstant, and is referred to as the *dimension* of the manifold. As discussed earlier, the topological structure on manifolds can sometimes be stiffened to smooth or piecewise–linear (PL) structures. The fundamental theorem pertinent to topological 3–manifolds is a theorem of Moise:

Theorem 4.1.27 (Moise). *Let M be a compact 3–manifold (possibly with boundary). Then $(M, \partial M)$ has a unique smooth or PL structure.*

For a proof, see [18].

4.1.3 Building 3–manifolds from tetrahedra

Informally speaking, a PL structure on a n manifold amounts to a description of the manifold as being built up from n–dimensional simplices in a combinatorial way, such that the local combinatorial pictures are "standard".

Definition 4.1.28. A *combinatorial* structure on a 3–manifold is a description of the manifold as the total space $|\Sigma|$ of a simplicial complex Σ satisfying the following conditions:

- Σ is made by gluing faces of a collection Δ_i of tetrahedra (3–simplices) in pairs
- There are only finitely many simplices glued together around each edge
- The link of every vertex is a sphere

Such conditions are certainly sufficient for $|\Sigma|$ to be a 3–manifold, and it is not too hard to verify that they are also necessary.

There is a useful combinatorial criterion which verifies that the link of every vertex is a sphere.

Lemma 4.1.29. *Let Σ be obtained from finitely many tetrahedra by glueing the faces in pairs. $|\Sigma|$ is a 3–manifold iff the Euler characteristic $\chi(\Sigma) = 0$.*

Proof. Recall that the Euler characteristic $\chi(\Sigma)$ is defined to be the alternating sum

$$\chi(\Sigma) = v - e + f - t$$

where v is the number of vertices, e edges, f faces and t tetrahedra in the glued up complex.

Since faces are identified in pairs, $f = 2t$. Moreover, each edge contributes 2 vertices, each face 3 edges, each tetrahedron 4 faces to the union of links of vertices. It follows that

$$\sum_{v_i} \chi(\text{link}(v_i)) = 2e - 3f + 4t$$

Note that this sum is equal to $2v$ iff all links of vertices are spheres, since $\chi(S) \leq 2$ for every surface S, with equality iff S is a sphere. Hence $2v - 2e + 3f - 4t = 0$ iff all links are spheres. Equivalently, $v - e + 1\frac{1}{2}f - 2t = 0$. Since $f = 2t$, this sum is equal to $v - e + f - t = \chi(\Sigma)$, and the lemma is proved. □

If M is a 3–manifold with boundary ∂M a (possibly disconnected) surface other than a sphere, one can often think of the interior of M as given by a simplicial complex minus some (or all) of its vertices. The links of removed vertices correspond to removed boundary components of M, and are thus not required to be spheres. If all boundary components of M are tori, a simplicial complex Σ with all vertices removed and all links of vertices tori, whose underlying topological space is homeomorphic to the interior of M, is called an *ideal triangulation* of M.

4.1.4 Heegaard decompositions

If one thinks of the standard embedding $\Sigma_g \to \mathbb{R}^3$ of the surface of genus g in $\mathbb{R}^3$, the bounded component of $\mathbb{R}^3 - \Sigma_g$ is a standard 3–manifold called the *handlebody of genus g*. If Γ is a finite graph embedded in some 3–manifold M, then a thickened neighborhood $N(\Gamma)$ is homeomorphic to a handlebody of some genus.

Definition 4.1.30. A finite graph $\Gamma \subset H$ is called a *standard graph* if $H - \Gamma$ is homeomorphic to $\partial H \times [0, 1)$.

A handlebody admits many combinatorially distinct standard graphs.

Definition 4.1.31. A system $\mathcal{D}$ of pairwise disjoint discs in a handlebody H with $\mathcal{D} \cap \partial H = \partial\mathcal{D}$ is called a *meridian system* in H if $H - (\partial H \cup \mathcal{D})$ is homeomorphic to an (open) 3–ball.

To any standard graph $\Gamma \subset H$ is associated (up to topological equivalence) a meridian system $\mathcal{D}_\Gamma$. The system is obtained by first fixing a spanning tree T for Γ and then choosing the disk system dual to the edges of $\Gamma - T$. Conversely, a disk system gives rise to a standard graph by taking the midpoints of the disks and coning them to the midpoint of the complementary ball.

Definition 4.1.32. Let H_1, H_2 be two handlebodies of genus g. A description of a 3–manifold $M = H_1 \cup H_2$ where $H_1 \cap H_2 = \partial H_1 = \partial H_2$ is called a *Heegaard decomposition* of M. i.e. a Heegaard decomposition of M is a decomposition of the manifold into two handlebodies (necessarily of the same genus).

Theorem 4.1.33 (Heegaard). *Let M be a closed, oriented 3–manifold. Then M admits a Heegaard decomposition.*

Proof. Triangulate $M = \cup_i \Delta_i$. Let H_1 be a thickened neighborhood of the 1–skeleton. Then H_1 is a handlebody. The complement is a thickened neigborhood of the 1–skeleton of the dual cell decomposition, and therefore also a handlebody. □

We can think of a handlebody as being composed of neighborhoods of the disks in a meridian system, together with the complementary 3–ball. Gluing H_2 to H_1 to make a 3–manifold M can be done in two steps — gluing on the disks, and then attaching the 3–ball. If H_2 is of genus g, a meridian system $\mathcal{D}$ contains g disks. The boundaries of the disks $D_i \in \mathcal{D}$ are g circles $\gamma_i \subset \partial H_2$. Specifying the image of these g circles in ∂H_1 determines how the D_i are attached to H_1 in M.

Topologically, a set $\{\gamma_i\}$ of circles in ∂H_1 is the image of the boundaries of a meridian system of an attaching handlebody H_2 iff the circles satisfy the following conditions:

- the circles are pairwise disjoint and embedded
- the number of circles in the collection is equal to the genus of the handlebody
- the complement of the circles is connected

These conditions ensure that if we build a 3–manifold

$$M' = H_1 \cup_i N(D_i)$$

where the D_i are glued along γ_i, and thickened slightly, then $\partial M' = S^2$. It follows that we can attach the remaining 3–ball to this boundary sphere to make a closed 3–manifold M; manifestly, M has a handlebody decomposition into H_1 and its complement. To see that M is well–defined it suffices to show that there is a unique way to attach an n–disk D^n along its boundary; this is sometimes known as *Alexander's trick*.

Lemma 4.1.34 (Alexander's trick). *Let X be a topological space, and $S \subset X$ an embedded sphere of dimension $n-1$. Let ϕ_1, ϕ_2 be two homeomorphisms from ∂D^n to S. Then the result of the glueings $X \cup_{\phi_1} D^n$ and $X \cup_{\phi_2} D^n$ are homeomorphic.*

Proof. Let

$$\psi = \phi_2 \phi_1^{-1} : \partial D^n \to \partial D^n$$

Then ψ can be extended to a homeomorphism $\Psi : D^n \to D^n$ by the formula $\Psi(r, \theta) = (r, \psi(\theta))$ where we use radial co–ordinates (r, θ) on D^n, where $\theta \in S^{n-1}$.

Then we have a homeomorphism

$$\mathrm{id} \cup \Psi : X \cup_{\phi_1} D^n \to X \cup_{\phi_2} D^n$$

□

With $n = 2$, this shows that the disks are attached in an essentially unique way, given their image. With $n = 3$, this shows that the ball is at the end is attached uniquely. It follows that any closed oriented 3–manifold M can be described uniquely by a collection of g embedded circles in the boundary of a genus g handlebody H whose complement in ∂H is connected. Such a presentation is called a *Heegaard diagram* for M.

4.1.5 Knots, Links and Surgery

The idea of surgery — building up a manifold by balls of the appropriate dimension, attached along canonical pieces of their boundary — is the foundation of the theory of high dimensional manifolds.

Definition 4.1.35. An *n–dimensional k–handle* is an n–ball D^n thought of as a product $D^k \times D^{n-k}$. It's boundary has a natural decomposition

$$\partial D^k \times D^{n-k} = S^{k-1} \times D^{n-k} \cup D^k \times S^{n-k-1}$$

The subset $S^{k-1} \times D^{n-k}$ is called the *attaching boundary* of the k–handle, and the sphere $S^{k-1} \times 0$ is called the *core* of the attaching boundary.

If W is an n–manifold with boundary, and $S \subset \partial W$ is a $n-k-1$–sphere, we can attach a k–handle to ∂W so that the core of the attaching boundary of the handle goes along S.

With this terminology, we can think of a Heegaard decomposition as giving a decomposition of a 3–manifold into 3 dimensional handles: the first handlebody H_1 can be built from a 0–handle (a solid ball) and g 1–handles. Then g 2–handles are attached, corresponding to a meridian system for H_2, and finally the result is "capped off" with a 3–handle.

Framings

In general, we need to know how to identify the attaching boundary $S^{k-1} \times D^{n-k}$ with the neighborhood of S in ∂W; such an identification is called a *framing* of S. The different choices of framings are given by the different trivializations of this neighborhood.

For $k = 1$ and $n = 4$, we are attaching a $D^2 \times D^2$ along a solid torus $S^1 \times D^2$ whose core goes around a knot K in the 3–manifold ∂W. A framing of K amounts to specifying how many times the circle $S^1 \times p$ in the attaching boundary, for some $p \neq 0$ wraps around the core $S^1 \times 0$ when it is identified with a neighborhood $N(K) \subset \partial W$.

The boundary $\partial N(K)$ is a torus. There is a natural choice of meridian for this torus — a small circle which bounds a disk D intersecting the knot in a single point. A *longitude* for $\partial N(K)$ is any curve l which intersects the meridian m transversely in a single point. It is easy to see that there are $\mathbb{Z}$ choices for the longitude: if l is some choice of longitude, $l + nm$ is another choice of longitude; here $l + nm$ denotes the circle which follows l around once, then winds n times around the meridian m before joining up. It is the image $t_m^n(l)$ of l under the nth power of a Dehn twist in the curve m. We usually orient the meridian and longitude so that at the crossing c , the tangent vectors (m', l') are a positively oriented basis for the tangent space $T_c\partial N(K)$.

In this language, a framing for a knot amounts to a choice of a longitude for $\partial N(K)$.

Knot and Link projections

A *knot* in a 3–manifold is an embedded circle. Usually we are interested in *tame knots* — i.e. those embedded circles which have a neighborhood homeomorphic to an embedded solid torus. A tame knot can also be thought of as a polygonal embedded circle with respect to the unique PL structure on the manifold. A knot in $\mathbb{S}^3$ can be thought of as sitting in $\mathbb{R}^3$, by removing a point from the complement of the knot. Since the complement is connected, any point will do. Imagine a far away light source, whose rays are almost parallel when they reach the knot: the (thickened) knot casts a shadow on a horizontal plane in $\mathbb{R}^3$. If the knot is in general position, this shadow is an immersed circle in $\mathbb{R}^2$ whose singularities occur when the circle crosses itself transversely at finitely many places. If we record at each such crossing the relative position of the two knot segments in $\mathbb{R}^3$ — i.e. which one is closer, and which further away from the light source — we can reconstruct the embedding of the knot in $\mathbb{R}^3$ from this picture. Such a picture — an immersed circle in $\mathbb{R}^2$ with resolutions of crossings into over and under — is called a *knot diagram.* Two such knot diagrams describe equivalent knots iff they are related by a sequence of *Reidemeister moves*, which describe the topological changes in the picture when the knot moves from one general position to another by a generic isotopy. These moves are illustrated below.

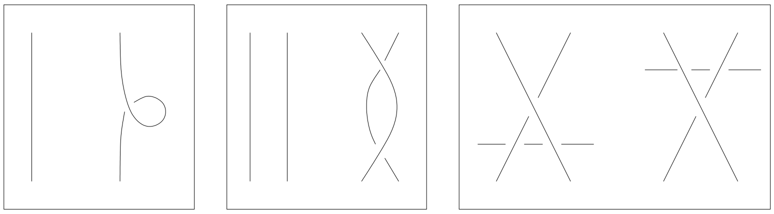

Figure 4.1: Two knot diagrams represent the same topological knot iff they are related by a sequence of Reidemeister moves.

For more details and a proof, consult [20].

If $M^3 = \partial W^4$ and $K \subset M$ is a knot, adding a 2–handle to W along $N(K) \subset M$ has the effect on M of removing a solid torus neighborhood of K, and gluing in a new solid torus along the resulting boundary. The meridian of the glued in solid torus goes once around the longitude of $\partial N(K)$ and some number of times around the meridian of $\partial N(K)$. The exact number of times it winds around the meridian depends exactly on the choice of framing of K.

Seifert surfaces, linking number and natural framings

It is convenient to have a canonical choice of "base" framing for a knot, so that other framings can be calculated relative to it. Such a base framing is provided by a geometric object called a *Seifert surface* for a knot.

Definition 4.1.36. Let $K \subset \mathbb{S}^3$ be a knot. A *Seifert surface* for K is an orientable surface Σ properly embedded in $M - K$ such that $\partial\Sigma = K$.

A Seifert surface is thus a surface in the knot complement whose boundary is the knot. For instance, if K is the unknot, a Seifert surface is a spanning disk for K.

Definition 4.1.37. Let K be a knot. The *natural longitude* on $\partial N(K)$ is the (isotopy class of) intersection $\Sigma \cap \partial N(K)$ where Σ is a Seifert surface for K.

This definition should raise many questions. Why does a Seifert surface exist for a knot? Why should two different Seifert surfaces give rise to topologically equivalent longitudes? The reasons for these facts are related to the topology of $\mathbb{S}^3$; for a knot in an arbitrary 3–manifold M, there are no "naturally" defined longitudes.

Definition 4.1.38. A *link* in a 3–manifold is a disjoint collection of knots.

We usually require our links to be tame. A link gives rise to a link diagram, much like a knot diagram. Two link diagrams represent the same topological link iff they are related by a sequence of Reidemeister moves.

If L is a link with two components K_1, K_2, we can define the *linking number* of K_1 with K_2. This definition requires a choice of orientation on each of K_1, K_2, and is defined as follows

Definition 4.1.39. Let v be a non–zero vector at some point in $\mathbb{R}^3$. The unit length vector parallel to v centered at the origin can be identified with an element of the unit sphere, $g(v) \in S^2$. This defines the *Gauss map*

$$g : T\mathbb{R}^3 - 0 \to S^2$$

Note that 0 denotes the subspace of $T\mathbb{R}^3$ consisting of zero vectors based at any point.

Let $T = S^1 \times S^1$ be parameterized by points on the knots K_1, K_2. Define $g : T \to S^2$ by sending the point (θ, ϕ) to the image of the Gauss map of the vector from the point $\theta \in K_1$ to $\phi \in K_2$.

The *linking number* of K_1 with K_2 is defined to be the degree of the map g; that is, the number of points (counted with sign) in the preimage of a generic point $s \in S^2$.

An easy way to count this number is by looking at a link projection. If s is the south pole of S^2, the preimages of s correspond to the pairs of points $p \in K_1, q \in K_2$ such that p is vertically above q — that is, such that the knot K_1 crosses over the knot K_2 in the projection. So we can define the linking number to be the sum over the undercrossings of K_2 under K_1 in a diagram, of either ± 1, where the count is determined by the following figure:

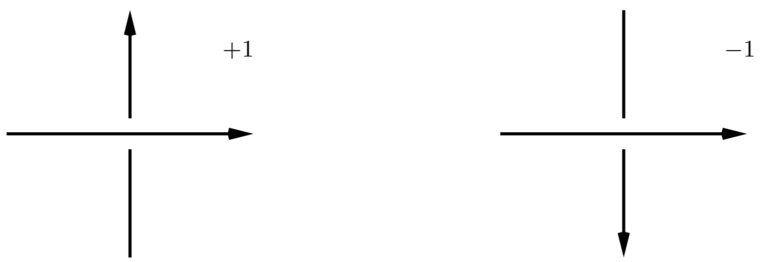

Figure 4.2: The linking number of K_1 and K_2 is the sum over undercrossings of K_2 under K_1 of the sign of the crossing, defined as in this figure.

Now we may define the *natural longitude* of a knot K as the choice of longitude on $\partial N(K)$ which has linking number 0 with K.

Surgery diagrams

Definition 4.1.40. A *surgery diagram* is a link $L = K_1 \cup K_2 \cup \cdots \cup K_n$ in $\mathbb{S}^3$, where each component K_i may or may not be decorated with an element r_i of $\mathbb{Q} \cup \infty$. Such a diagram represents a 3–manifold, possibly with some number of torus boundary components, defined by surgering the decorated components by (p_i, q_i) surgery with respect to a natural meridian and longitude, where $r_i = p_i/q_i$ is a representation of the rational number r_i as a ratio of coprime integers, and $(p_i, q_i) = (1, 0)$ if $r_i = \infty$.

Surgery diagrams are certainly not unique: a fixed 3–manifold admits infinitely many representations by a surgery diagram. Moreover, there is no known algorithm which will decide if two diagrams represent the same manifold; in fact, this seems to be a difficult question to decide in practice, as well as in theory.

There are some standard "moves" on diagrams which generate equivalent manifolds; these basic moves always suffice to move between two different diagrams of the same manifold.

Example 4.1.41. 1 surgery on the figure 8 knot complement is equivalent to -1 surgery on the right–handed trefoil:

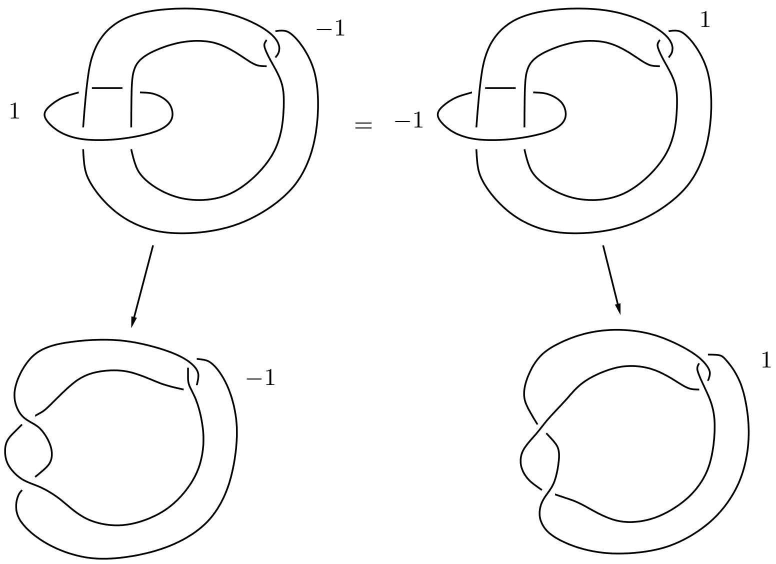

Figure 4.3: This sequence of moves demonstrates the equivalence of the two manifolds obtained by 1 and -1 surgery on the figure 8 knot and right–handed trefoil respectively

These moves are as follows:

1. *Twist an unknotted component.*

   If $K \subset L$ is an unknot, it bounds a disk D which can be chosen to be transverse to $L - K$. The manifold obtained from $\mathbb{S}^3 - K$ by cutting along D, twisting through 2π in the anticlockwise direction and then regluing is obviously topologically still $\mathbb{S}^3 - K$, but in terms of the diagram, we have "twisted" the components of L passing through D through one full twist. This has the effect of changing the framings on these components as follows: if K' is a component that passes through D, orient K' and D, and let u, d be the number of times K' passes through D is the positive (resp. negative) direction. Then the framing on K' changes by $(u - d)^2$. This framing change is obviously independent of the orientation of K' and D. Moreover, this twist changes the framing of the boundary of a neighborhood of K. In particular, the natural choice of meridian and longitude for K change by $(m, l) \to (m + l, l)$, so if K is decorated with a rational number, this number must be suitably modified to reflect the change of basis.

2. *Add or remove a knot with framing ∞*

   Doing ∞ surgery on a knot corresponds to removing a solid torus neighborhood of a knot, and replacing it so that the meridian of the added torus goes to the meridian of the knot. In other words, this operation does nothing to the manifold.

These moves suffice to generate all equivalences between manifolds described by surgery diagrams.

As explained above, *integral* surgery on a component of a link in a 3–manifold M corresponds to adding a 2–handle to a 4 manifold W with boundary M. If K_1, K_2 with framings n_1, n_2 are the cores of the attaching handles, we can imagine attaching 2–handles in some particular order — first along K_1, and then along K_2. Let W_1 be the result of adding a 2–handle to W along K_1. Isotopic attaching maps give rise to homeomorphic manifolds, so we can "slide" K_2 around in $M_1 = \partial W_1$ without changing the end result of surgery. If we slide K_2 over the 2–handle used to attach K_1, the resulting knot K_2' will not be isotopic *in* $M - K_1$ (except under very special circumstances) to K_2. The effect of such a slide is illustrated below.

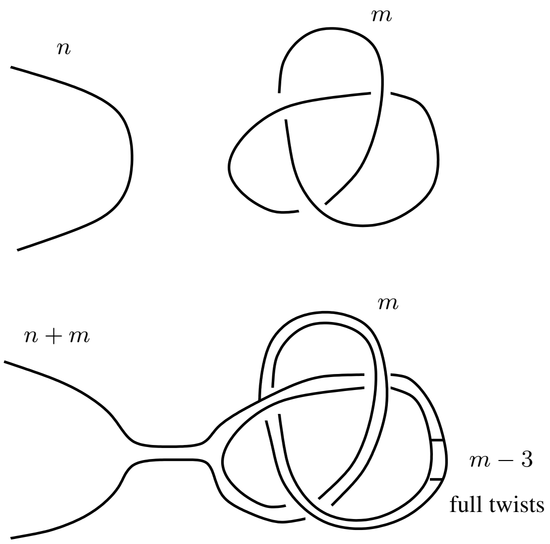

Figure 4.4: The result of sliding K_2 with framing n over K_1 with framing m, both integers

Surgery diagrams with integer framings can equally well represent 3–manifolds or the 4–manifolds that they bound. In the 4–dimensional context, the rules for equivalence of diagrams is called *Kirby Calculus*. A brief exposition can be found in [17]. A much more detailed and thorough account is found in [14].

4.2 Tessellations

4.2.1 Discrete groups

Definition 4.2.1. Let Γ be a group of symmetries of a space X which is one of $\mathbb{S}^n, \mathbb{E}^n, \mathbb{H}^n$ for some n. Γ is *properly discontinuous* if, for each closed and bounded subset K of X, the set of $\gamma \in \Gamma$ such that $\gamma(K) \cap K \neq \emptyset$ is finite. Γ acts *freely* if no $\gamma \in \Gamma$ has a fixed point; that is, if $\gamma(p) = p$ for any p, then $\gamma = \text{id}$.

For a point $p \in X$, the subgroup of Γ which fixes p is called the *stabilizer* of p, and is typically denoted by $\Gamma(p)$. If Γ acts properly discontinuously, then $\Gamma(p)$ is finite for any p.

Definition 4.2.2. A subgroup Γ of a Lie group G is *discrete* if $K \cap \Gamma$ is finite for any compact subset $K \subset G$.

Exercise 4.2.3. *Show that if G is a Lie group of symmetries of a space X as above, then a subgroup Γ is discrete iff it acts properly discontinuously on X.*

Suppose Γ acts on X freely and properly discontinuously. We can define a quotient space $M = X/\Gamma$ where two points $x, y \in X$ are identified exactly when there is some $\gamma \in \Gamma$ such that $\gamma(x) = y$.

Theorem 4.2.4. *If Γ acts on X freely and properly discontinuously, where X is one of $\mathbb{S}^n, \mathbb{E}^n, \mathbb{H}^n$ for some n, then the projection $X \to X/\Gamma$ is a covering space, and X is the universal cover of X/Γ.*

Proof. Pick a point $x \in X$ and let U be a neighborhood of x which intersects only finitely many translates $\gamma(U)$. Then there is a smaller $V \subset U$ which is disjoint from all its translates. Then under the quotient map, each translate $\gamma(V)$ is mapped homeomorphically to its image. Thus X is a covering space of X/Γ, and since it is simply–connected, it is the universal cover. $\square$

Fundamental domains

Let Γ act on X properly discontinuously, where X is one of the spaces $\mathbb{S}^n, \mathbb{E}^n, \mathbb{H}^n$.

Definition 4.2.5. A *fundamental domain* for the action of Γ is a polygon $P \subset X$ such that for all p in the interior of P, $\alpha(p) \cap P = \emptyset$ unless $\alpha = \text{id}$, and such that the faces of P are paired by the action of Γ.

The translates of a fundamental domain are disjoint except along their boundaries. Moreover, these translates cover all of X. Thus they give a tessellation of X, whose symmetry group contains Γ as a subgroup. For "generic" fundamental domains, there are no "accidental" symmetries, and the group of symmetries of the tessellation is exactly Γ. In general, fundamental domains can be *decorated* with some extra marking which destroys any additional extra symmetries of a fundamental domain.

Definition 4.2.6. Choose a point $p \in X$. The *Dirichlet domain of* Γ *centered at* p, is the set

$$D = \{q \in X \text{ such that } d(p,q) \leq d(p, \alpha(q)) \text{ for all } \alpha \in \Gamma\}$$

In dimension $2, 3$, in the cases we are interested in, D will be a locally finite polygon in X whose faces are paired by the action of Γ; but in general, D might not be a polygon.

A Dirichlet domain is a fundamental domain for Γ.

4.2.2 Lattices in $\mathbb{E}^2$

Discrete groups in Isom$(\mathbb{E}^2)$

Perhaps the most important theorem about discrete subgroups of Isom$(\mathbb{E}^n)$ is Bieberbach's theorem, that such a group has a free abelian subgroup of finite index. In dimension two, this can be refined as follows:

Theorem 4.2.7. *Let Γ act properly discontinuously on $\mathbb{E}^2$ by isometries. Then Γ has a subgroup Γ^+ of index at most 2 which is orientation–preserving. Moreover, Γ^+ is one of the following:*

- *Γ^+ is a subgroup of stab(p) for some p. In this case, $\Gamma^+ \cong \mathbb{Z}/n\mathbb{Z}$ consists of powers of a single rotation.*
- *Γ^+ is a semi–direct product*
$$\Gamma^+ = \mathbb{Z}/n\mathbb{Z} \ltimes \mathbb{Z}$$
where the $\mathbb{Z}$ factor is generated by a translation, and n is 1 or 2. In the second case, the conjugation action takes $x \to -x$.
- *Γ^+ is a semi–direct product*
$$\Gamma^+ = \mathbb{Z}/n\mathbb{Z} \ltimes (\mathbb{Z} \oplus \mathbb{Z})$$
where the $\mathbb{Z}\oplus\mathbb{Z}$ factor is generated by a pair of linearly indpendent translations, and $n = 1, 2, 3, 4$ or 6. If $n = 4$, the $\mathbb{Z}\oplus\mathbb{Z}$ is conjugate to the group of translations of the form $z \to z + n + mi$ for integers n, m. If $n = 3$ or 6, the $\mathbb{Z}\oplus\mathbb{Z}$ is conjugate to the group of translations of the form $z \to z+n+m\frac{1+i\sqrt{3}}{2}$ for integers n, m.

Proof. First, there is a homomorphism
$$o : \text{Isom}(\mathbb{E}^2) \to \mathbb{Z}/2\mathbb{Z}$$
where $o(\alpha)$ is 0 or 1 depending on whether or not α is orientation preserving. The intersection of Γ with the kernel of o is Γ^+, and it has index at most 2. Next, there is a homomorphism
$$a : \text{Isom}^+(\mathbb{E}^2) \to S^1$$
given by the action of isometries on equivalence classes of parallel lines, where the image is the angles of rotation. There is an induced homomorphism
$$a : \Gamma^+ \to S^1$$

The kernel K of a in Γ^+ consists of a group of translations, and is therefore abelian.

Suppose $\theta = a(\alpha)$ for some $\alpha \in \Gamma^+$, and θ nonzero. Then α is a rotation, and therefore fixes some p. Since Γ^+ is properly discontinuous, either $\Gamma^+ \subset \text{stab}(p)$ in which case Γ^+ is cyclic and consists of the powers of a fixed rotation, or there is a (non–unique) closest image $q \neq p$ of p under some β. Since $q = \beta(p)$, no translate of p is closer to q than p. If $|\theta| < \frac{2\pi}{6}$, then

$$0 < d(q, \alpha(q)) = 2\sin\Big(\frac{\theta}{2}\Big) d(p,q) < d(p,q)$$

which is a contradiction. Since the same must also be true for each power of α, the order of α is ≤ 6. If the order is 5, then $0 < d(\alpha(q), \beta\alpha^{-1}\beta^{-1}(p)) < d(p,q)$, which is a contradiction. Hence the image $a(\Gamma^+)$ is $\mathbb{Z}/n\mathbb{Z}$ where $n = 1, 2, 3, 4$ or 6. In any case, the image is cyclic, and is therefore generated by a single $a(\alpha)$ where α is a rotation, so Γ^+ is a semi–direct product.

The kernel K of a in Γ^+ is a properly discontinuous group of translations. If K is nontrivial and the elements of K are linearly dependent, they are all powers of the element α of shortest translation length. The image $a(\Gamma^+)$ must preserve the translation of α; in particular, this image must be trivial or $\mathbb{Z}/2\mathbb{Z}$.

If K is nontrivial and the elements of K are linearly independent, they are generated by two elements of shortest translation length. The image $a(\Gamma^+)$ must preserve the set of nonzero elements of shortest length, so if this image is $\mathbb{Z}/4\mathbb{Z}$ the group K is generated by two perpendicular vectors of equal length. If the image is $\mathbb{Z}/3\mathbb{Z}$ or $\mathbb{Z}/6\mathbb{Z}$, K is generated by two vectors at angle $\frac{2\pi}{6}$ of equal length. □

The two special lattices (i.e. groups of translations generated by two linearly independent elements) appearing in theorem 4.2.7 are often called the *square* and *hexagonal* lattices respectively.

Integral quadratic forms

A good reference for the material in this and the next section is [2].

Definition 4.2.8. A *quadratic form* is a homogeneous polyonomial of degree 2 in some variables. That is, a function of variables $x_1, \ldots, x_n$ such that

$$f(\lambda_1 x_1, \ldots, \lambda_n x_n) = \prod_{i=1}^{n} \lambda_i^2 f(x_1, \ldots, x_n)$$

A quadratic form is *integral* if its coefficients as a polynomial are integers.

We will be concerned in the sequel with integral quadratic forms of two variables, such as $3x^2 + 2xy - 7y^2$ or $-x^2 - 3xy$.

Notice for every quadratic form $f(\cdot,\cdot)$ there corresponds uniquely a symmetric matrix M_f such that $f(x,y) = \begin{bmatrix} x & y \end{bmatrix} M_f \begin{bmatrix} x \\ y \end{bmatrix}$. In particular, if $f(x,y) = ax^2 + bxy + cy^2$ then

$$M_f = \begin{bmatrix} a & \frac{b}{2} \\ \frac{b}{2} & c \end{bmatrix}$$

Definition 4.2.9. We say that a quadratic form $f(x, y)$ *represents* an integer n if there is some assignment of integer values n_1, n_2 to x, y for which

$$f(n_1, n_2) = n$$

Given an integral quadratic form, it is a natural question to ask what interger values it represents. Observe that there are some elementary transformations on quadratic forms which do not change the set of values represented. If we substitute $x \to x \pm y$ or $y \to y \pm x$ we get a new quadratic form. Since this substitution is *invertible*, the new quadratic form obtained represents exactly the same set of values as the original. For instance,

$$x^2 + y^2 \longleftrightarrow x^2 + 2xy + 2y^2$$

This substitution defines an equivalence relation on quadratic forms.

The most general form of substitution allowed is a transformation of the form

$$x \to px + qy, \ y \to rx + sy$$

For integers p, q, r, s. Again, since this substitution must be invertible, we should have $ps - qr = \pm 1$. That is, the matrix $\begin{bmatrix} p & q \\ r & s \end{bmatrix}$ is in $\pm SL(2, \mathbb{Z})$. Since these forms are homogeneous of degree 2, the substitution $x \to -x, \ y \to -y$ does nothing. One can easily check that every other substitution has a nontrivial effect on some quadratic form.

In particular, we have the following theorem:

Theorem 4.2.10. *Integral quadratic forms up to equivalence are parameterized by equivalence classes of matrices of the form* $\begin{bmatrix} a & \frac{b}{2} \\ \frac{b}{2} & c \end{bmatrix}$ *for integers* a, b, c *modulo the conjugation action of* $PSL(2, \mathbb{Z})$.

Observe that what is really going on here is that we are evaluating the quadratic form f on the integral lattice $\mathbb{Z} \oplus \mathbb{Z}$. The group $SL(2, \mathbb{Z})$ acts by automorphisms of this lattice, and therefore permutes the set of values attained by the form f.

There is nothing special about the integral lattice here; it is obvious that $SL(2, \mathbb{Z})$ acts by automorphisms of *any* lattice L. In particular, if $L = \langle e_1, e_2 \rangle$ then $M = \begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix} \in SL(2, \mathbb{Z})$ acts on elements of this lattice by

$$M \cdot re_1 + se_2 \to (\alpha r + \beta s)e_1 + (\gamma r + \delta s)e_2$$

Moduli of tori, continued fractions and $PSL(2, \mathbb{Z})$

Definition 4.2.11. Let r be a real number. A *continued fraction expansion* of r is an expression of r as a limit of a (possibly terminating) sequence

$$n_1, n_1 + \frac{1}{m_1}, n_1 + \frac{1}{m_1 + \frac{1}{n_2}}, n_1 + \frac{1}{m_1 + \frac{1}{n_2 + \frac{1}{m_2}}}, \ldots$$

where each of the n_i, m_i is a positive integer.

A continued fraction expansion of r can be obtained inductively by Euclid's algorithm. First, n_1 is the biggest integer $\leq r$. So $0 \leq r - n_1 < 1$. If $r - n_1 = 0$ we are done. Otherwise, $r' = \frac{1}{r-n_1} > 1$ and we can define m_1 as the biggest integer $\leq r'$. So $0 \leq r' - m_1 < 1$. continuing inductively, we produce a series of integers $n_1, m_1, n_2, m_2, \ldots$ which is the *continued fraction expansion of* r. If r is rational, this procedure terminates at a finite stage. The usual notation for the continued fraction expansion of a real number r is

$$r = n_1 + \frac{1}{m_1+}\frac{1}{n_2+}\frac{1}{m_2+}\frac{1}{n_3+}\ldots$$

The following theorem is quite easy to verify:

Theorem 4.2.12. *If $n_1, m_1, \ldots$ is a continued fraction expansion of r, then the successive approximations*

$$n_1, n_1 + \frac{1}{m_1}, n_1 + \frac{1}{m_1 + \frac{1}{n_2}}, \ldots$$

denoted $r_1, r_2, r_3, \ldots$ satisfy

$$|r - r_i| \leq |r - p/q|$$

for any integers (p, q), where $q <$ the denominator of r_{i+1}.

Thus, the continued fraction approximations of r are the best rational approximations to r for a given bound on the denominator.

Let T be a flat torus. Then the isometry group of T is transitive (this is not too hard to show). Pick a point p, and cut T up along the two shortest simple closed curves which start and end at p. This produces a Euclidean parallelogram P. After rescaling T, we can assume that the shortest side has length 1. We place P in $\mathbb{E}^2$ so that this short side is the segment from 0 to 1, and the other side runs from 0 to z where Im$(z) > 0$. By hypothesis, $|z| \geq 1$. Moreover, if $|\text{Re}(z)| \geq \frac{1}{2}$ we can replace z by $z+1$ or $z-1$ with smaller norm, contradicting the choice of curves used for the decomposition. Let D be the region in the upper half–plane bounded by the two vertical lines Re$(z) = \frac{1}{2}$, Re$(z) = -\frac{1}{2}$ and the circle $|z| = 1$.

The group $PSL(2, \mathbb{Z})$ acts naturally on $\mathbb{H}$ as a subgroup of $PSL(2, \mathbb{R})$. The action there is properly discontinuous. The action of $PSL(2, \mathbb{Z})$ permutes the sides of D. The element $\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ pairs the two vertical sides, and the element $\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$ preserves the bottom side, interchanging the left and right pieces of it. In particular, D is a fundamental domain for $PSL(2, \mathbb{Z})$. The quotient is topologically a disk, but with two "cone points" of order 2 and 3 respectively, which correspond to the points i and $\frac{1+i\sqrt{3}}{2}$ respectively, whose stabilizers are $\mathbb{Z}/2\mathbb{Z}$ and $\mathbb{Z}/3\mathbb{Z}$ respectively.

This quotient is an example of an *orbifold*.

Definition 4.2.13. An n–dimensional *orbifold* is a space which is locally modelled on $\mathbb{R}^n$ modulo some finite group.

A 2–dimensional orbifold looks like a surface except at a collection of isolated points p_i where it looks like the quotient of a disk by the action of $\mathbb{Z}/n_i\mathbb{Z}$, a group of rotations centered at p_i. The point p_i is a *cone point*, sometimes also called an *orbifold point*. The finite group is part of the data of the orbifold. One can think of the orbifold combinatorially as a surface (in the usual sense) with a finite number of distinguished points, each of which has an integer attached to it. Geometrically, this point looks like a "cone" made from a wedge of angle $2\pi/n_i$.

We can define an *orbifold fundamental group* $\pi_1^o(\cdot)$ for a surface orbifold. Thinking of our orbifold Σ as X/Γ for the moment where Γ acts properly discontinuously but not freely, the orbifold fundamental group of Σ should be exactly Γ. This means that a small loop around an orbifold point p_i should have order n_i in $\pi_1^o(\Sigma)$. Note that we are being casual about basepoints here, so we are only thinking of these groups up to isomorphism.

In any case, the orbifold $\mathbb{H}^2/PSL(2,\mathbb{Z})$ should have orbifold fundamental group isomorphic to $PSL(2,\mathbb{Z})$. There is an element of order 2 corresponding to the loop around the order 2 point; a representative of this element in $PSL(2,\mathbb{Z})$ is $\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$. There is an element of order 3 corresponding to the loop around the order 3 point; a representative of this element in $PSL(2,\mathbb{Z})$ is $\begin{bmatrix} 1 & 1 \\ -1 & 0 \end{bmatrix}$. Note that these elements have order 2 and 3 respectively in $PSL(2,\mathbb{Z})$, even though the corresponding matrices have orders 4 and 6 respectively in $SL(2,\mathbb{Z})$.

Every loop in the disk can be shrunk to a point; it follows that every loop in the orbifold $\mathbb{H}^2/PSL(2,\mathbb{Z})$ can be shrunk down to a collection of small loops around the two cone points in some order. That is, the group $PSL(2,\mathbb{Z})$ is generated by these two elements.

Theorem 4.2.14. *A presentation for $PSL(2,\mathbb{Z})$ is given by*

$$PSL(2,\mathbb{Z}) \cong \langle \alpha, \beta | \alpha^2, \beta^3 \rangle = \mathbb{Z}/2\mathbb{Z} * \mathbb{Z}/3\mathbb{Z}$$

where representatives of α and β are the two matrices given above.

Proof. By the discussion above, all that needs to be established is that there are no other relations that do not follow from the relations $\alpha^2 = \mathrm{id}$ and $\beta^3 = \mathrm{id}$. That is, there is a homomorphism $\phi : G \to PSL(2,\mathbb{Z})$ sending α, β to the two matrices given; all we need to check is that the kernel of this homomorphism consists of the identity element.

A general element of $G = \langle \alpha, \beta | \alpha^2, \beta^3 \rangle$ is a product $\alpha^{a_1}\beta^{b_1}\alpha^{a_2}\beta^{b_2}\dots\alpha^{a_n}\beta^{b_n}$ where each of the a_i, b_i are integers. We reduce the a_i mod 2 and the b_i mod 3; after rewriting of this kind, we are left with a product of the form above where every $a_i = 1$ and every b_i is 1 or 2. We write $L = \alpha\beta$ and $R = \alpha\beta^2$, so that every nontrivial element of G is of the form $w, \beta^{\pm 1}w, w\alpha, \beta^{\pm 1}w\alpha$ where w is a word in the letters L and R. Furthermore, we have the relation $(\alpha\beta\alpha\beta^2)^3$

We show that no word w in the letters L and R is trivial in the group $PSL(2,\mathbb{Z})$. A similar argument works for elements of G of the other forms.

Now, $\phi(L) = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$ and $\phi(R) = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$ in $PSL(2, \mathbb{Z})$. Suppose that

$$w = L^{m_1} R^{n_1} L^{m_2} R^{n_2} \dots L^{m_k} R^{n_k}$$

where all the m_i, n_i are nonzero, say. Then we can calculate

$$\phi(w) = \begin{bmatrix} p & r \\ q & s \end{bmatrix}$$

where

$$\frac{p}{q} = \frac{1}{m_1+} \frac{1}{n_1+} \frac{1}{m_2+} \frac{1}{n_2+} \dots \frac{1}{m_k}$$

$$\frac{r}{s} = \frac{1}{m_1+} \frac{1}{n_1+} \frac{1}{m_2+} \frac{1}{n_2+} \dots \frac{1}{m_k+} \frac{1}{n_k}$$

where the notation is for a continued fraction expansion. That is, the alternating coefficients m_i, n_i give the continued fraction expansions of $\frac{p}{q}$ and $\frac{r}{s}$. In particular, $\phi(w) \neq \mathrm{Id}$ unless w is the empty word, and ϕ is an isomorphism. $\square$

Notice actually that this method of proof does considerably more. We have shown that every element obtained by a product of *positive* multiples of L and R is non–trivial. It is not true that the *group* generated by L and R is free, since LR^{-1} has order 3. In fact, L and R together generate the entire group $PSL(2, \mathbb{Z})$. But the group generated by L^2 and R^2 is free, since a fundamental domain for its action is the domain D' bounded by the lines

$$\mathrm{Re}(z) = 1, \ \mathrm{Re}(z) = -1$$

and the semicircles

$$|z - 1/2| = 1/2, \ |z + 1/2| = 1/2$$

Let Γ denote this subgroup of $PSL(2, \mathbb{Z})$. The domain D' is a regular ideal hyperbolic quadrilateral; the quotient $\mathbb{H}^2/\Gamma$ is topologically a thrice punctured sphere. By the argument of the previous section, a presentation is

$$\pi_1(\text{thrice punctured sphere}) \cong \langle \alpha, \beta | \ \rangle$$

That is, $\Gamma \cong \mathbb{Z} * \mathbb{Z}$.

Another description of Γ is the following: there is an obvious homomorphism $\psi : PSL(2, \mathbb{Z}) \to PSL(2, \mathbb{Z}/2\mathbb{Z})$ given by reducing the entries mod 2. The image group has order 6 and the surjection is onto, so the kernel is a subgroup of index 6. Since the fundamental domain D' can be made from 6 copies of D, it follows that the index of Γ in $PSL(2, \mathbb{Z})$ is 6. Moreover, Γ is certainly contained in the kernel of ψ. It follows that Γ is exactly equal to this kernel.

Γ is sometimes also denoted by $\Gamma(2)$ (for "reduction mod 2") and is of considerable interest to number theorists, who like to refer to it as the *principal congruence subgroup of level* 2.

Notice that the domain D' is obtained from two ideal triangles. The union of all the translates of D' by $\Gamma(2)$ gives a tessellation of $\mathbb{H}^2$ by regular ideal quadrilaterals; a subdivision of D' into two ideal triangles gives a subdivision of $\mathbb{H}^2$ into ideal triangles. If we choose the subdivision along the line Re$(z) = 0$, the ideal triangulation T of $\mathbb{H}^2$ so obtained admits reflection symmetry along every edge. The 1–skeleton of the dual cell–decomposition to this ideal triangulation is the *infinite* 3*–valent tree*. There is a natural action of $PSL(2, \mathbb{Z})$ on this tree, where the elements of order 3 are the stabilizers of vertices and the elements of order 2 are the stabilizers of edges. This description of $PSL(2, \mathbb{Z})$ as a group of automorphisms of a tree gives another way to see that it is isomorphic to $\mathbb{Z}/2\mathbb{Z} * \mathbb{Z}/3\mathbb{Z}$.

For a rational point p/q we can consider the straight line l perpendicular to the real axis given by Re$(z) = p/q$. As this line l moves from ∞ to p/q it crosses through many different triangles of T, and therefore determines a word w in the letters R, L and their inverses. By induction, it is easy to show that the word w is of the form

$$w = L^{m_1} R^{n_1} L^{m_2} R^{n_2} \dots L^{m_k} R^{n_k}$$

where

$$\frac{p}{q} = \frac{1}{m_1+} \frac{1}{n_1+} \frac{1}{m_2+} \frac{1}{n_2+} \cdots \frac{1}{m_k}$$

An irrational point r determines an infinite word

$$w = L^{m_1} R^{n_1} L^{m_2} R^{n_2} \dots$$

where

$$r = \frac{1}{m_1+} \frac{1}{n_1+} \frac{1}{m_2+} \frac{1}{n_2+} \cdots$$

is an infinite continued fraction expansion of r.

Notice that this word w is *eventually periodic* exactly when r is of the form $a + \sqrt{b}$ for rational numbers a, b.

4.2.3 Finite subgroups of $SO(3)$ and $\mathbb{S}^3$

The "fair dice"

A *die* is a convex 3–dimensional polyhedron. We can ask under what conditions a die is *fair* — that is, the probability that the die will land on a given side is $1/n$ where n is the number of sides. This is a very hard problem to treat in full generality, since it is very hard to calculate these probabilities for a generic polyhedron. But there are certain circumstances under which it is easy to show that these probabilities are all equal; if for any two faces f_1, f_2 of a die D there is a symmetry of D to itself taking f_1 to f_2 then the die is manifestly fair. The group G of all symmetries of D is a subgroup of the group of permutations of the vertices. Any symmetry of D extends to an isometry of $\mathbb{E}^3$, in particular it is an affine map. It follows that if the vertices of D are at the vectors $v_1, v_2, \dots, v_n$ then the images of these vertices under a symmetry σ are the same set of vectors in permuted order. Thus the symmetry fixes the center of gravity of D; as a vector this is $\frac{\sum_{i=1}^n v_i}{n}$.

Translating this center of gravity to the origin in $\mathbb{E}^3$, we see that G is a finite subgroup of $O(3)$. That is, G is a properly discontinuous group of isometries of $\mathbb{S}^2$.

Spherical orbifolds

Of course, any properly discontinuous group Γ of isometries of $\mathbb{S}^2$ has a subgroup Γ^+ of index at most two which consists of orientation–preserving elements. Every orientation–preserving isometry of $\mathbb{S}^2$ has a fixed point, so Γ^+ does not act freely unless it is trivial. In any case, the quotient $\mathbb{S}^2/\Gamma^+$ will be a *spherical orbifold* Σ. This orbifold is topologially a surface with finitely many cone points $p_1, \ldots, p_m$ of orders $n_1, \ldots, n_m$. The Gauss–Bonnet formula gives

$$\text{area}(\Sigma) = \int_\Sigma \kappa = 2\pi \left(\chi(\Sigma) - \sum_{i=1}^m \frac{n_i - 1}{n_i} \right)$$

Since the area is positive, Σ must be topologically a sphere, since that is the only surface with positive Euler characteristic. Each $\frac{n_i-1}{n_i}$ term is at least $1/2$, so it follows that there can be at most 3 cone points. Notice too that if Σ has two cone points, small loops around them are isotopic, and therefore should represent the same element of the orbifold fundamental group; in particular, they should have the same order. Similarly, Σ cannot have a single cone point, since a nontrivial element of the orbifold fundamental group could be shrunk to a point.

We therefore have the following theorem:

Theorem 4.2.15. *Let Γ be a properly discontinuous group of isometries of $\mathbb{S}^2$. Then Γ has a subgroup Γ^+ of index at most 2 which is orientation–preserving. The following are the possibilities for Γ^+:*

- *Γ^+ fixes a pair of antipodal points. $\Gamma^+ \cong \mathbb{Z}/n\mathbb{Z}$ and is generated by a single rotation.*
- *Γ^+ is generated by two rotations r_1, r_2 of order 2 whose axes are at an angle of $\frac{2\pi}{n}$ to each other. The group $\Gamma^+ = \langle r_1, r_2 \rangle$ is the dihedral group D_n.*
- *The quotient $\mathbb{S}^2/\Gamma^+$ is a sphere with 3 cone points of orders $(2,3,3)$, $(2,3,4)$ or $(2,3,5)$. Γ^+ in these cases is the group of orientation–preserving symmetries of the regular tetrahedron, octahedron, and dodecahedron respectively. As a group, Γ^+ is isomorphic to A_4, S_4, A_5 respectively.*

An orbifold Σ whose underlying surface is a sphere with three cone points p_1, p_2, p_3 is called a *triangle orbifold*. For the sake of generality, we can think of a puncture as a "cone point of order ∞", so that $\mathbb{H}^2/PSL(2,\mathbb{Z})$ is the triangle orbifold with cone points of order $(2,3,\infty)$. The triangle orbifold with cone points of order (p,q,r) will be denoted $\Delta(p,q,r)$.

A presentation for the triangle orbifold with cone points (p,q,r) is

$$\pi_1^o(\Delta(p,q,r)) \cong \langle \alpha, \beta | \alpha^p, \beta^q, (\alpha\beta)^r \rangle$$

Lemma 4.2.16. *For $r = 3,4,5$ there is an isomorphism*

$$\pi_1^o(\Delta(2,3,r)) \to PSL(2, \mathbb{Z}/r\mathbb{Z})$$

where in each case, the image of the small loops α, β around the cone points of order $2, 3$ correspond to the equivalence classes of matrices

$$\alpha \to \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \beta \to \begin{bmatrix} 1 & 1 \\ -1 & 0 \end{bmatrix}$$

Proof. There are certainly homomorphisms from $\pi_1^o(\Delta(2,3,r))$ onto $PSL(2,\mathbb{Z}/r\mathbb{Z})$ determined by the maps in question, since the relations $\alpha^2 = \text{id}$ and $\beta^3 = \text{id}$ hold in $PSL(2,\mathbb{Z}/n\mathbb{Z})$ for any n, and $\alpha\beta = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$ which has order r in $PSL(2,\mathbb{Z}/r\mathbb{Z})$.

To see that these maps are injective, observe that the orders of the groups for $r = 3, 4, 5$ are both equal to $12, 24$ and 60 respectively, so the maps are isomorphisms. □

For $r > 5$, the group $\pi_1^o(\Delta(2,3,r))$ is infinite, and therefore cannot be isomorphic to $PSL(2,\mathbb{Z}/r\mathbb{Z})$. But for $r = \infty$ we have seen

$$\pi_1^o(\Delta(2,3,\infty)) \cong PSL(2,\mathbb{Z})$$

The homomorphism $PSL(2,\mathbb{Z}) \to PSL(2,\mathbb{Z}/r\mathbb{Z})$ for $3 \leq r \leq 5$ is induced by the map from $\Delta(2,3,\infty)$ to $\Delta(2,3,r)$ which is the identity away from the special points, sends the order $2, 3$ points to order $2, 3$ points respectively, and "sends the puncture" to the cone point of order r.

Reflection groups, Coxeter diagrams

If P is a polyhedron in X one of $\mathbb{S}^n, \mathbb{E}^n, \mathbb{H}^n$ whose dihedral angles between top dimensional faces are all of the form π/m_i for integers m_i, the group G_P generated by reflections in these faces of P acts properly discontinuously on X with fundamental domain P. G_P has a subgroup of index 2 consisting of orientation preserving elements, which has as fundamental domain a copy of P and its mirror image P'. This follows from a theorem called *Poincaré's polyhedron theorem*. A precise statement and discussion are found in [7].

If $X = \mathbb{S}^n$, we can think of $\mathbb{S}^n \subset \mathbb{R}^{n+1}$ as the unit sphere, and reflections through hyperplanes in $\mathbb{S}^n$ correspond to reflections in hyperplanes through the origin in $\mathbb{R}^{n+1}$. If π_i, π_j are two of these hyperplanes, and the corresponding reflections are denoted r_i, r_j then the composition r_i, r_j is a rotation through an angle $2\theta_{ij}$, where θ_{ij} is the angle between the planes π_i and π_j, and the rotation is in the plane spanned by the two perpendiculars to π_i, π_j. A presentation for the group G generated by reflections in the sides of P is

$$G_P \cong \langle r_1, r_2, \ldots, r_n | (r_1)^2, (r_2)^2, \ldots, (r_n)^2, (r_1 r_2)^{m_{12}}, \ldots, (r_i r_j)^{m_{ij}}, \ldots \rangle$$

where the angle between π_i and π_j is π/m_{ij}.

The subgroup G_P^+ of orientation–preserving elements of G_P is generated by the elements of the form $r_i r_j$, which is to say, G_P^+ consists of products of even numbers of reflections.

Definition 4.2.17. A *Coxeter group G* is an abstract group defined by a group presentation of the form $\langle r_i | (r_i r_j)^{k_{ij}} \rangle$ where

- the indices vary over some index set I
- the exponent $k_{ij} = k_{ji}$ is either a positive integer or ∞ for each pair i, j
- $k_{ii} = 1$ for each i
- $k_{ij} > 1$ for each $i \neq j$

If $k_{ij} = \infty$ for some i, j then the corresponding relation is meaningless and may be deleted from the presentation.

Definition 4.2.18. The *Coxeter graph* of the Coxeter group G is a labelled graph Γ with vertices corresponding to the index set I and edges $\langle (i, j) : k_{ij} > 2 \rangle$ labelled by k_{ij}.

For simplicity, edges with $k_{ij} = 3$ are usually left unlabelled.

Theorem 4.2.19. *Finite Coxeter groups can be realized as properly discontinuous spherical reflection groups.*

Notice that fundamental groups of triangle orbifolds are index 2 subgroups of reflection groups whose Coxeter graphs have three vertices.

"Bad" orbifolds

If Σ is a spherical orbifold with two cone points of order $p, q > 1$ where $p \neq q$, the orbifold Euler characteristic of Σ is $2 - \frac{p-1}{p} - \frac{q-1}{q} > 0$, so the universal cover of Σ should be $\mathbb{S}^2$. But we have seen that this is impossible; the loop around the point of order p is freely homotopic to the loop of order q, so a presentation for $\pi_1^o(\Sigma)$ is $\langle \alpha | \alpha^p, \alpha^q, \alpha^{p-q} \rangle$. This group is $\mathbb{Z}/d\mathbb{Z}$ where d is the greatest common divisor of p and q; but every $\mathbb{Z}/d\mathbb{Z}$ subgroup of $SO(3, \mathbb{R})$ has as quotient the spherical orbifold with two cone points of order d.

Thus Σ is not obtained from a smooth surface by the action of a properly discontinuous group. An orbifold with no manifold cover is called a *bad orbifold*. A spherical orbifold with two cone points of unequal order is a bad orbifold; similarly, a spherical orbifold with one cone point is a bad orbifold.

It turns out that any orbifold whose underlying surface is not the sphere, but a surface of higher genus, is a good orbifold, and is obtained as a quotient X/Γ for X one of $\mathbb{S}^2, \mathbb{E}^2, \mathbb{H}^2$ and Γ a properly discontinuous group of isometries.

4.2.4 Discrete subgroups of $PSL(2, \mathbb{R})$

Glueing hyperbolic polygons

Gluing up hyperbolic polygons to make a closed hyperbolic surface is not essentially different from glueing up spherical or flat polygons. If Σ_g denotes the unique (up to homeomorphism) closed orientable surface of genus $g > 1$, then we can decompose Σ_g (nonuniquely) into *pairs of pants*.

Definition 4.2.20. A *pair of pants* is the topological surface obtained from a disk by removing two subdisks — that is, a disk with two holes.

A pair of pants can also be thought of as a sphere minus three subdisks. The Euler characteristic of a pair of pants is -1. Since the Euler characteristic of its boundary is 0, a surface obtained from glueing n pairs of pants has Euler characteristic $-n$. So Σ_g can be decomposed (in many different ways) into $2g-2$ pairs of pants.

Exercise 4.2.21. *Show that the number of decompositions of Σ_g into pairs of pants, up to combinatorial equivalence, is equal to the number of graphs with $2g-2$ vertices with 3 edges at every vertex. Such graphs are called* trivalent graphs. *Show that the number of such graphs is positive for $g > 1$, and enumerate such graphs for $g \leq 5$ (you might need to write a computer program . .)*

For α a closed loop in Σ_g, a choice of hyperbolic metric on Σ_g determines a unique shortest loop α_g — a geodesic — which is homotopic to α. For, if $p \in \alpha$ and $\widetilde{\alpha}$ denotes an arc in $\mathbb{H}^2$, the universal cover of Σ_g, whose endpoints project to p and such that $\widetilde{\alpha}$ projects to α under the covering map, then there is a unique isometry $\gamma \in PSL(2,\mathbb{R})$ corresponding to an element of $\pi_1(\Sigma_g)$ taking one end of $\widetilde{\alpha}$ to the other. If $\widetilde{\alpha_g}$ denotes the invariant axis of γ, then $\widetilde{\alpha_g}/\gamma = \alpha_g$ a geodesic in Σ_g.

Exercise 4.2.22. *If α is an essential simple closed curve in Σ_g — that is, it is embedded and does not bound a disk, then α_g is also simple. Furthermore, if α, β are disjoint essential simple closed curves, their geodesic representatives α_g, β_g are disjoint.*

By the exercise, a combinatorial decomposition of Σ_g into pairs of pants determines, for a hyperbolic metric on Σ_g, a (combinatorially equivalent) decomposition of that surface into hyperbolic pairs of pants with geodesic boundary. Call such an object a *geodesic pair of pants*.

Let P be a geodesic pair of pants with boundary circles α, β, γ, and δ an embedded arc joining two distinct boundary components α, β. Then we can let Q be the surface, topologically a torus with two disks removed, obtained from two copies of P with opposite orientations glued along the pairs of circles corresponding to α, β. Then the two copies of δ make up a closed loop $\hat{\delta}$ which has a unique geodesic representative $\hat{\delta}_g \subset Q$. There is an orientation–reversing map i from Q to itself which fixes $\alpha \cup \beta$. By uniqueness, $\hat{\delta}_g$ is invariant under i, and therefore it intersects the boundary curves in right angles. We obtain an arc δ_g in P perpendicular to α and β. There are two other arcs ϵ_g, λ_g in P perpendicular to α, γ and β, γ. These decompose P into two right angled hyperbolic hexagons H_1, H_2. The alternate sides of H_1 and H_2 are equal, and therefore they are isometric, by an orientation–reversing isometry.

Exercise 4.2.23. *Prove the claim made in the previous paragraph. That is, show that a right–angled hyperbolic hexagon is determined up to isometry by the lengths of three nonadjacent sides. Conversely show that for any three number $p, q, r > 0$ there is a right–angled hexagon with three nonadjacent sides with those lengths.*

In short we have proved the following fact:

Lemma 4.2.24. *Let Σ_g be a surface of genus g. A combinatorial decomposition into pairs of pants and a hyperbolic metric on Σ_g determine a decomposition of Σ_g into $2g - 2$ geodesic pairs of pants. The geometry of these pairs of pants is determined uniquely by the lengths of the closed geodesics along which Σ_g was decomposed.*

It remains to understand how the pairs of pants can be put back together to give Σ_g. For P_1, P_2 a pair of geodesic pairs of pants with boundary components $\alpha_1 \subset \partial P_1$ and $\alpha_2 \subset \partial P_2$ with the same length, for any two points $p_1 \in \alpha_1$ and $p_2 \in \alpha_2$ there is a unique way to glue P_1 to P_2 by identifying α_1, α_2 so that p_1, p_2 match up. There is a 1–parameter family of glueings, parameterized by the amount of "twisting" of these geodesics. In particular, the geometry of Σ_g is determined uniquely by the $3g - 3$ lengths of the geodesics along which it is decomposed, together with $3g - 3$ twist parameters.

Thus we have a correspondence:

$$(\text{metric on } \Sigma_g, \text{pair of pants decomposition}) \longleftrightarrow (\mathbb{R}^+)^{3g-3} \times (\mathbb{R}/\mathbb{Z})^{3g-3}$$

Here the pair of pants decomposition is thought of as *ordered*, in the sense that the $3g - 3$ curves are given specific labels, which correspond to the $3g - 3$ co–ordinates on the right.

Although this is a nice characterization, the information contained in a pair of pants decomposition is both too little and too much — too little because we have not resolved the $\mathbb{Z}$–ambiguity in the twist parameters, and too much because it does not answer the question of what the space of hyperbolic structures on a surface is parameterized by. We address these issues now.

Definition 4.2.25. Fix a base surface Σ of genus $g > 1$. The space of *marked hyperbolic structures on* Σ, denoted $\mathscr{MH}(\Sigma)$ is the space of equivalence classes of pairs (f, Σ') where Σ' is a hyperbolic surface and $f : \Sigma \to \Sigma'$ is a homeomorphism, and two such pairs (f_1, Σ_1) and (f_2, Σ_2) are equivalent if there is an isometry $i : \Sigma_1 \to \Sigma_2$ such that $f_1 \circ i$ is homotopic to f_2 as a map from Σ to Σ_2.

Exercise 4.2.26. *Show that the relation defined in the definition of marked hyperbolic structure is really an equivalence relation. That is, show it is symmetric, reflexive and transitive.*

Theorem 4.2.27. *For Σ a surface of genus g, there is a 1–1 correspondence*

$$\mathscr{MH}(\Sigma) \longleftrightarrow (\mathbb{R}^+)^{3g-3} \times \mathbb{R}^{3g-3}$$

The correspondence is defined as follows: there is a pair of pants decomposition along essential simple closed curves $\alpha_1, \ldots, \alpha_{3g-3}$ for Σ, and a collection of loops $\beta_1, \ldots, \beta_{3g-3}$ transverse to the α_i such that if (f, Σ') is an element in $\mathscr{MH}(\Sigma)$, the corresponding co–ordinates are given by

$$(\textit{length}((f(\alpha_1))_g), \ldots, \textit{length}((f(\alpha_{3g-3}))_g), \textit{twist}((f(\alpha_1))_g), \ldots, \textit{twist}((f(\alpha_{3g-3}))_g))$$

where the twist parameters are normalized so that twist 0 corresponds, for fixed lengths of $(f(\alpha_i))_g$, to the unique marked surface for which the length of β_i is minimized.

This requires some explanation. The image of a fixed pair of pants decomposition of Σ under f determines one in Σ', and therefore the lengths of the decomposed geodesics are well–defined and the twist parameters are well–defined mod 2π. Let P_1, P_2 be two geometric pairs of pants glued along boundaries to make a sphere with 4 disks removed Q. If α_i is the common loop in $\partial P_1 \cap \partial P_2$, then β_i is a dual curve which cuts Q into two other pairs of pants P_1', P_2' such that P_1' has one boundary component in common with each of P_1, P_2 and similarly for P_2'. Then twisting α_i through 2π replaces β_i with a new curve $t_{\alpha_i}(\beta_i)$, the curve obtained by a *Dehn twist around* α_i. Briefly: β_i decomposes into two arcs δ, ϵ along α_i, and α_i decomposes into two arcs μ, ν along β_i. Then $t_{\alpha_i}(\beta_i)$ is the simple closed curve homotopic to $\delta * \mu * \epsilon * \nu$, wher $*$ denotes concatenation of arcs. Imagine β_i as a rubber band on the surface Q. When the two sides P_1, P_2 are twisted independently along α_i, the rubber band becomes twisted up, and when P_1 and P_2 return to their original configuration, the rubber band detects how many full rotations the two sides went through. It is true, though we don't prove it here, that there is a unique rotation for which the length of the geodesic representative of β_i is minimized. For a reference, see [1]. Thus for fixed sets of lengths of the geodesic representatives of all the α_j, we have well–defined twist parameters which detect the amount of twisting relative to this minimal twist. This shows the map to parameter space is well–defined. Conversely, such a collection of parameters defines a collection of geodesic pairs of pants and instructions for glueing them together to give a marked hyperbolic structure on Σ. Thus the two sets are the same and the theorem is proved.

Definition 4.2.28. Let Σ be a closed surface. The *mapping class group* of Σ, denoted $\mathrm{MC}(\Sigma)$, is defined to be the quotient group

$$\mathrm{MC}(\Sigma) = \mathrm{Homeo}(\Sigma)/\mathrm{Homeo}_0(\Sigma)$$

where $\mathrm{Homeo}_0(\Sigma)$ denotes the normal subgroup of self–homeomorphisms of Σ which are homotopic to the identity.

Exercise 4.2.29. *Show Homeo*$_0(\Sigma)$ *is a normal subgroup of Homeo*(Σ).

Notice that $\mathrm{Homeo}(\Sigma)$ acts on $\mathscr{MH}(\Sigma)$ by

$$\psi : (f, \Sigma') \to (\psi \circ f, \Sigma')$$

Moreover, if $\psi \in \mathrm{Homeo}_0(\Sigma)$, then

$$\psi(f, \Sigma') \sim (f, \Sigma')$$

with respect to the equivalence relation defined on representatives. That is, $\mathrm{MC}(\Sigma)$ acts on $\mathscr{MH}(\Sigma)$. Moreover, the quotient space is exactly the space of equivalence classes of elements in $\mathscr{MH}(\Sigma)$ where $(f_1, \Sigma_1) \sim (f_2, \Sigma_2)$ if and only if there is an isometry $i : \Sigma_1 \to \Sigma_2$. That is, two marked hyperbolic structures have the same orbits under $\mathrm{MC}(\Sigma)$ if and only if the underlying hyperbolic structures (forgetting the marking) are equivalent. In particular, there is a corresponding action of $\mathrm{MC}(\Sigma)$ on $(\mathbb{R}^+)^{3g-3} \times \mathbb{R}^{3g-3}$ and therefore a correspondence

$$\text{hyperbolic structures on } \Sigma \longleftrightarrow \{(\mathbb{R}^+)^{3g-3} \times \mathbb{R}^{3g-3}\}/\mathrm{MC}(\Sigma)$$

The action of MC(Σ) on $\mathbb{R}^{6g-6}$ is properly discontinuous, but it is not free. Thus the space of hyperbolic structures on Σ is best thought of as an orbifold. This quotient space is also known as *moduli space*.

Exercise 4.2.30. *Verify the claims made above. In particular, show that the action of MC(Σ) on $\mathscr{M}\mathscr{H}(\Sigma)$ is well–defined, independently of the choice of representative of an element in $\mathscr{M}\mathscr{H}(\Sigma)$.*

The space $\mathscr{M}\mathscr{H}(\Sigma)$ is also known as the *Teichmüller space* of Σ, and denoted Teich(Σ).

Definition 4.2.31. Let Σ be a closed surface.

Let Homeo$^+(\Sigma)$ denote the subgroup of Σ consisting of *orientation–preserving* homeomorphisms. Then define

$$\text{MC}^+(\Sigma) = \text{Homeo}^+(\Sigma)/\text{Homeo}_0(\Sigma)$$

Notice that in this definition we use implicitly the fact that for a *closed* surface, the subgroup of self–homeomorphisms homotopic to the identity are all orientation–preserving. This is not true for surfaces with boundary without some extra constraints on the boundary behaviour of these homeomorphisms.

Exercise 4.2.32. *Let Σ be the unit disk. Find a self–homeomorphism homotopic to the identity which is orientation–reversing. Do the same with Σ an annulus. What about if Σ is a punctured surface of genus $g \geq 2$?*

4.2.5 Dehn twists and Lickorish's theorem

Definition 4.2.33. An oriented (polyhedral) simple closed curve c in a surface Σ and an annulus neighborhood A of c parameterized as $S^1 \times I$ define a homeomorphism $t_c : \Sigma \to \Sigma$ by

$$t_c : \begin{cases} x \to x & \text{for } x \text{ outside } A \\ (\theta, t) \to (\theta - 2t\pi, t) & \text{for } (\theta, t) \in A \end{cases}$$

This homeomorphism is known as a *Dehn twist* about c. As an element of MC(Σ), it depends only on the isotopy class of c.

Note that $[t_c]^{-1} = [t_{c'}]$ where c' denotes c with the opposite orientation.

Exercise 4.2.34. *If $h : \Sigma \to \Sigma$ is a homeomorphism, p a simple closed loop in Σ, and $h(p) = q$, then*

$$t_p = h^{-1} t_q h$$

The following theorem is proved in [4], and is often referred to as the *Lickorish twist theorem*:

Theorem 4.2.35 (Lickorish). *If Σ_g denotes the oriented surface of genus g, then the group $MC^+(\Sigma_g)$ is generated by Dehn twists in $3g-1$ (explicitly described) simple closed curves. In particular, this group is finitely generated.*

Proof. The method of proof proceeds as follows: let c be a simple closed curve in Σ, and let A be a collection of simple closed curves in Σ. Then either c intersects each element α of A not at all, exactly once, or exactly twice with opposite orientations, or there exists a loop d which intersects α fewer times than c, and each element of A at most as many times as c, such that $t_\alpha(c)$ has fewer intersections with α and the same number or fewer intersections with each other element of A. Proceeding inductively, we see that if C is a maximal collection of disjoint essential simple closed curves and ψ is a homeomorphism of Σ, then there are a sequence of Dehn twists $t_1, t_2, \dots$ such that $t_n t_{n-1} \dots t_1 \psi(C)$ intersects C in one of finitely many possibilities. After twisting some more in elements of C, we can assume the image of C is one of finitely many possibilities, which can be explicitly identified. In short, ψ can be written as a product of Dehn twists.

Now, for each such twist t_c, we can replace t_c by $t_d t_{t_d(c)} t_d^{-1}$ where each $d, t_d(c)$ intersect C more simply than c. In this way, each t_c can be expanded as a product of Dehn twists in curves which intersect C very simply. After twisting in C, it follows that these involve only finitely many possibilities, which can be explicitly enumerated. □

Remark 4.2.36. Casson has shown that the number of twist generators required is at most $2g + 2$. Furthermore, it is known that $\mathrm{MC}^+(\Sigma)$ is generated by only 2 elements (which are not Dehn twists).

Chapter 5

Riemannian geometry

5.1 Liouville's theorem

In two dimensions, conformal geometry is very flexible. Any (nonsingular) holomorphic or antiholomorphic map between open regions in $\mathbb{C}$ is conformal.

However, in three dimensions and higher, conformal geometry becomes very rigid. The fundamental theorem in this subject is Liouville's theorem, which characterizes conformal maps between open subsets of $\mathbb{E}^n$.

Theorem 5.1.1 (Liouville). *Let U, V be open sets in $\mathbb{E}^n$ with $n \geq 3$ and suppose $\phi : U \to V$ is conformal. Then ϕ is of the form*

$$\phi : x \to \lambda A i(x) + t$$

where $\lambda > 0$ is a dilation, t is a translation, A is orthogonal, and i is either an inversion in a sphere or the identity.

Proof. By the definition of a conformal map, there is a positive valued function μ on U such that

$$|df(v)| = \mu(x)|v|$$

for all $x \in U$ and $v \in T_x\mathbb{E}^n$. Define $\rho(x) = \frac{1}{\mu(x)}$.

It is clear that if we know ρ, then we also know μ, and therefore can measure length of vectors in $\phi(U)$. In particular, the Riemannian metric in $\phi(U)$ is intrinsically determined by ϕ. It follows that if we know that $\phi(U)$ is isometric to a subset of $\mathbb{E}^n$, then ϕ itself is determined by ρ, up to composition with an isometry of $\mathbb{E}^n$. Furthermore, ϕ is of the form $x \to \lambda A(x) + t$ iff ρ is constant, and ϕ is of the form $x \to \lambda A i(x) + t$ for a nontrivial inversion i iff $\rho = \eta|x - x_0|^2$ for some non–zero constant η and basepoint x_0.

We introduce some notation as follows. For each $p \in U$,

$$d\phi : T_pU \to T_{\phi(p)}V$$

denotes the induced map on tangent spaces. For X, Y vector fields, let $\nabla_X Y$ denote the derivative of the vector field Y in the direction tangent to X. If X and Y commute, then this expression is symmetric in X and Y. Then define

$$d^2\phi(X,Y) = \nabla_{d\phi(X)} d\phi(Y) - d\phi(\nabla_X Y)$$

If $[X, Y] = 0$, then $[d\phi(X), d\phi(Y)] = 0$, and therefore this expression is symmetric in X and Y for commuting vector fields.

Now, $\mathbb{E}^n$ admits a family of orthonormal parallel commuting vector fields e_i. Since this family is parallel, $\nabla_{e_i} e_j = 0$ for all i, j. It follows that

$$d^2\phi(e_i, e_j) = \nabla_{d\phi(e_i)} d\phi(e_j)$$

We can also define $d^2\rho : T_U \times T_U \to \mathbb{R}$. Note that $d^2\rho$ is just another name for the *Hessian* of ρ.

Now for any vectors $u, v \in T_pU$, we can extends u, v, w to parallel commuting vector fields on $\mathbb{E}^n$, and by abuse of notation, define $d^2\phi(u, v) \in T_{\phi(p)}U$ in this way.

Claim 5.1.2. *Suppose $u \perp w$. Then $d^2\rho(u, w) = 0$.*

Proof. Let v be another vector orthogonal to both u and w. We extend u, v, w to vector fields locally which are constant in local co–ordinates. Now, we have the identity

$$\langle d\phi(u), d\phi(v)\rangle = 0$$

Taking the partial derivative of both sides with respect to w, we get

$$\begin{aligned} 0 &= d\phi(w)\langle d\phi(u), d\phi(v)\rangle \\ &= \langle \nabla_{d\phi(w)} d\phi(u), d\phi(v)\rangle + \langle d\phi(u), \nabla_{d\phi(w)} d\phi(v)\rangle \end{aligned}$$

which we can rewrite as

$$\langle d^2\phi(u, w), d\phi(v)\rangle = -\langle d\phi(u), d^2\phi(v, w)\rangle$$

Note that the left side is symmetric in u, w and the right side is symmetric in v, w. On the other hand the equality implies both sides are antisymmetric in u, v. So both sides must be identically zero.

Since v is arbitrary, it follows that $d^2\phi(u, w)$ has no component normal to the space spanned by $d\phi(u)$ and $d\phi(w)$. It follows that there are functions α, β such that

$$d^2\phi(u, w) = \alpha(x) d\phi(u) + \beta(x) d\phi(w)$$

By definition, we have

$$\langle d\phi(u), d\phi(u)\rangle = \mu(x)^2\langle u, u\rangle$$

Taking the partial derivatives of both sides with respect to w, we get

$$\langle d^2\phi(u, w), d\phi(u)\rangle = \mu(x) d\mu(w)|u|^2$$

Notice that $|u|$ is constant as a function of x by our assumption.

It follows that

$$\alpha(x) = \frac{d\mu(w)}{\mu(x)}$$

and

$$\beta(x) = \frac{d\mu(u)}{\mu(x)}$$

Since $d\rho(z) = -\frac{d\mu(z)}{\mu(x)^2}$ we have

$$\rho(x)d^2\phi(u,w) + d\rho(w)d\phi(u) + d\rho(u)d\phi(w) = 0$$

Taking partial derivatives with respect to v we get

$$\begin{aligned} d\rho(v)d^2\phi(u,w) + \rho(x)d^3\phi(u,w,v) + d^2\rho(w,v)d\phi(u) &\quad + \\ d\rho(w)d^2\phi(u,v) + d^2\rho(u,v)d\phi(w) + d\rho(u)d^2\phi(w,v) &\quad = \quad 0 \end{aligned}$$

Now, terms 2,4,5 are symmetric in u, v. Moreover the sum of terms 1 and 6 is also symmetric in u, v. It follows that term 3 is symmetric in u, v. That is,

$$d^2\rho(w,v)d\phi(u) = d^2\rho(w,u)d\phi(v)$$

Since $d\phi(u)$ and $d\phi(v)$ are orthogonal, $d^2\rho(w,u) = 0$. □

It follows that at each point, $d^2\rho$ is a symmetric bilinear form, and if e_i is the canonical basis for $\mathbb{E}^n$, $d^2\rho(e_i, e_j) = k_i\delta^i_j$ for some constants k_i. Moreover

$$0 = d^2\rho(e_i + e_j, e_i - e_j) = k_i - k_j$$

so these constants are equal, and $d^2\rho$ is k times the usual Euclidean inner product, and we can write

$$d^2\rho(u,w) = \eta(x)\langle u, w\rangle$$

for some $\eta(x)$.

Taking the partial derivative of this expression with respect to v, we get

$$d^3\rho(w,u,v) = d\eta(v)\langle w, u\rangle$$

where now w, u, v are arbitrary. Since d^3 is symmetric, we have

$$\langle d\eta(v)w - d\eta(w)v, u\rangle = 0$$

which implies that $d\eta(v)w = d\eta(w)v$ and therefore $d\eta = 0$; i.e. η is a constant.

It follows that $d^2\rho(u,w) = \eta\langle u, w\rangle$ for some constant η, and therefore ρ is a quadratic function. By examining the possibilities, it is easy to show ϕ is one of the two possibilities raised above. □

Bibliography

[1] R. Benedetti and C. Petronio, *Lectures on Hyperbolic Geometry*, Springer–Verlag Universitext (1992)

[2] J. Conway, *The sensual quadratic form*, Math. Ass. Amer. Carus mathematical monographs **26**, (1997)

[3] M. Gromov, *Metric structures for Riemannian and non–Riemannian spaces*, Birkhauser (1999)

[4] R. Lickorish, *A finite set of generators for the homeotopy group of a* 2*–manifold*, Proc. Camb. Phil. Soc. **60** (1964), pp. 769–778

[5] J. Montesinos, *Classical tessellations and three–manifolds*, Springer–Verlag (1987)

[6] T. Rado, *Über den Begriff der Riemannsche Fläche*, Acta Univ. Szeged **2** (1924–26), pp. 101–121

[7] J. Ratcliffe, *Foundations of hyperbolic manifolds*, Springer–Verlag GTM **149** (1994)

[8] K. Schütte and B. L. van der Waerden, *Auf welcher Kugel haben* 5, 6, 7, 8 *oder* 9 *Punkte mit Mindestabstand Eins Platz*, Math. Ann. **123** (1951), pp. 96–124

[9] J. Stallings, *Category language*, notes; available at `http://www.math.berkeley.edu/~stall`

[10] W. Thurston, *Three–dimensional geometry and topology, vol. 1*, Princeton University Press, Princeton Math. Series 35 (1997)

[11] M. do Carmo, *Riemannian geometry*, Birkhauser (1992)

[12] J. Conway and N. Sloane, *Sphere packings, lattices and codes*, Springer–Verlag

[13] D. Gabai, R. Meyerhoff and N. Thurston, *Homotopy implies isotopy for hyperbolic manifolds*

[14] R. Gompf, 4*–manifolds and Kirby Calculus??*

[15] J. Hocking and G. Young, *Topology*, Addison–Wesley (1961)

[16] B. Jessen, *The algebra of polyhedra and the Dehn–Sydler theorem*, Math. Scand. **22** (1968), pp. 241–256

[17] R. Kirby, *The topology of* 4*–manifolds*, Springer–Verlag LNM 1374 (1989)

[18] E. Moise, *Geometric Topology in Dimensions* 2 *and* 3, Springer–Verlag GTM **47** (1977)

[19] D. Mostow, *Rigidity of rank* 1 *locally symmetric spaces*

[20] K. Reidemeister, *Knot Theory*, translated from the German; BCS associates (1983)

[21] J–P. Sydler, *Conditions nécessaires et suffisantes pour l'équivalence des polyèdres de l'espace euclidien à trois dimensions*, Comm. Math. Helv. **40** (1965), pp. 43–80

[22] J. Tits, *Free subgroups in linear groups*, J. Algebra **20** (1972), pp. 250–270

[23] F. Warner, *Foundations of differentiable manifolds and Lie groups*, Scott, Foresman and co. (1971)