A.B.Sossinsky

GEOMETRIES

September 2011. First 13 chapters for Russian students. Note that "chapter" means "lecture".

Chapter 1
TOY GEOMETRIES AND MAIN DEFINITIONS

In this chapter, we study five toy examples of geometries (symmetries of the equilateral triangle, the square, the cube, and the circle) and a model of the geometry of the so-called elliptic plane. These examples are followed by the main definition of this course: a geometry in the sense of Klein is a set with a transformation group acting on it. We then present some useful general notions related to transformation groups. Finally, we study the relationships (called morphisms or equivariant maps) between different geometries, thus introducing the category of all geometries. The notions introduced in this chapter are illustrated by some problems (dealing with toy models of geometries) collected at the end of the chapter.

But before we begin with these topics, we briefly recall some terminology from elementary Euclidean geometry.

1.1. Isometries of the Euclidean plane and space

We assume that the reader is familiar with the basic notions and facts of Euclidean geometry in the plane and in space. One can think of Euclidean geometry as an axiomatic theory (not too rigorously taught in high school) or as a small chapter of linear algebra (the plane $\mathbb{R}^2$ and the space $\mathbb{R}^3$ supplied with the standard metric). It is irrelevant for us which of these two points of view is adopted by the reader, and the aim of this subsection is merely to fix some terminology common to the two approaches.

An *isometry* of the Euclidean plane $\mathbb{R}^2$ (or space $\mathbb{R}^3$) is a map $f : \mathbb{R}^2 \to \mathbb{R}^2$ (respectively $f : \mathbb{R}^3 \to \mathbb{R}^3$) which preserves the distance d between points, i.e., $d(f(P), f(Q)) = d(P, Q)$ for any pair of points P, Q of the plane (resp. of space). There are two types of isometries: those which preserve orientation (they are called *motions*) and those that reverse orientation (*orientation-reversing isometries*).

In the plane, examples of motions are *parallel translations* (determined by a fixed *translation vector*) and *rotations* (determined by a pair (C, α), where C is the *center of rotation* and α is *angle of rotation*. In space, examples of motions are *parallel translations* and *rotations* (about an axis). Rotations in space are determined by pairs (l, α), where l is the *axis of rotation*, i.e., a straight line with a specified direction on it, and α is the *angle of rotation*; the rotation (l, α) maps any point M in space to the point M' obtained by

rotating M in the plane Π perpendicular to l by the angle α counterclockwise if one looks at the plane from "above", i.e., from some point of l obtained from the point $l \cap \Pi$ by moving in the direction specified on the axis.

Examples of orientation-reversing isometries in the plane are *reflections* (i.e., symmetries with respect to a line). In space, examples of orientation-reversing isometries are given by *mirror symmetries* (i.e., reflections with respect to planes) and *point symmetries* (i.e., reflections with respect to a point).

All other isometries of the Euclidean plane and space are *compositions* of those listed above.

The reader who feels uncomfortable with the notions mentioned in this subsection is invited to study Appendix I at the end of the book.

1.2. Symmetries of some figures

1.2.1. *Symmetries of the equilateral triangle.* Consider all the isometries of the equilateral triangle $\triangle = ABC$, i.e., all the distance-preserving mappings of this triangle onto itself. (To be definite, we assume that the letters A, B, C have been assigned to vertices in counterclockwise order.) Denote by s_A, s_B, and s_C the reflections in the bisectors of angles A, B, C of the triangle. Denote by r_0, r_1, r_2 the counterclockwise rotations about its center of gravity by 0, 120, 240 degrees, respectively. Thus r_1 takes the vertex A to B, B to C, and C to A. These six transformations are all called *symmetries* of triangle ABC and the set that they constitute is denoted by $\mathrm{Sym}(\triangle)$. Thus

$$\mathrm{Sym}(\triangle) = \{r_0,\, r_1,\, r_2,\, s_A,\, s_B,\, s_C\}.$$

There are no other isometries of $\triangle$. Indeed, any isometry takes vertices to vertices, each one-to-one correspondence between vertices entirely determines the isometry. (For example, the correspondence $A \to B$, $B \to A$, $C \to C$ determines the reflection s_C.) But there are only six different ways to assign the letters A, B, C to three points, so there cannot be more than 6 isometries of $\triangle$.

In a certain sense, $\mathrm{Sym}(\triangle)$ is the same thing as the family of all permutations of the three letters A, B, C; this remark will be made precise in the next chapter.

We will use the symbol $*$ to denote the *composition* (or *product*) of isometries, in particular of elements of $\mathrm{Sym}(\triangle)$, and understand expressions such

as $r_1 * s_A$ to mean that r_1 is performed first, and then followed by s_A. Obviously, when we compose two elements of Sym($\triangle$), we always obtain an element of Sym($\triangle$).

What element is the composition of two given ones can be easily seen by drawing a picture of the triangle ABC and observing what happens to it when the given isometries are successively performed, but this can also be done without any pictures: it suffices to follow the "trajectory" of the vertices A, B, C. Thus, in the example $r_1 * s_A$, the rotation r_1 takes the vertex A to B, and then B is taken to C by the symmetry s_A; similarly, $B \to C \to B$ and $C \to A \to A$, so that the vertices A, B, C are taken to C, B, A in that order, which means that $r_1 * s_A = s_B$.

The order in which symmetries are composed is important, because the resulting symmetry may change if we inverse the order. Thus, in our example, $s_A * r_1 = s_C \neq s_B$ (as the reader will readily check), so that $r_1 * s_A \neq s_A * r_1$. So for elements of Sym($\triangle$), composition is *noncommutative*.

The compositions of all possible pairs of symmetries of $\triangle$ can be conveniently shown in the following *multiplication table*:

$*$	r_0	r_1	r_2	s_A	s_B	s_C
r_0	r_0	r_1	r_2	s_A	s_B	s_C
r_1	r_1	r_2	r_0	s_B	s_C	s_A
r_2	r_2	r_0	r_1	s_C	s_A	s_B
s_A	s_A	s_C	s_B	r_0	r_1	r_2
s_B	s_B	s_A	s_C	r_2	r_0	r_1
s_C	s_C	s_A	s_B	r_1	r_2	r_0

Here (for instance) the element s_V at the intersection of the fifth column and the third row is $s_B = r_1 * s_A$, the composition of r_1 and s_A in that order (first the transformation r_1 is performed, then s_A).

As we noted above, composition is *noncommutative*, and this is clearly seen from the table (it is not symmetric with respect to its main diagonal).

The composition operation $*$ in Sym($\triangle$) is (obviously) *associative*, i.e., $(i * j) * k = i * (j * k)$ for all $i, j, k \in$ Sym($\triangle$). The set Sym($\square$) contains the *identity* transformation r_0 (also denoted id or $\mathbf{1}$). Any element i of Sym($\square$) has an *inverse* i^{-1}, i.e., an element such that $i * i^{-1} = i^{-1} * i = \mathbf{1}$.

The set Sym($\triangle$) supplied with the composition operation $*$ is called the *symmetry group of the equilateral triangle*.

1.2.2. *Symmetries of the square.* Consider all the isometries of the unit square $\square = ABCD$, i.e., all the distance-preserving mappings of the square to itself.

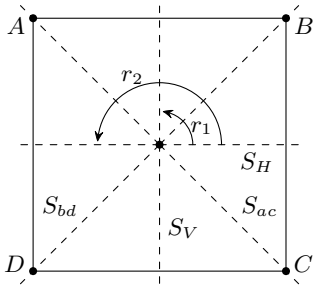


Figure 1.1. Symmetries of the square

Let us denote by s_H, s_V, and s_{ac}, s_{bd} the reflections in the horizontal and vertical mid-lines, and in the diagonals AC, BD, respectively. Denote by r_0, r_1, r_2, r_3 the rotations about the center of the square by 0, 90, 180, 270 degrees, respectively. These eight transformations are all called *symmetries* of the square. We write

$$\mathrm{Sym}(\square) = \{r_0,\, r_1,\, r_2,\, r_3,\, s_H,\, s_V,\, s_{ac},\, s_{bd}\}.$$

Just as in the case of the equilateral triangle, the composition of any two symmetries of the square is a symmetry of the square, and a *multiplication table*, indicating the result of all pairwise compositions, can be drawn up:

$*$	r_0	r_1	r_2	r_3	s_H	s_V	s_{ac}	s_{bd}
r_0	r_0	r_1	r_2	r_3	s_H	s_V	s_{ac}	s_{bd}
r_1	r_1	r_2	r_3	r_0	s_{ac}	s_{bd}	s_V	s_H
r_2	r_2	r_3	r_0	r_1	s_V	s_H	s_{bd}	s_{ac}
r_3	r_3	r_0	r_1	r_2	s_{bd}	s_{ac}	s_H	s_V
s_H	s_H	s_{bd}	s_V	s_{ac}	r_0	r_2	r_3	r_1
s_V	s_V	s_{ac}	s_H	s_{bd}	r_2	r_0	r_1	r_3
s_{ac}	s_{ac}	s_H	s_{bd}	s_V	r_1	r_3	r_0	r_2
s_{bd}	s_{bd}	s_V	s_{ac}	s_H	r_3	r_1	r_2	r_0

Here (for instance) the element s_V at the intersection of the sixth column and the fourth row is $s_V = r_2 * s_H$, the composition of r_2 and s_H in that

order (first the transformation r_2 is performed, then s_V). Composition is *noncommutative*.

Obviously, composition is *associative*. The set Sym($\square$) contains the *identity* transformation r_0 (also denoted id or **1**). Any element i of Sym($\square$) has an *inverse* i^{-1}, i.e., an element such that $i * i^{-1} = i^{-1} * i = \mathbf{1}$..

The set Sym($\square$) supplied with the composition operation is called the *symmetry group of the square*.

1.2.3. *Symmetries of the cube.* Let

$$I^3 = \{(x, y, z) \in \mathbb{R}^3 |\ 0 \le x \le 1, 0 \le y \le 1, 0 \le z \le 1\}$$

be the unit cube. A *symmetry* of the cube is defined as any isometric mapping of I^3 onto itself. The composition of two symmetries (of I^3) is a symmetry. How many are there?

Let us first count the orientation-preserving isometries of the cube (other than the identity), i.e., all its rotations (about an axis) by nonzero angles that take the cube onto itself.

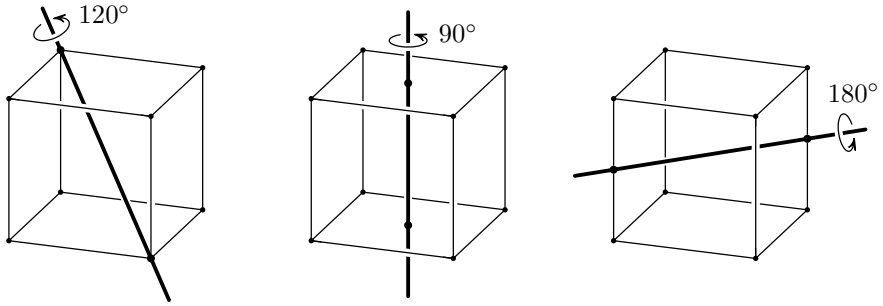


Figure 1.2. Rotations of the cube

There are three axes of rotation joining the centers of opposite faces, and the rotation angles for each are $\pi/2$, π, $3\pi/2$. There are four axes of rotation joining opposite vertices, the rotation angles for each being $2\pi/3$ and $4\pi/3$. There are six axes of rotation joining midpoints of opposite edges, with only one nonzero rotation for each (by π). This gives us a total of $(3 \times 3) + (4 \times 2) + (6 \times 1) = 23$ orientation-preserving isometries, or 24 if we include the identity.

There are no other orientation-preserving isometries; at this point, we could prove this fact by a tedious elementary geometric counting argument,

but we postpone the proof to Chapter 3, where it will be the immediate result of more general and sophisticated algebraic method.

There are also 24 orientation-reversing isometries of the cube. Listing them all is the task prescribed by Exercise 1.2 (see the end of the chapter), a task which requires little more than a bit of spacial intuition.

Thus the cube has 48 isometries. All their pairwise compositions constitute a multiplication table, which is a 49 by 49 array of symbols, much too unwieldy to fit in a book page.

The set $\mathrm{Sym}(I^3)$ of all 48 symmetries of the cube supplied with the composition operation is called the *symmetry group of the cube*; it is associative, noncommutative, has an identity, and all its elements have inverses, just as the symmetry groups in the two previous examples.

1.2.4. *Symmetries of the circle.* Let

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

be the unit circle. Denote by $\mathrm{Sym}(\bigcirc)$ the set of all its isometries. The elements of $\mathrm{Sym}(\bigcirc)$ are of two types: the rotations r_φ about the origin by angles φ, $\varphi \in [0, 2\pi)$, and the reflections in lines passing through the origin, s_α, $\alpha \in [0, \pi)$, where α denotes the angle from the x-axis to the line (in the counterclockwise direction). The composition of rotations is given by the (obvious) formula

$$r_\phi * r_\psi = r_{(\phi+\psi) \bmod 2\pi},$$

where mod 2π means that we subtract 2π from the sum $\phi + \psi$ if the latter is greater than or equal to 2π.

The composition of two reflections s_α and s_β is a rotation by the angle $|\alpha - \beta|$,

$$s_\alpha * s_\beta = r_{2(\alpha-\beta)}.$$

The interested reader will readily verify this formula by drawing a picture and comparing the angles that will appear when the two reflections are composed.

The set of all isometries of the circle supplied with the composition operation is called the *symmetry group of the circle* and is denoted by $\mathrm{Sym}(\bigcirc)$. The group $\mathrm{Sym}(\bigcirc)$ has an infinite number of elements. As before, this group is associative, noncommutative, has an identity, and all its elements have inverses.

1.2.5. *Symmetries of the sphere.* Let

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

be the unit sphere. Denote by $\mathrm{Sym}(\mathbb{S}^3)$ the set of all its isometries and by $\mathrm{Rot}(\mathbb{S}^3)$ the set of all its rotations (by different angles about different axes passing through the center of the sphere). Besides rotations, the transformation group $\mathrm{Sym}(\mathbb{S}^3)$ contains reflections in different planes passing through the center of the sphere, its symmetry with respect to its center, and the composition of these transformations with rotations.

Reflections in planes, unlike rotations, reverse the orientation of the sphere. This means that a little circle oriented clockwise on the sphere (if we are looking at it from the outside) is transformed by any reflection into a counterclockwise oriented circle, and the picture of a left hand drawn on the sphere becomes that of a right hand. Now a reflection in a line passing through the sphere's center does not reverse orientation (unlike reflections in the plane!) because a reflection of the sphere in a line is exactly the same transformation as a rotation about this line by 180°. On the other hand, a reflection of the sphere with respect to its center reverses its orientation (again, this is not the case for reflections of the plane with respect to a point).

Note that the composition of two reflections in planes is a rotation (see Exercise 1.11), while the composition of two rotations is another rotation (by what angle and about what axis is the question discussed in Exercise 1.12).

The set of all isometries of the sphere supplied with the composition operation is called the *symmetry group of the sphere* and is denoted by $\mathrm{Sym}(\mathbb{S}^3)$. The group $\mathrm{Sym}(\mathbb{S}^3)$ has an infinite number of elements. As before, this group is associative, noncommutative, has an identity, and all its elements have inverses.

1.2.6. *A model of elliptic plane geometry.* Consider the set $Ant(\mathbb{S}^2)$ of all pairs of antipodal points (i.e., points symmetric with respect to the origin) on the unit sphere $\mathbb{S}^2$); thus elements of $Ant(\mathbb{S}^2)$ are *not* ordinary points, but *pairs of points*. Now consider the family (that we denote O(3)) of all isometries of the space $\mathbb{R}^3$ that do not move the origin [1]. Clearly, any such isometry takes pairs of antipodal points to pairs of antipodal points, thus it maps the set $X = Ant(\mathbb{S}^2)$ to itself.

The family O(3) of transformations of the set $Ant(\mathbb{S}^2)$ is called the *isometry group of the Riemannian elliptic plane*. This is a much more complicated object than the previous "toy geometries". We will come back to its study in Chapter 6.

[1] In linear algebra courses such transformations are called *orthogonal* and O(3) is called the *orthogonal group*.

1.3. Transformation groups

1.3.1. *Definitions and notation.* Let X be a set (finite or infinite) of arbitrary elements called *points*. By definition, a *transformation group* G *acting on* X is a (nonempty) set G of bijections of X supplied with the composition operation $*$ and satisfying the following conditions:

(i) G is closed under composition, i.e., for any transformations $g, g' \in G$, the composition $g * g'$ belongs to G;

(ii) G is closed under taking inverses, i.e., for any transformation $g \in G$, its inverse g^{-1} belongs to G.

These conditions immediately imply that G contains the identity transformation. Indeed, take any $g \in G$; by (ii), we have $g^{-1} \in G$; by (i), we have $g^{-1} * g \in G$; but $g^{-1} * g =$id (by definition of inverse element), and so id$\in G$. Note also that composition in G is associative (because the composition of mappings is always associative).

If $x \in X$ and $g \in G$, then by xg we denote the image of the point x under the transformation g. (The more usual notation $g(x)$ is not convenient: we have $x(g * h) = (xg)h$, but $(g * h)(x) = h(g(x))$, with g and h appearing in reverse order in the right-hand side of this equality.)

1.3.2. *Examples.* The five toy geometries considered in the previous section all give examples of transformation groups. The five transformation groups Sym act (by isometries) on the equilateral triangle, the square, the cube, the circle, and the sphere, respectively. In the last example (1.2.5), the orthogonal group O(3) acts on pairs of antipodal points on the sphere, these pairs being regarded as "points" of the "elliptic plane".

More examples are given by the transformation group consisting of all the bijections Bij(X) of any set X. By definition of transformation groups, Bij(X) is the largest (by inclusion) transformation group acting on the given set X. At the other extreme, any set X has a transformation group consisting of a single element, the identity transformation.

When the set X is finite and consists of n objects, the group Bij(X) of all its bijections is called the *permutation group* on n objects and is denoted by Σ_n. This group is one of the most fundamental notions of mathematics, and plays a key role in abstract algebra, linear algebra, and, as we shall see already in the next chapter, in geometry.

1.3.3. *Orbits, stabilizers, class formula.* Let $(X : G)$ be some transformation group acting on a set X and let $x \in X$. Then the *orbit* of x is defined

as

$$\mathrm{Orb}(x) := \{xg | g \in G\} \subset X,$$

and the *stabilizer* of x is

$$\mathrm{St}(x) := \{g \in G | xg = x\} \subset G.$$

For example, if $X = \mathbb{R}^2$ and G is the rotation group of the plane about the origin, then the set of orbits consists of the origin and all concentric circles centered at the origin; the stabilizer of the origin is the whole group G, and the stabilizers of all the other points of $\mathbb{R}^2$ are trivial (i.e., they consist of one element – the identity $\mathrm{id} \in G$).

Suppose $(X : G)$ is an action of a finite transformation group G on a finite set X. Then the number of points of G is (obviously) given by

$$\boxed{|G| = |\mathrm{Orb}(x)| \times |\mathrm{St}(x)|} \tag{1.1}$$

for any $x \in X$. Now let $A \subset X$ be a set that intersects each orbit at exactly one point. Then the number of points of X is given by the formula

$$\boxed{|X| \ = \ \sum_{x \in A} \frac{|G|}{|\mathrm{St}(x)|}}, \tag{1.2}$$

called the *class formula*. This formula, just as the previous one, follows immediately from definitions.

1.3.4. *Fundamental domains.* If X is a subset of $\mathbb{R}^n$ (e.g. $\mathbb{R}^n$ itself) and G is a transformation group acting on X, then a subset $F \subset X$ is called a *fundamental domain* of the action of G on X if

- F is an open set in X ;
- $F \bigcap Fg = \varnothing$ for any $g \in G$ (except $g =$id).;
- $X = \bigcup_{g \in G} \mathrm{Clos}(Fg)$, where $\mathrm{Clos}(.)$ denotes the closure of a set.

For example, in the case of the square, a fundamental domain of the action of $\mathrm{Sym}(\square)$ is the interior of the triangle AOM, where O is the center of the square and M is the midpoint of side AB; of course $\mathrm{Sym}(\square)$ has many other fundamental domains. Thus fundamental domains are not necessarily

unique. Moreover, fundamental domains don't always exist: for instance, $\mathrm{Sym}(S^1)$ (and other "continuous" geometries) do not have any fundamental domains.

1.3.5. *Morphisms.* According to one of the main principles of the category approach to mathematics, as soon as an important class of objects is defined, one must define their *morphisms*, i.e., the natural of class of relationships between them. Following this principle, we say any mapping of transformation groups $\alpha : G \to H$ is a *homomorphism* if α respects the product (composition) structure, i.e.,

$$\alpha(g_1 * g_2) = \alpha(g_1) * \alpha(g_2) \quad \text{for all} \quad g_1, g_2 \in G. \tag{1.3}$$

Let us look at a few examples of homomorphisms:

(i) the mapping μ :$\mathrm{Sym}(\square) \to \mathrm{Sym}(I^3)$ obtained by placing the square on top of the cube and extending its isometries to the whole cube in the natural way (e.g. assigning the rotation by 90° about the vertical axis passing through the centers of the horizontal faces of the cube to the 90° rotation of the square);

(iv) the mapping ν :$\mathrm{Sym}(\triangle)\to\mathrm{Sym}(\bigcirc)$ assigning to each rotation of the triangle the rotation of the circle by the same angle and, to the reflections s_A, s_B, s_C, the reflections $s_0, s_{2\pi/3}, s_{4\pi/3}$ of the circle;

(iii) the mapping π :$\mathrm{Rot}(I^3) \to \mathrm{Rot}(\square)$ induced by the projection of the cube on its bottom horizontal face Φ, i.e., assigning the identity element to all isometries of the cube that do not map Φ to Φ, and assigning, to all the other isometries of the cube, their restriction to Φ;

(iv) the mapping $\iota : S_3 \to\mathrm{Sym}(\triangle)$ assigning to each permutation of the symbols A, B, C the isometry that performs that permutation of the vertices A, B, C of the triangle.

The proof of the fact that these mappings are indeed homomorphisms, i.e., relation (1.3) holds, is a straightforward verification left to the reader.

A homomorphism α of transformation groups is said to be a *monomorphism* if the mapping α is injective (i.e., takes different elements to different ones). Examples of monomorphisms are the homomorphisms μ and ν above. A homomorphism α of transformation groups is said to be an *epimorphism* if α is surjective (i.e., is an onto map). An example is the mapping π above. A homomorphism α of transformation groups is said to be an *isomorphism*

if it is both a monomorphism and an epimorphism, i.e., if the mapping α is bijective.

Two transformation groups G and H acting on two sets X and Y (the case $X = Y$ is not forbidden) are called *isomorphic* if there exists an isomorphism $\phi : G \to H$. If two isomorphic groups are finite, then they necessarily have the same number of elements (but the number of points in the sets on which they act can differ, as for example in the case of the isomorphic groups $\mathrm{Sym}(\triangle)$ and S_3). Note in this context that $\mathrm{Sym}(\square)$ and S_4 are *not* isomorphic, because the first of these groups consists of 8 elements, while the second has $4! = 24$.

1.3.6. *Order.* The *order* of a transformation group G is, by definition, the number of its elements; we denote it by $|G|$. Thus $|\mathrm{Sym}(\triangle)| = 6$, $|\mathrm{Sym}(\square)| = 8$, and $|\mathrm{Sym}(\bigcirc)| = \infty$.

The *order* of an element g of a transformation group G is, by definition, the least positive integer k such that the element $g * g * \cdots * g$ (k factors) is the identity; this integer is denoted by $\mathrm{ord}(g)$; if there is no such integer, then g is said to be of *infinite order*. For example, the rotation by 30° in $\mathrm{Sym}(\bigcirc)$ has order 12, while the rotation by $\sqrt{2}\pi$ is of infinite order. (The last fact follows from the irrationality of $\sqrt{2}$)

1.3.6. *Subgroups.* Many important classes of objects have naturally defined "subobjects" (e.g. spaces and subspaces, manifolds and submanifolds, algebras and subalgebras). Transformation groups are no exception: if G is a transformation group and H is a subset of G, then H is called a *subgroup* of G if H itself is a group with respect to the composition operation $*$, i.e., if it satisfies the two conditions

(i) H is closed under composition, i.e., $g, g' \in H \Longrightarrow g * g' \in H$;

(ii) H is closed under taking inverses, i.e., $g \in G \Longrightarrow g^{-1} \in G$.

According to this definition, any transformation group G has at least two subgroups: G itself and its one-element subgroup, i.e., the group $\{\mathrm{id}\}$ consisting of the identity element. We will call these two subgroups *trivial*, and all the others, *nontrivial*.

For example, the subset of all rotations of the group $\mathrm{Sym}(\square)$ is a (nontrivial) subgroup of $\mathrm{Sym}(\square)$ (of order 4), the set consisting of the identity element and a reflection s_α is a subgroup of order 2 in $\mathrm{Sym}(\bigcirc)$, while the set of all rotations of $\mathrm{Sym}(\bigcirc)$ is a subgroup of infinite order.

If g is an element of order k in a transformation group G, then the set of k elements $\{g, g * g, \ldots, g * g * \cdots * g = \mathrm{id}\}$ is a subgroup of G of order

k; it is called the *cyclic subgroup of G generated by g*. This terminology is also used when g is an element of infinite order, but then the subgroup $\{\mathrm{id},\, g,\, g*g, \ldots,\, g*g*\cdots*g, \ldots\}$ is also of infinite order.

1.4. The category of geometries

In this section, we present the main definition of this course (that of a geometry) and define some related basic concepts.

1.4.1. *Geometries in the sense of Klein.* A pair $(X : G)$, where X is a set and G is a transformation group acting on X will be called a *geometry in the sense of Klein*. The five examples in Sec.1.2 define the geometry of the equilateral triangle, the geometry of the square, the geometry of the cube, the geometry of Riemann's elliptic plane. Another example is the set $\mathrm{Bij}(X)$ of all bijections of any set X.

1.4.2. *The Erlangen program.* The idea that geometries are sets of objects with transformation groups acting on them was first stated by the German mathematician Felix Klein in 1872 in a famous lecture at Erlangen. In that lecture (for an English translation, see [10]), he enunciated his views on geometries in the framework of what became known as the "Erlangener programme".

There is no doubt that all the geometries known in the times of Klein satisfy the property that he gave in his lecture, and so do all the geometries that were developed since then. However, this property can hardly be said to *characterize* geometries: it is much too broad. Thus, in the sense of the formal definition from the previous subsection, the permutation group is a geometry, and so is any topological space, any abstract group, even any set.

Nevertheless, we will stick to the notion of geometry given in 1.4.1 for want of a more precise formal definition. Such a definition, if it existed, would require supplying $(X : G)$ with additional structures (besides the action of G on X), but it is unclear at this time what these structures ought to be. If one looks at such branches of mathematics as global differential geometry, geometric topology, and differential topology, there appears to be no consensus among the experts about where geometry ends and topology begins in those fields.

The definition in 1.4.1 may be too broad, but it has the advantage of being simple and leading to the definition of a very natural category.

1.4.3. *Morphisms.* According to the general philosophy underlying the category language, a morphism from one geometry to another should be de-

fined as a mapping of the set of points of one geometry to the set of points of the other that respects the actions of the corresponding transformation groups. More precisely, given two geometries $(G : X)$ and $(H : Y)$, a *morphism* (or an *equivariant map*) is any pair (α, f) consisting of a homomorphism of transformation groups $\alpha : G \to H$ and a mapping of sets $f : X \to Y$ such that

$$\boxed{f(xg) = (f(x))(\alpha(g))} \tag{1.4}$$

for all $x \in X$ and all $g \in G$. This definition is typical of the category approach in mathematics: at first glance, the boxed formula makes no sense at all (no wonder category theory is called abstract nonsense), but actually the definition is perfectly natural.

To see this, let us take any point $x \in X$ and let an arbitrary transformation $g \in G$ act on x, taking it to the point $xg \in X$. Under the map $f : X \to Y$, the point x is taken to the point $f(x) \in Y$ and the point xg is taken to the point $f(xg) \in Y$. How are these two points related? What transformation (if any) takes $f(x)$ to $f(xg)$? Clearly, if the pair of maps (f, α) respects the action of the transformation groups in X and Y, it must be none other than $\alpha(g)$, and this is precisely what the boxed formula says.

To check that the reader has really understood this definition, we suggest that she/he prove that $\alpha(\mathbf{1})=\mathbf{1}$ for any morphism (f, α).

1.4.4. *Isomorphic geometries.* In any mathematical theory, isomorphic objects are those which are equivalent, i.e., are not distinguished in the theory. Thus isomorphic linear spaces are not distinguished in linear algebra, sets of the same cardinality (i.e., sets for which there exists a bijective map) are equivalent in set theory, isomorphic fields are not distinguished in abstract algebra, congruent triangles are the same in Euclidean plane geometry, and so on. What geometries should be considered equivalent? We hope that the following definition will seem natural to the reader.

Two geometries $(X : G)$ and $(Y : H)$ are called *isomorphic*, if there exist a bijection $f : X \to Y$ and an isomorphism $\alpha : G \to H$ such that

$$f(xg) = (f(x))(\alpha(g)) \quad \text{for all} \quad x \in X \quad \text{and all} \quad g \in G.$$

In the definition, the displayed formula is a repetition of relation (1.4), so it expresses the requirement that an isomorphism be a morphism (must satisfy the equivariance condition, i.e., respect the action of the transformation

groups), the conditions on α and f say that they are equivalences, so what this definition is saying is that $(X : G)$ and $(Y : H)$ are the same.

At this stage we have no meaningful examples of isomorphic geometries. They will abound in what follows. For instance, we will see (in Chapter 10) that the Poincaré half-plane model is isomorphic to the Cayley–Klein disk model.

1.4.5. *Subgeometries.* What are subobjects in the category of geometries? The reader who is acquiring a feel for the category language should have no difficulty in coming up with the following definition. A geometry $(G : X)$ is said to be a *subgeometry* of the geometry $(H : Y)$ if X is a subset of Y and G is a subgroup of H.

A closely related definition is the following. An *embedding* (or *injective morphism*) of the geometry $(X : G)$ to the geometry $(Y : H)$ is a morphism (f, α) such that $\alpha : G \to H$ is a monomorphism and $f : X \to Y$ is injective.

Examples of subgeometries and embeddings of geometries can easily be deduced from the examples of subgroups of transformation groups in Subsection 1.3.4.

1.5. Some philosophical remarks

The examples in Section 1.2 (square, cube, circle) were taken from elementary school geometry. This was done *to motivate* the choice of the action of the corresponding transformation group. But now, in the example of the cube, let us forget school geometry: instead of the cube I^3 with its vertices, edges, faces, angles, interior points and other structure, consider the abstract set of points $\{A, B, C, D, A', B', C', D'\}$ and define the "isometries" of the "cube" as a set of 48 bijections; for example, the "rotation by 270°" about the vertical axis is the bijection

$$A \mapsto B,\ B \mapsto C,\ C \mapsto D,\ D \mapsto A,\ A' \mapsto B',\ B' \mapsto C',\ D' \mapsto A',$$

and the 47 other "isometries" are defined similarly. Then (still forgetting school geometry), we can *define* vertices, edges (AB is an edge, but AC' is not), faces, prove that all edges are congruent, all faces are congruent, the "cube" can "rotate" about each vertex, etc.). The result is the *intrinsic geometry of the set of vertices* of the cube.

This geometry is not the same as the geometry of the cube, $(I^3, \mathrm{Sym}(I^3))$, described in Subsection 1.2.3. Of course the group G acting in these two geometries is the *same group* of order 48, but it acts on two *different sets*:

the (infinite) set of points of the cube I^3 and the (finite) set of its 8 vertices $A, B, C, D, A', B', C', D'$. Thus the algebra of the two situations is the same, but the geometry is different. The geometry of the solid cube I^3 is of course much richer than the geometry of the vertex set of the cube. For example, we can define line segments inside the cube, establish their congruence, etc.

Note also that the geometric properties of the cube I^3 *regarded as a subset of Euclidean space* $\mathbb{R}^3$ are richer than its properties coming from its own geometry $(I^3 : \mathrm{Sym}(I^3))$, e.g. segments of the same length inside the cube, which are always congruent in the geometry of $\mathbb{R}^3$, don't have to be congruent in the geometry of the cube!

Another example: the set of three points $\{A, B, C\}$ with two transformations, namely the identity and the "reflection"

$$A \mapsto A,\ B \mapsto C,\ C \mapsto B$$

is of course a geometry in the sense of Klein. What should it be called? An appropriate title, as the reader will no doubt agree, is "the intrinsic geometry of the vertex set of the isosceles triangle".

1.6. Problems

1.1. List all the elements (indicating their orders) of the symmetry group (i.e., isometry group) of the equilateral triangle. List all its subgroups. How many elements are there in the group of motions (i.e., orientation-preserving isometries) of the equilateral triangle.

1.2. Answer the same questions as in Problem 1.1 for

(a) the regular n-gon (i.e., the regular polygon of n sides); consider the cases of odd and even n separately;

(b) the regular tetrahedron;

(c) the cube;

(e)* the dodecahedron;

(f)* the icosahedron;

(g) the regular pyramid with four lateral faces.

1.3. Embed the geometry of the motion group of the square into the geometry of the motion group of the cube, and the geometry of the circle into the geometry of the sphere.

1.4. For what n and m can the geometry of the regular n-gon be embedded in the geometry of the regular m-gon?

1.5. Let G be the symmetry group of the regular tetrahedron. Find all its subgroups of order 2 and describe their action geometrically.

1.6. Let G^+ be the group of motions of the cube. Indicate four subsets of the cube on which G^+ acts by all possible permutations.

1.7. Let G be the symmetry group of the dodecahedron. Indicate subsets of the dodecahedron on which G acts by all possible permutations.

1.8. Find a minimal system of generators for the symmetry group of
(a) the regular tetrahedron;
(b) the cube.

1.9. Describe fundamental domains of the symmetry group of
(a) the cube;
(b) the icosahedron;
(c) the regular tetrahedron.

1.10. Describe the Möbius band as a subset of $\mathbb{R}P^2$.

1.11. Show that the composition of two reflections of the sphere in planes passing through its center is a rotation. Determine the axis of rotation and, if the angle between the planes is given, the angle of rotation.

1.12. Given two rotations of the sphere, describe their composition.

Chapter 2
ABSTRACT GROUPS AND GROUP PRESENTATIONS

In order to study geometries more complicated than the toy models with which we played in the previous chapter, we need to know much more about group theory. Accordingly, in this chapter we present the relevant facts of this theory (they will constantly be used in what follows).

The theory of transformation groups began in the work of several great mathematicians: Lagrange, Abel, Galois, Sophus Lie, Felix Klein, Élie Cartan, Herman Weyl. At the beginning of the 20th century, algebraists decided to generalize this theory to the formal theory of *abstract groups*. In this chapter, we will study this formal theory and learn that it is not a generalization at all: Cayley's Theorem (which concludes this chapter) says that all abstract groups are actually transformation groups. We will also learn that two important classes of groups (*free groups* and *permutation groups*) have certain universality properties. Finally, we will find out how *to present groups* by means of generators and relations; this allows to replace computations with groups by games with words.

2.1. Abstract groups

2.1.1. *Groups: definition and manipulation.* By definition, an (abstract) *group* is a set G of arbitrary elements supplied with a binary operation $*$ (usually called *multiplication*) if it obeys the following rules:

- (*neutral element axiom*) there exists a unique element $e \in G$ such that $g * e = e * g = g$ for any $g \in G$;
- (*inverse element axiom*) for any $g \in G$ there exists a unique element $g^{-1} \in G$, called *inverse to* g, such that $g * g^{-1} = g^{-1} * g = e$;
- (*associativity axiom*) $(g * h) * k = g * (h * k)$ for all $g, h, k \in G$.

A group $(G, *)$ is called *commutative* or *Abelian* if $g * h = h * g$ for all $g, h \in G$ (in that case the operation is usually called a *sum* and the inverse element is usually denoted by $-g$ instead of g^{-1}).

Note that the elements of an abstract group can be objects of any nature, they are not necessarily bijections of something and the operation $*$ is not

necessarily composition, while the notation g^{-1} for inverse elements is purely formal, it does not mean that g^{-1} is the inverse of a bijection.

The three axioms for groups listed above are much stronger than necessary. For example, the uniqueness condition in the inverse element axiom can be omitted without changing the class of objects defined by these axioms. The definition can be weakened further, but this is an not important fact from the point of view of geometry, so we do not dwell on it further.

The group axioms have some obvious consequences that are useful when performing calculations with elements of groups. In these calculations and further on, we omit the group operation symbol, i.e., we write gh instead of $g * h$.

The first immediate consequence of the group axioms are the *left* and *right cancellation rules*, which say that one can cancel equal terms on the two sides of an equations, provided they both appear at the left (or at the right) of the corresponding expression, i.e.,

$$\forall g, h, k \in G \quad gh = gk \Longleftrightarrow h = k, \quad hg = kg \Longleftrightarrow h = k.$$

The implications in these formulas are two-sided; reading them from right to left, we can say that one can multiply both sides of an equation by the same element *from the same side*. The phrase in italics is of course important, because for non-Abelian groups the cancellation of equal terms on different side of an equation can result in a false statement.

Another simple but important consequence of the axioms is the *rule for solving linear equations*, i.e.,

$$\forall g, h, x \in G \quad gx = h \Longleftrightarrow x = g^{-1}h, \quad xg = h \Longleftrightarrow x = hg^{-1},$$

which are proved by multiplying both sides by the element g^{-1} (it exists by the inverse element axiom) from the left and the right, respectively, using associativity and the neutral element axiom.

These two rules are constantly used when performing manipulations with equations in groups, as the reader will see in solving some of the exercises at the end of this chapter.

2.1.2. *Examples of groups.* It is easy to see that any transformation group is a group. Indeed, the three axioms of abstract groups listed above, although they do not appear explicitly in the definition of transformation groups, hold automatically for the latter, because their elements are not

arbitrary objects, they are bijections, the multiplication operation is not arbitrary (it is composition): for them associativity and the neutral element axiom hold automatically.

Here are some other important examples of groups.

(i) *The standard numerical groups*: the integers under addition $(\mathbb{Z}, +)$, as well as the rational, real, and complex numbers under addition $(\mathbb{Q}, +)$, $(\mathbb{R}, +)$, and $(\mathbb{C}, +)$; the nonzero rational, real, and complex numbers under multiplication $(\mathbb{Q} \setminus \{0\}, \times)$, $(\mathbb{R} \setminus \{0\}, \times)$, and $(\mathbb{C} \setminus \{0\}, \times)$. Note that the nonzero integers under multiplication are *not* a group (no inverse elements!), neither are the natural numbers $\mathbb{N}$ under addition (for the same reason). Another nice numerical group is formed by the *unimodular complex numbers* under multiplication $\mathbb{S}^1 := \{z \in \mathbb{C} : |z| = 1\}$.

(ii) *The group of residues modulo* m, $(Z_m, \oplus)$ (also known as the m-element *cyclic group*); its elements are the m infinite sets of integers that have the same remainder under division by the natural number m; we denote these sets by $\overline{0}, \overline{1}, \dots \overline{m-1}$; their sum $\oplus$ is defined by

$$\overline{i} \oplus \overline{j} := \overline{(i+j) \bmod m},$$

where $(\cdot) \bmod m$ stands for the remainder under division by m. The sum operation $\oplus$ is well-defined, i.e., does not depend on the choice of the representatives i and j in the classes $\overline{i}$ and $\overline{j}$. Indeed, if we take $i + rm$ instead of i and $j + sm$ instead of j, then

$$\overline{(i+rm)+(j+sm)} = \overline{(i+j+(r+s)m)} = \overline{(i+j)}.$$

(iii) *The group of permutations of* n *objects* S_n: its elements are bijections of a set of n elements that we denote by natural numbers $(\{1, 2, \dots n\})$; we will write bijections $s \in S_n$ in the form

$$s = [i_1, i_2, \dots, i_n], \quad \text{where} \quad i_1 = s(1),\ i_2 = s(2), \dots i_n = s(n);$$

multiplication in S_n is the composition of bijections. This group is extremely important not only in geometry, but also in linear algebra, combinatorics, representation theory, mathematical physics, etc. We will come back to permutation groups later in this chapter.

(iv) *The free group* $\mathbb{F}_n = \mathbb{F}(a_1, \dots, a_n)$ *on* n *generators*; its elements are equivalence classes of words and the group operation is concatenation; a detailed definition of $\mathbb{F}_n$ appears in Subsection 2.6.2 below.

(v) *The group $GL(n)$ of nondegenerate linear operators on $\mathbb{R}^n$*; its elements are n by n matrices with nonzero determinant and the group operation is matrix multiplication (or, which is the same thing, composition of operators).

(vi) *The groups of orthogonal and special orthogonal operators on $\mathbb{R}^n$*, standardly denoted by $O(n)$ and $SO(n)$. We assume that the reader is familiar with the groups $GL(n)$, $O(n)$, and $SO(n)$ at least for $n = 2$ and $n = 3$; if this is not the case, he/she is referred to Appendix I.

2.1.3. *Order of a group and of its elements, generators.* The notions of *order* (of elements of a group and of the group itself) and of *generator* for abstract groups are defined exactly as for transformation groups (see 1.3). In this book, $|G|$ denotes the *order of the group* G (i.e., the number of its elements), $\mathrm{ord}(g)$ denotes the *order of the element* $g \in G$, i.e. the least natural number k such that $g^k = e$. For example: $|\mathbb{Z}_5| = 5$; $|\mathrm{Sym}(\bigcirc)| = \infty$; for $\overline{3} \in \mathbb{Z}_{15}$, we have $\mathrm{ord}(\overline{3}) = 5$; for any nonzero real number x in the additive group $\mathbb{R}$, we have $\mathrm{ord}(x) = \infty$.

A family of *generators* of a group G is a (finite or infinite) set of its elements $g_1, g_2, \dots$ in terms of which any element g of G can be expressed, i.e., written in the form $g = g_1^{\varepsilon_1} \dots g_k^{\varepsilon_k}$, where the ε_i's equal ± 1 and g_i^{+1} stands for g_i. For example, any nonzero element of $\mathbb{Z}_p$, where p is prime, constitutes a (one-element) family of generators for $\mathbb{Z}_p$, while $\mathrm{Sym}(\bigcirc)$ does not have a finite family of generators. If g is an element of order m of a group G, then the set $\{g, g^2, \dots g^{m-1}, g^m = e\}$ is also a group (it is a "subgroup" of G, see the definition in 2.3.1), and its order is m. This justifies the use of the same term "order" for groups and their elements, i.e., for notions that seem very different at first glance.

2.2. Morphisms of Groups

In accord with the traditions of the category language, as soon as we have defined an interesting class of objects, in this case groups, we should define their morphisms.

2.2.1. *Definitions.* Suppose $(G, *)$ and $(H, \star)$ are groups; a mapping $\phi : G \to H$ is called a *homomorphism* (or a *morphism of groups*) if it respects the operations, i.e., $\varphi(g_1 * g_2) = \varphi(g_1) \star \phi(g_2))$. Thus the inclusion $\mathbb{Z} \to \mathbb{R}$, $n \mapsto n$, is a morphism, while the inclusion $(\mathbb{Q} \setminus \{0\}, \times) \to (\mathbb{Q}, +)$ is not (the operations are not respected, e.g. $2 \times 3 \neq 2 + 3$).

By definition, a homomorphism φ is a *monomorphism* (respectively, an *epimorphism* or an *isomorphism*) if the mapping φ is injective (resp., surjective or bijective). From the point of view of abstract algebra, isomorphic groups are identical.

2.2.2. *Examples.* The group $\mathrm{Sym}(\triangle)$ of isometries of the equilateral triangle is isomorphic to the permutation group S_3, the group $\mathrm{Sym}(\bigcirc)$ is isomorphic to $\mathrm{SO}(2)$; there are obvious monomorphisms of the rotation group $\mathrm{Rot}(\square)$ into $\mathrm{SO}(2)$ and of $\mathbb{Z}_3$ into $\mathbb{Z}_{15}$; there is a no less obvious epimorphism of $\mathbb{Z}$ onto $\mathbb{Z}_{17}$.

2.3. Subgroups

Worthwhile mathematical objects should not only be related by morphisms, they should have naturally defined subobjects.

2.3.1. *Definitions and examples.* A *subgroup* H of a group G is a subset of G which satisfies the group axioms. Note that in order to check that H is a subgroup of G, it is not necessary to verify all the group axioms: it suffices to check that H is closed under the group operation and under taking inverses. Any group G has at least two subgroups: the one-element subgroup consisting of the neutral element $e \in G$ and the group G itself. These two subgroups are sometimes called *trivial*, and of course in the study of the structure of groups we are interested in *nontrivial* subgroups.

Examples: $\mathrm{Rot}(\bigcirc)$ is a subgroup of $\mathrm{Sym}(\bigcirc)$, the set $\{[1234], [2134]\}$ is a subgroup of S_n, the set $\{\overline{0}, \overline{5}, \overline{10}\}$ is a subgroup of $\mathbb{Z}_{15}$, while $\{\overline{0}, \overline{5}, \overline{11}\}$ is not.

2.3.2. *Partition of a group into cosets.* If H is a subgroup of G, then the (*left*) *coset* $gH \subset G$, for some $g \in G$, is the set of all elements of the form gh for $h \in H$. Right cosets Hg are defined similarly. Right cosets as well as *left cosets form a partition of the set of elements of a group*, i.e., two cosets either do not intersect or coincide.

To prove this, it suffices to show that if two cosets have a common element $\overline{g} \in g_1H \cap g_2H$, then any element of g_1H belongs to g_2H and vice versa. So suppose that $\widetilde{g} \in g_1H$ (which means that $\widetilde{g} = g_1\widetilde{h}$ for some $\widetilde{h} \in H$); we must show that $\widetilde{g} \in g_2H$, i.e., we must find an $h_x \in H$ such that $\widetilde{g} = g_2h_x$.

Since $\overline{g} \in g_1H \cap g_2H$, there exist elements $\overline{h}_1, \overline{h}_2 \in H$ for which we have $g_1\overline{h}_1 = \overline{g} = g_2\overline{h}_2$, which implies that $g_1 = g_2\overline{h}_2(\overline{h}_1)^{-1}$. Now we can write

$$\widetilde{g} = g_1\widetilde{h} = g_2\overline{h}_2(\overline{h}_1)^{-1}\widetilde{h} = g_2\left(\overline{h}_2(\overline{h}_1)^{-1}\widetilde{h}\right) = g_2h_x,$$

where we have defined h_x as $\overline{h}_2(\overline{h}_1)^{-1}\widetilde{h}$, and since h_x belongs to H (as the product of elements of H), we have proved that $\widetilde{g} \in g_1H \Longrightarrow \widetilde{g} \in g_2H$. The reverse implication is proved by a symmetric argument (interchange the indices 1 and 2).

Thus we have obtained the partition of G into left cosets. The partition into right cosets is obtained similarly.

Note also that all cosets have the same number of elements (finite or infinite), because there is an obvious bijection between any coset and the subgroup H. This bijection for left cosets is given by $gH \ni gh \mapsto h \in H$.

2.4. The Lagrange Theorem

The corollary to the elementary theorem proved below is the first structure theorem about abstract groups. It was proved (for transformation groups) almost two centuries ago by Lagrange.

Theorem 2.4.1. *If H is a subgroup of a finite group G, then the order of H divides the order of G.*

Proof. The cosets of H in G form a partition of the set of elements of G (see 2.3.2) and all have the same number of elements as H. □

Corollary 2.4.2. *Any group G of prime order p is isomorphic to $\mathbb{Z}_p$.*

Proof. Let $g \in G$, $g \neq e$. Let m be the smallest positive integer such that $g^m = e$. Then it is easy to see hat $H := \{e, g, g^2, \ldots, g^{m-1}\}$ is a subgroup of G. By Theorem 1, m divides p. This is impossible unless $m = p$, but then $H = G$ is obviously isomorphic to $\mathbb{Z}_p$. □

2.5. Quotient groups

Nice mathematical objects often have naturally defined "quotient objects" obtained by "dividing out" the given object by some subobject (examples that may be familiar to the reader are quotient spaces in linear algebra). The construction of "quotient groups" along those lines works only when the subgroup used is in a sense "nice", and we begin by defining such subgroups.

2.5.1. *Normal subgroups.* A subgroup $H \subset G$ is *normal* if $gHg^{-1} = H$ for any $g \in G$, i.e., for any $h \in H$ and any $g \in G$ we have $g^{-1}hg \in H$.

An example of a normal subgroup is the set $\{\overline{0}, \overline{5}, \overline{10}\}$ in $\mathbb{Z}_{15}$. More generally, any subgroup of an Abelian group is (obviously!) normal.

To see an example of a subgroup which is *not* normal, consider the subset $D := \{e = [1, 2, 3, 4], [2, 1, 3, 4]\}$ in the permutation group S_4. The set D is

obviously a subgroup (isomorphic to $\mathbb{Z}_2$) of S_4, but it is not normal, because

$$[4,1,2,3]\,[2,1,3,4]\,[2,3,4,1] = [1,3,2,4] \notin D.$$

2.5.2. *Construction of quotient groups.* If H is a normal subgroup of G, there is a well-defined operation in the family of cosets: the product of two cosets is the coset containing the product of any two elements of these cosets. For left cosets this may be written as $g_1H\,g_2H := g_1g_2H$.

To prove that this is an operation well-defined on cosets, we must show that if we replace g_1 by another element $\overline{g}_1$ from Hg_1 and replace g_2 by another element $\overline{g}_2$ from Hg_2, then $\overline{g}_1H\,\overline{g}_2H = g_1H\,g_2H$. Without loss of generality, it suffices to consider the case in which only one of the two elements is replaced, say g_1. Then we have $\overline{g}_1 = g_1\overline{h}_1$ for some $\overline{h}_1 \in H$. We must prove that $\overline{g}_1g_2 \in g_1g_2H$, i.e., that there exists an h_x such that $\overline{g}_1g_2 = g_1g_2h_x$. Replacing $\overline{g}_1$ by its expression $g_1\overline{h}_1$ (see above), we can rewrite the previous equation as

$$g_1\overline{h}_1g_2 = g_1g_2h_x.$$

Solving this (linear) equation for h_x, we obtain $h_x = g_2^{-1}\overline{h}_1g_2$. Recall that H is normal, therefore the right-hand side of the previous equality is an element of H. Thus we have found the required element $h_x \in H$, thereby proving that the product of cosets is well defined.

The family of cosets supplied with this product operation is called the *quotient group* of G by H and is denoted by G/H. It is easy to show that G/H satisfies the axioms for groups.

Example: in the additive group of integers $(\mathbb{Z},+)$, elements of the form $5k$, $k \in \mathbb{Z}$, constitute a normal subgroup (of infinite order), denoted $5\mathbb{Z}$; the corresponding quotient group $\mathbb{Z}/5\mathbb{Z}$ is isomorphic to the group $\mathbb{Z}_5$.

2.6. Free groups and permutations

In this section, we study two classes of groups: the free groups (which have the "least structure") and the permutation groups (which have the "most structure").

2.6.1. *Free groups.* Let $\{a_1,\dots,a_k\}$ be a set of symbols. Then the set of formal symbols (called *letters*)

$$A := \{e, a_1, \dots, a_k, a_1^{-1}, \dots, a_k^{-1}\}$$

will be our *alphabet*. A string of letters from our alphabet will be called a *word*. Two words w_1 and w_2 are called *equivalent*, if one can be obtained from the other by using the following *trivial relations* $a_i a_i^{-1} = a_i^{-1} a_i = e$ for any i and $ae = ea = a$ for any $a \in A$; for example

$$a_1 a_3^{-1} \sim a_1 a_3^{-1} e \sim a_1 a_3^{-1} a_2 a_2^{-1} \sim a_1 a_3^{-1} a_2 e a_2^{-1}.$$

The *product* of two words is defined as their concatenation (i.e., the result of writing then one after the other). The *free group* with generators $a_1, \dots, a_k$ is defined as the set of equivalence classes of words supplied with the product (concatenation) operation and is denoted by $\mathbb{F}_k = \mathbb{F}[a_1, \dots, a_k]$. The fact that concatenation is well-defined on the equivalence classes (i.e., the concatenation of equivalent elements produces an element from the same equivalence class) is obtained by an straightforward verification that we omit.

For example, $\mathbb{F}[a]$ is isomorphic to $(\mathbb{Z}, +)$, while $\mathbb{F}[a_1, a_2]$ is not commutative.

2.6.2. *Permutation groups.* The *permutation group* S_n on n objects was defined in 2.1.2 as the family of all bijections of the set $\{1, 2, \dots, n\}$ supplied with the operation of composition; S_n consists of $n! = 1 \cdot 2 \cdot \dots \cdot n$ elements denoted by $[i_1, \dots, i_n]$, where $i_k := \beta(k)$ and β is the bijection defining the given permutation.

Geometrically, the permutation group S_3 can be interpreted as the isometry group $\mathrm{Sym}(\triangle)$ of the equilateral triangle, while S_4 is isomorphic to the isometry group of the regular tetrahedron (as we shall see in the next chapter).

2.6.3. *Universality theorem.* It turns out that permutation groups and free groups have important "universality" properties.

Theorem 2.6.3. (i) *For any finite group G there exists a monomorphism of G into S_n for some n.*

(ii) *For any group G with a finite number n of generators there exists an epimorphism of the free group $\mathbb{F}_n$ onto G.*

Proof. (i) Let $|G| = n$ and $g_0 \in G$; then the mapping

$$\beta_{g_0} : G \to S_n \quad \text{given by} \quad G \ni g \mapsto g g_0 \in G$$

is a monomorphism. Indeed, it is obviously a homomorphism (indeed, we have $\beta_{g_0} \beta_{g_1} = \beta_{g_0 g_1}$, because both maps are given by the rule $g \mapsto g g_0 g_1$).

The homomorphism β_{g_0} is injective, because $g_0 g = g_0 g'$ implies $g = g'$ by the cancellation rule.

(ii) Let $g_1, \ldots, g_n$ be a set of generators of G. Then the mapping

$$\alpha : \mathbb{F}[a_1, \ldots, a_n] \to G \quad \text{given by} \quad \alpha(a_i) = g_i,\ i = 1, \ldots, n$$

is obviously a homomorphism. It is also surjective, because to each element $g_{i_1}^{\varepsilon_1} g_{i_2}^{\varepsilon_2} \ldots g_{i_m}^{\varepsilon_m} \in G$, where the ε_i's are equal to ± 1, the mapping α takes the element $a_{i_1}^{\varepsilon_1} a_{i_2}^{\varepsilon_2} \ldots a_{i_m}^{\varepsilon_m} \in \mathbb{F}[a_1, \ldots, a_n]$. $\square$

2.7. Group presentations

A presentation of a group is a way of defining the group by means of equations (called *defining relations*) in the generators of the group. This reduces concrete calculations in the group to the formal editing of words according to simple rules. The formal definition of the notion of group presentation is easy to state but perhaps difficult to grasp, so we begin with some examples.

2.7.1. *Examples of group presentations.* (i) Consider all words in the three-letter alphabet $\{e, a, a^{-1}\}$, i.e., expressions such as $eaa^{-1}aae$, $a^{-1}aeaaa$, etc. Let us say that two words are *equivalent* if one can transform one word into another by means of the *trivial relations* $aa^{-1} = e = a^{-1}a$ and $ae = ea = a$ and the relation $a^5 = 1$ (as usual, a^5 stands for $aaaaa$). This is obviously an equivalence relation in the technical sense, i.e. it is reflexive, symmetric, and transitive, so that the set of all words splits into equivalence classes. Define the product of two equivalence classes as the class containing the concatenation of any two elements of the given classes. It is easy to see that this product is well defined, i.e., does not depend on the choice of representatives in the classes. Obviously, there will be 5 equivalence classes (determined by the elements $a, a^2, a^3, a^4, a^5 = e$) and they form a group under the product operation defined above. The group obtained is clearly isomorphic to $\mathbb{Z}_5$.

(ii) Now consider words in the five-letter alphabet $\{e, s_1^{\pm}, s_2^{\pm 1}\}$. Let us say that two words are *equivalent* if one can be transformed into the other by means of the *trivial relations* (which we won't write out again) and the relations $s_1^2 = s_2^2 = e$ and $s_1 s_2 s_1 = s_2 s_1 s_2$ (the latter is known as the *Artin relation*). Defining the product of the corresponding equivalence classes as in the previous example, we obtain a group which is isomorphic to S_3 (see Exercise 2.9 below).

2.7.3. *Formal definition.* The definition of a group presentation is the following. An expression of the form

$$G = \langle g_1, \dots, g_n \ : R_1, \dots, R_k \rangle,$$

where $R_1, \dots R_k$ are words in the alphabet $A = \{g_1, \dots, g_n g_1^{-1}, \dots, g_n^{-1}\}$, is called a *presentation* of the group G; the words R_j are called *relators*; the group G is defined by its presentation as the quotient group

$$\mathbb{F}[g_1, \dots, g_n]/\{R_1, \dots, R_k\},$$

where $\{R_1, \dots, R_k\}$ is the minimal (by inclusion) normal subgroup of the free group $\mathbb{F}[g_1, \dots, g_n]$ containing the elements (relators) $R_1, \dots, R_k$.

This formal definition may be difficult to understand. But the notion of group presentation is simple. The elements of the group G that it defines are words in the alphabet A defined up to the trivial relations (see 2.6.1 above) and up to all the *defining relations* $R_1 = e, \dots, R_k = e$; the product is concatenation (and is well defined).

Here are some examples:

(i) $\mathbb{Z}_m = \langle a \ : a^m \rangle$ is the m-element cyclic group;

(ii) $\mathbb{F}[g_1, \dots, g_n] = \langle g_1, \dots, g_n \ : \ \rangle$ is the free group on n generators (nothing appears after the colon in the angle brackets because the free group has no defining relations);

(iii) the permutation group on four elements can be presented as $S_4 = \langle s_1, s_2, s_3 : s_1^2, s_2^2, s_3^2, \ s_1 s_2 s_1^{-1} s_2^{-1}, \ s_1 s_2 s_1 s_2^{-1} s_1^{-1} s_2^{-1}, \ s_2 s_3 s_2 s_3^{-1} s_2^{-1} s_3^{-1} \rangle$.

More details and examples appear in the problem section of this chapter.

2.8. Cayley's theorem

The following theorem (due to the British mathematician Arthur Cayley) shows that the notion of abstract group is not a real generalization: all groups are in fact transformation groups!

Theorem 2.8.1. *Any group G is a transformation group acting on the set G by right multiplication: $g \mapsto gg_0$ for any $g_0 \in G$.*

Proof. First, must show that the assignment $g \mapsto gg_0$ is a bijection for any $g_0 \in G$. But this is obvious: it is injective (by the cancellation rule) and surjective (to any element $h \in G$ the element g_0 assigns the element hg_0^{-1}). Further we must verify the transformation group axioms (see 1.3.1). This

verification is also obvious: the transformations defined by elements of G are closed under composition (because so are elements of G) and under taking inverse elements (the transformation inverse to the one given by g_0 is the one given by g_0^{-1}). $\square$

Corollary 2.8.2. *Any group is a geometry in the sense of Klein* (*i.e., in the sense of formal definition given in* 1.4.1).

This corollary shows (as we mentioned previously) that the definition of geometry given in 1.4.1 is of course too general; additional restrictions on the set of elements and the transformation group are needed to obtain an object about which most mathematicians will agree that it is a *bona fide* geometry. However, there seems to be no formal agreement on this subject, so that the "additional restrictions" to be imposed are a matter of opinion, and we will not specify any (at least on the formal level) in this course.

2.7. Problems

2.1. Describe all the finite groups of order 6 or less and supply each with a geometric interpretation.

2.2. Describe all the (nontrivial) normal subgroups and the corresponding quotient groups of

(a) the isometry group of the equilateral triangle;

(b) the isometry group of the regular tetrahedron.

2.3. Let G be the motion group of the plane, P its subgroup of parallel translations, and R its subgroup of rotations with fixed center O. Prove that the subgroup P is normal and the quotient group G/P is isomorphic to R.

2.4. Prove that if the order of a subgroup is equal to half the order of the group (i.e., the subgroup is of *index* 2), then the subgroup is normal.

2.5. Find all the orbits and stabilizers of all the points of the group $G \subset S_{10}$ generated by the permutation $[5, 8, 3, 9, 4, 10, 6, 2, 1, 7] \in S_{10}$ acting on the set $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$.

2.6. Find the maximal order of elements in the group (a) S_5; (b) S_{13}.

2.7. Find the least natural number n such that the group S_{13} has no elements of order n.

2.8. Prove that the permutation group S_n is generated by the transposition $(1\,2) := [2, 1, 3, 4, \ldots, n]$ and the cycle $(1\,2\ldots n) := [2, 3, \ldots, n, 1]$.

2.9. Present the symmetry group of the equilateral triangle by generators and relations in two different ways.

2.10. How many homomorphisms of the free group in two generators into the permutation group S_3 are there? How many of them are epimorphisms?

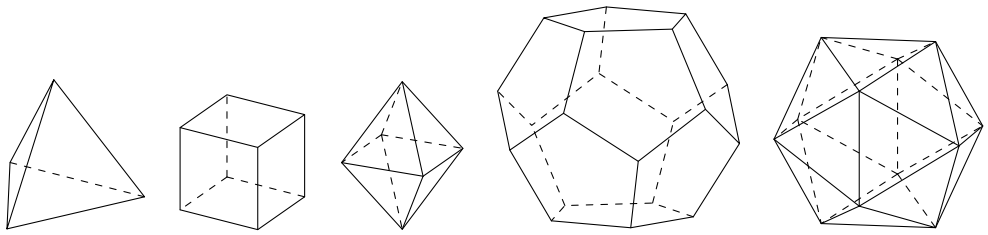
2.11. Prove that the group presented as $\langle a, b \mid a^2 = b^n = a^{-1}bab = 1\rangle$ is isomorphic to the dihedral group $\mathbb{D}_n$ (defined in Chapter 3).

2.12. Show that if the elements a and b of a group satisfy the relations $a^5 = b^3 = 1$ and $b^{-1}ab = a^2$, then $a = 1$.

Chapter 3

FINITE SUBGROUPS OF SO(3) AND THE PLATONIC BODIES

This chapter is devoted to the classification of regular polyhedra (the five "Platonic bodies") pictured below:



The proof of the classification theorem given here is based on group theory, more precisely on the study of finite subgroups of the isometry group of the two-dimensional sphere.

3.1. The Platonic bodies in art, philosophy, and science

The perfection of the shape of regular polyhedra attracted the great artist and thinker Leonardo da Vinci, who pictured them in various media. Figure 3.1 reproduces his engravings of two of them.

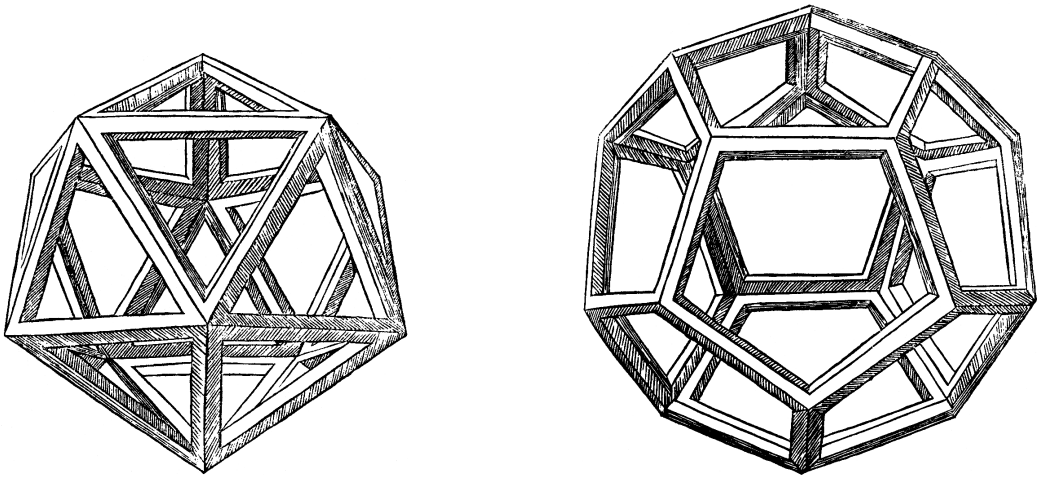


Figure 3.1. Da Vinci engravings: the icosahedron and dodecahedron

Some philosophers and scientists felt an almost mystical attraction to these amazingly symmetric shapes. Thus the great astronomer Kepler believed that the distances from the planets to the Sun could be calculated from a system of nested inscribed Platonic bodies (see his weird engraving reproduced in Fig.3.2).

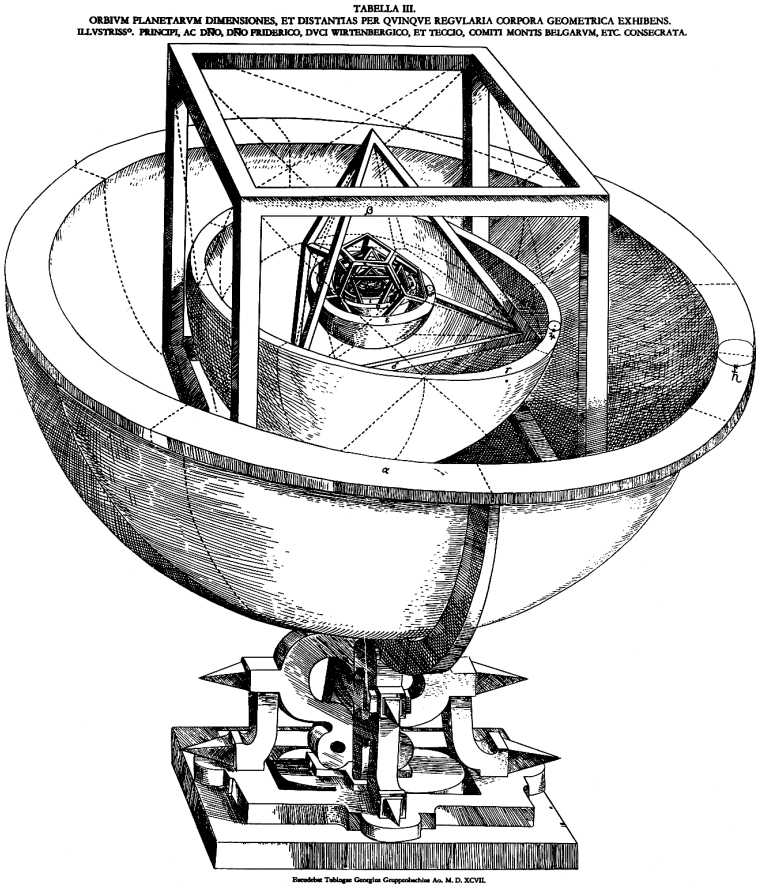


Figure 3.2. Kepler's theory of planetary orbits

The engraving shows a cube inscribed in a sphere, then a smaller sphere inscribed in the cube, a tetrahedron inscribed in that second sphere, a third

sphere inscribed in the tetrahedron, followed by successively inscribed sphere, dodecahedron, sphere, octahedron, sphere, icosahedron. Kepler claimed that the distances from the five planets to the Sun were proportional to the distances from the vertices of the five nested polyhedra to their common center of symmetry. He regarded this "discovery" as his main scientific achievement, far more important than the three fundamental astronomical laws that bear his name. Fortunately for his self-esteem, he did not live to see the day when more exact measurements of the distances between the Sun and the planets showed that Kepler's theory was erroneous.

The five regular polyhedra were known to the ancient Greeks, in particular to the philosopher Plato, who expressed his admiration for their unique perfection so beautifully that today they are often called "Platonic bodies". Of course Plato cannot be credited with their discovery (they were known before his time), but who the actual discoverers were is not clear. It is also unclear whether the ancient Greeks had a proof of the fact that there are no other regular polyhedra, or indeed felt that such a proof was necessary. We can only conjecture that Archimedes had such a proof, or that it was possibly known to the Pythagorean school.

We do know that Pythagoras was interested in the regular polyhedra in connection with his theory of the "singing spheres". In the 20th century, his theory was revived in the work of the German physicist Heisenberg, but the relevant ideas lie outside the scope of a mathematical textbook.

3.2. Finite subgroups of $\mathbf{SO}(3)$

As we mentioned above, the main goal of this chapter is to prove that the only regular three-dimensional polyhedra are the five Platonic bodies. The proof that we give here is essentially group theoretic (we reduce the classification problem of regular polyhedra to classifying finite subgroups of the special orthogonal group SO(3), or, which is the same thing, the group of motions of the sphere $\mathbb{S}^2$). This proof is quite natural and more geometric, in a deeper sense, than the tedious and eclectic space geometry proof anterior to the appearance of the notion of transformation group in mathematics.

Let us return to the geometry (briefly studied in Chapter 1, see 1.1.6) of the two-dimensional sphere

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

defined by the action of its isometry group $\mathrm{Sym}\,(\mathbb{S}^2)$. (In linear algebra courses this group is defined in a different (but equivalent) way, is called the

orthogonal group, and usually denoted by O(3).) Here we will be dealing with the subgroup of $\mathrm{O}(3) = \mathrm{Sym}\,(\mathbb{S}^2)$ consisting of rotations, namely the group $\mathrm{Rot}(\mathbb{S}^2)$ each element of which is a rotation of the sphere about some axis passing through the origin by some angle ϕ, $0 \le \phi < 2\pi$. In linear algebra courses this group is defined in a different (but equivalent) way, is called the *special orthogonal group* and is usually denoted by SO(3).

Our goal is to find the finite subgroups of SO(3). We begin with some examples of finite subgroups of O(3) and SO(3).

3.2.1. *The monohedral group $\mathbb{Z}_n$ for any $n \ge 2$.* Its n elements are rotations about an axis by angles of $2k\pi/n$, where $k = 0, \dots, n-1$.

3.2.2. *The dihedral group $\mathbb{D}_n$ for any $n \ge 2$.*

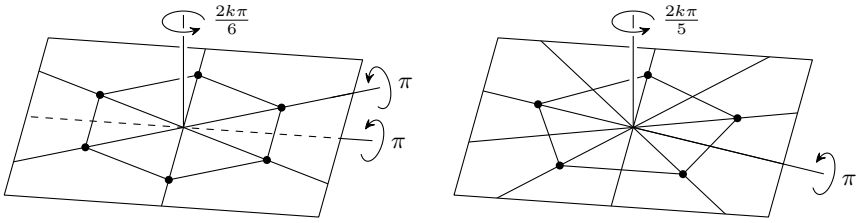


Figure 3.3. The dihedral group $\mathbb{D}_n$ for $n = 6$ and $n = 5$

This $2n$-element group is the isometry group of the regular n-gon (lying in the horizontal plane Oxy and inscribed in the sphere $\mathbb{S}^2$); $\mathbb{D}_n$ consists of n rotations (by angles of $2k\pi/n$, $k = 0, 1, \dots, n-1$) and n reflections in the horizontal lines passing through the center of the sphere, the vertices, and the midpoints of the sides (be careful: these lines are different when n is even or odd – look at the figure!). Note that the reflections of $\mathbb{D}(n)$ in the horizontal lines are actually *rotations* by 180° in space about these lines.

3.2.3. *The isometry group of the regular tetrahedron.* It consists of 24 elements, it is denoted by $\mathrm{Sym}(\Delta^3)$ and its (12 element) rotation subgroup is:

$$\mathrm{Rot}(\Delta^3) = \mathrm{Sym}^+(\Delta^3) \subset \mathrm{Sym}(\Delta^3);$$

$\mathrm{Sym}(\Delta^3)$ consists of 8 rotations about 4 axes (containing one vertex) by angles of $2\pi/3$ and $4\pi/3$, of three rotations by π about axes joining the midpoints of opposite edges and of the identity. It is easy to see that $\mathrm{Sym}(\Delta^3)$ is isomorphic to the permutation group S_4. But here we think of this group geometrically, regarding the tetrahedron as inscribed in the sphere $\mathbb{S}^2$ and the elements of $\mathrm{Sym}(\Delta^3$ as acting on the sphere as well.

3.2.4. *The isometry group* $\mathrm{Sym}(I^3)$ *of the cube.* It has 48 elements (see 1.2.3); its rotation subgroup consists of 24 elements:

$$\mathrm{Rot}(I^3) = \mathrm{Sym}^+(I^3) \subset \mathrm{Sym}(I^3).$$

If we join the center of each of the 6 faces of the cube by segments to the four neighboring centers, we obtain the carcass of the *octahedron* dual to the cube (see Fig.3.4). The octahedron has 6 vertices and 8 triangular faces; its isometry group is obviously the same as that of the cube.

3.2.5. *The isometry group* $\mathrm{Sym}(\mathrm{Dod})$ *of the dodecahedron.* It has 120 elements and possesses a (60 element) rotation subgroup:

$$\mathrm{Rot}(\mathrm{Dod}) = \mathrm{Sym}^+(\mathrm{Dod}) \subset \mathrm{Sym}(\mathrm{Dod}).$$

The dodecahedron is the (regular) polyhedron (inscribed in the sphere $\mathbb{S}^2$) with 12 faces (congruent regular pentagons), 30 edges, and 20 vertices (see Fig.3.4). The existence of such a polyhedron will be proved at the end of this chapter. Joining the centers of the faces of the dodecahedron having a common edge (look at Fig.3.4 again), we get the *icosahedron* dual to the dodecahedron; it has 20 faces, 30 edges, and 12 vertices. Its transformation group is the same as that of the dodecahedron.

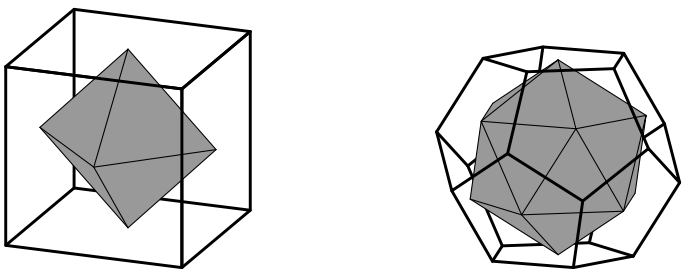


Figure 3.4 Dual pairs of regular polyhedra

The following theorem states that SO(3) has no finite subgroups other than those listed above.

3.2.6. Theorem *Any finite nontrivial subgroup of* $G^+ \subset \mathrm{Sym}^+(\mathbb{S}^2)$=$SO(3)$ *is isomorphic to one of the following groups:*

(i) $\mathbb{Z}_n$, $n \geq 2$, (ii) $\mathbb{D}_n$, $n \geq 2$, (iii) $\mathrm{Rot}(\Delta^3)$, (iv) $\mathrm{Sym}^+(I^3)$, (v) $\mathrm{Sym}^+(Dod)$.

Proof. We know that any element of SO(3) (and hence of G^+) is a rotation about a diameter of the sphere $\mathbb{S}^2$ and has two fixed points (the ends of the diameter). Let F be the set of fixed points of the group G^+:

$$F = \big\{x \in \mathbb{S}^2 \mid \exists g \in G^+ - \mathrm{id},\ xg = x\big\}.$$

For example, for the group $\mathbb{Z}_n$, F consists of two points, while for the rotation group $\mathrm{Rot}(\Delta^3)$ of the tetrahedron it has 14, namely the 4 vertices, the 4 intersection points of the axes of rotations of the faces with the sphere, and the 6 intersection points of the three axes of rotations passing through the midpoints of opposite edges of the tetrahedron.

Consider the (finite) geometry $(F : G^+)$ and let A be a set containing one point in each orbit of G^+ in F. First we claim that the number of points in F is

$$|F| = |A||G^+| - 2(|G^+| - 1).$$

The proof of this fact is the object of Exercise 3.3 at the end of the present chapter. Using the class formula (1.2), we can write

$$|F| = \sum_{x \in A} \frac{|G^+|}{v(x)}, \quad \text{where} \quad v(x) := |\mathrm{St}(x)|$$

Note that $v(x)$ is the order of the rotation subgroup of G^+ generated by the rotations about the axis containing x and its antipodal point. Replacing $|F|$ by its value found above and dividing by $|G^+|$, we obtain

$$\boxed{2 - \frac{2}{|G^+|} \;=\; \sum_{x \in A} \Big(1 - \frac{1}{v(x)}\Big)}, \tag{3.1}$$

or solving for $|G^+|$,

$$|G^+| = \Big[1 - \frac{1}{2} \cdot \sum_{x \in A} \Big(1 - \frac{1}{v(x)}\Big)\Big]^{-1}. \tag{3.2}$$

The left-hand side of the boxed formula is less than 2, but greater than or equal to 1; hence so is the sum in the right-hand side, and thus the summation over A cannot contain 4 summands or more (because $v(x) \geq 2$); therefore there can be only 2 or 3 orbits of the action of G^+ on F.

First let us consider the case in which $|F| = 2$, i.e., when there is only one rotation axis (with intersection points x_1 and x_2 with the sphere). In that

case there are two orbits in F, each consisting of one point, namely $\{x_1\}$ and $\{x_2\}$. Is such a situation possible? Of course it is, but only if G^+ consists of rotations about the unique axis x_1x_2. But then it follows that $G^+ \cong \mathbb{Z}_n$ for some $n \geq 2$. So the theorem is proved for the case $|F| = 2$. Note that in this case $v(x_1) = v(x_2) = n = |G^+|$.

It is easy to see that if the action of G^+ on F produces only two orbits, then the stabilizers of points from these two orbits have the same number of elements and we are in the case $|F| = 2$ considered above. Thus for the rest of the proof, we can assume that there are three orbits.

Denote by x_1, x_2, x_3 points of these three orbits, so that $A = \{x_1, x_2, x_3\}$, and denote by v_1, v_2, v_3 the values of $v(x)$ (the number of elements in the stabilizers, or which is the same thing, the order of the corresponding rotation axis) at these points, numbered so that $v_1 \leq v_2 \leq v_3$.

We can assume that $|F| > 2$ (the case $|F| = 2$ was considered above), i.e., there are two rotation axes or more; but then the composition of the two rotations gives a rotation about a third axis and so $G^+ \geq 6$. We now claim that *there is an orbit with stabilizer equal to 2.*

Indeed, if, in contradiction with our claim, all the v_i were greater than 2, the right-hand side of formula (3.1) would be greater than or equal to 2, which we know is impossible.

Thus it remains to consider the situation in which $v_1 = 2$ and there are three orbits of the action of G^+ on F.

The rest of the proof is a case-by-case analysis of this situation depending on the possible values of the v_i. These values must satisfy relation (3.1), whose right-hand side is, as we remember, less than 2. Thus we must have the inequality

$$3 - \frac{1}{2} - \frac{1}{v_2} - \frac{1}{v_3} < 2. \tag{3.3}$$

When is this inequality possible? Since v_2 and v_3 are both integers greater than or equal to 2, this can happen only in the cases 2–5 indicated in the following table (in it, the column for the number of elements of G^+ was filled by using formula (3.2)):

	v_1	v_2	v_3	$\lvert G^+\rvert$
case 1	n	n	-	n
case 2	2	2	n	$2n$
case 3	2	3	3	12
case 4	2	3	4	24
case 5	2	3	5	60

In the rest of the proof, we consider each case separately and distinguish (among the points of F) the vertices of a (possibly degenerate) polyhedron on which G^+ acts. We then show that this action is one of those listed in the claim of the theorem, i.e., the distinguished polyhedron either degenerates into a regular polygon or is the tetrahedron, or the cube, or the dodecahedron.

Case 1 is the case in which $|F| = 2$ considered above, and we showed that it yields the group $\mathbb{Z}_n$, $n \geq 2$.

Case 2: assume that $v_2 = 2$. Then we have two rotation axes l_1, l_2 of order 2, i.e., such that the rotation angle is 180°. Consider the line l_3 perpendicular to these two axes. One of its intersection points with the sphere will be x_3. Let n be the order of the axis l_3. It follows from formula (3.2) that the number of elements of G^+ is equal to $2n$. We can now identify the three orbits in F: the n-point orbit containing x_1, which lies in the plane perpendicular to l_3 passing through the center of the sphere, the n-point orbit containing x_2, lying in the same plane, and the 2-point orbit consisting of x_3 and its antipodal point. It is now clear that in our case G^+ is isomorphic to the dihedral group $\mathbb{D}_n$.

Case 3: $v_2 = v_3 = 3$. Then the number of elements of our group can be computed from formula (3.2), and is equal to 12. Consider the axis of rotation l_3 passing through x_2; it is of order 3. Let x_3' and x_3'' be the two points to which the rotations about l_2 takes the point x_3. The rotation about the axis l_1 containing the point x_1 is of order 2, hence at least one of the three points x_3, x_3', x_3'' must be taken to a point (that we denote by x_3''') that does not coincide with one of those three. Thus we obtain a tetrahedron x_3, x_3', x_3'', x_3''', which, as we will soon see, turns out to be regular. Taking the composition of the rotations about l_3 and the rotation about l_1, we get another rotation of order 3, from which we conclude that another face of the tetrahedron is an equilateral triangle, and therefore the tetrahedron x_3, x_3', x_3'', x_3''' is regular. Taking the composition of two order three rotations, we obtain another order two rotation and, continuing in the same vein, we describe all 12 rotations of G^+ and can conclude that G^+ is isomorphic to $\mathrm{Rot}(\Delta^3)$.

Case 4: assume that $v_2 = 2$ and $v_3 = 4$. Here the strategy of proof is similar to the one in Case 3, except that now we find the 8 vertices of a cube (rather than those of a tetrahedron) among the points of F. To do this, we begin with the order 4 rotation, obtaining two squares inscribed in the sphere, then use the other rotations to show that the two squares are actually opposite faces of a cube, and finally verify that the 24 elements of G^+ are the symmetries of this cube, so that G^+ is isomorphic to $\mathrm{Sym}^+(I^3)$.

Case 5: assume that $v_2 = 3$ and $v_3 = 5$. Here the strategy of proof is similar to that used in Cases 3 and 4, except that now we construct a dodecahedron from points of F and obtain an isomorphism between G^+ and $\mathrm{Sym}^+(Dod)$. We relegate the details to Exercise 3.10.

Thus we see that the five cases correspond to the groups (i)–(v), respectively. The theorem is proved. □

3.2.7. Let us denote by $\widetilde{\mathbb{D}_n}$ the subgroup of SO(3) generated by the elements of $\mathbb{D}_n$ and the reflection ρ in the plane passing through the rotation axis of order n and one of the axes of order 2 in $\mathbb{D}_n$. Obviously the subgroup $\widetilde{\mathbb{D}_n}$ has $4n$ elements (because the compositions of ρ with different elements of $\mathbb{D}_n$ are all different from each other).

Note also that the subgroup of SO(3) generated by the elements of $\mathbb{Z}_n$ (interpreted as the motion group of the regular n-gon lying in the equatorial plane of the sphere and inscribed in it) and the reflection in a vertical plane passing through a vertex of the n-gon and the center of the sphere is $\mathbb{D}_n$.

3.2.8. Corollary. *Any finite subgroup G of O(3) is either isomorphic to one of the groups listed in Theorem* 3.2.6 *or to one of the following groups:*

(i) $\widetilde{\mathbb{D}}_n$, (ii) S_4, (iii) $\mathrm{Sym}(I^3)$, (iv) $\mathrm{Sym}(Dod)$.

Proof. Let G be a finite subgroup of SO(3) and let G^+ be its rotation subgroup. By Theorem 3.1, G^+ must be one of the five groups listed in the theorem. The whole group G is generated by the elements of G^+ and one reflection in a plane passing through the origin, so it must be one of the five groups listed in the statement of the corollary. □

3.3. The five regular polyhedra

A *regular polyhedron* is defined as a convex polyhedron (inscribed in the sphere $\mathbb{S}^2$) such that

(i) all its faces are congruent regular polygons of k sides for some $k > 2$;

(ii) the endpoints of all the edges issuing from each vertex lie in one plane and form a regular l-gon for some $l > 2$.

Theorem 3.3.1. *There are exactly five different regular polyhedra: the tetrahedron, the cube, the octahedron, the dodecahedron, and the icosahedron.*

Proof. This theorem follows from the Corollary to Theorem 3.1. Indeed, the definition implies that the isometry group of a regular polyhedron is finite and therefore must be one of the groups listed in Theorem 3.1. The two "series" (i) and (ii) do not give any (nondegenerate) polyhedra (why?). In case (iii), we get the tetrahedron (because its symmetry group is isomorphic to the permutation group S_4). In case (iv), we get the cube and its dual, the octahedron, in case (v), the dodecahedron and its dual, the icosahedron. □

Thus we obtain five geometries with three different group actions (tetrahedron, cube ∼ octahedron, dodecahedron ∼ icosahedron). To understand the group actions in these geometries, it is useful to have a look at their fundamental domains (Fig.3.5).

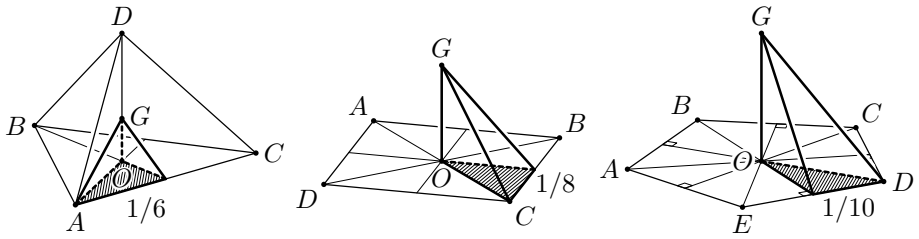


Figure 3.5. Fundamental domains of Platonic bodies

In all five cases, their fundamental domains are pyramids with vertex the center of the body and base the fundamental domain (a right triangle in all five cases) of the isometry group of a face. These triangles have acute angles 30° (tetrahedron, octahedron, icosahedron), 45° (cube), 54° (dodecahedron).

3.4. The five Kepler cubes

Kepler observed that the cube can be inscribed in five different ways into the dodecahedron. Here we will perform the opposite construction: starting

from the cube, we will construct a dodecahedron circumscribed to the cube. This will prove the existence of the dodecahedron.

Consider two copies $ABCDE$ and $A'B'C'D'E'$ of the regular pentagon with diagonals of length 1. Place these pentagons in the plane of the unit square $PQRS$ so that the diagonals BE and $B'E'$ are identified with PS and QR, respectively, and CD is parallel to $C'D'$. By rotating the pentagons in space about PS and QR, identify the sides CD and $C'D'$ above the square $PQRS$.

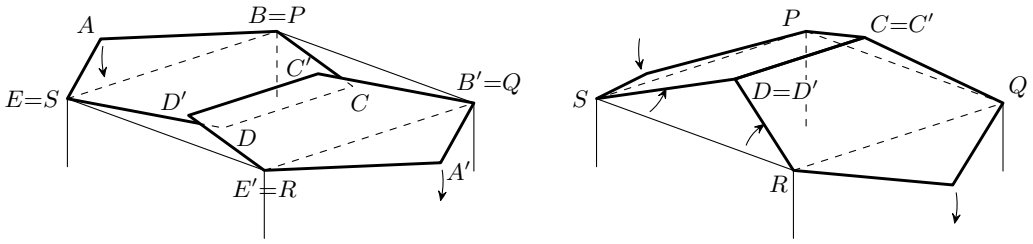


Figure 3.6. Constructing the dodecahedron

Now suppose $PQRS$ is the top face of the unit cube $PQRSP'Q'R'S'$. Place two more pentagons on the face $SRR'S'$ of the cube the same way as before, so that their parallel sides are parallel to SR. Now rotate these two pentagons until these parallel sides are identified. Then it is not hard to prove that the upper endpoint of the identified segment will coincide with one of the endpoints of the common (identified) segment of the first two pentagons. Perform similar constructions on the other faces of the cube. The polyhedron thus obtained will be the dodecahedron.

3.5. Regular polytopes in higher dimensions

In dimensions $n > 3$, there is a classification theorem for regular n-dimensional polytopes similar to that in three dimensions. Surprisingly, the number of types of polytopes decreases with the increase of n, changing from five (in dim=3) and six (dim=4) to three (when dim≥ 5). Thus, instead of the increased variety of regular bodies that might be expected in high dimensions, there are basically only three – the analogs of the tetrahedron, the cube, and the polytope dual to the cube.

In this section, after presenting the necessary definitions, we state the corresponding classification theorems without proof.

3.5.1. *Examples and definitions.* We begin with a simple example: the *four-dimensional cube*. In the Euclidean space $\mathbb{R}^4$, consider the 16 points $(\pm 1, \pm 1, \pm 1, \pm 1)$; their convex hull is by definition the 4-cube. A projection of the four-dimensional cube on the plane appears in Fig.3.7.

Even simpler (as its name indicates) is the regular *n-dimensional simplex* Δ^n, which is the n-dimensional analog of the tetrahedron, and is defined inductively: given the $(n-1)$-dimensional (regular) simplex Δ^{n-1} lying in $\mathbb{R}^{n-1}$, we construct a perpendicular from its center of gravity into the nth dimension (i.e., a line parallel to the basis vector $(0, \dots, 0, 1) \in \mathbb{R}^n \supset \mathbb{R}^{n-1}$) and take for the $n+1$st vertex of our simplex the point whose distance from the n vertices of Δ^{n-1} is equal to the length of the edges of Δ^{n-1}. It is easy to see that the transformation group of Δ^n is the permutation group Σ_{n+1}.

Regular n-dimensional polyhedra are defined recursively. The recursion begins for $n = 3$ and is that of a Platonic body (see Sect.3.3 above). If regular $(n-1)$-dimensional polyhedra have been defined, we define a *regular n-dimensional polyhedron* as a convex polyhedron (inscribed in the sphere $\mathbb{S}^{n-1} := \{(x_1, \dots, x_n) \in \mathbb{R}^n | x_1^2 + \dots x_n^2 = 1\}$) such that

(i) all its faces are congruent regular $(n-1)$-dimensional polyhedra;

(ii) the endpoints of all the edges issuing from each vertex lie in one hyperplane and form a regular $(n-1)$-dimensional polyhedron; all such polyhedra are congruent (but are not necessarily the same as those from item (i)).

To each regular polytope P, one can assign its *symbol*, defined (inductively) as the n-tuple of integers $(r_1, r_2, \dots, r_{n-1})$ in which r_1 is the number of edges of any one of the 2-dimensional faces Q of P, while $(r_2, \dots, r_{n-1})$ is the symbol of Q. For example, (4,3,3) is the symbol of the four-dimensional cube, (5,3) is that of the dodecahedron, (3,3,3,3) that of the five-dimensional regular simplex.

One can define the *dual* to any regular polytope in the natural way (similarly to the way it is done in dimension 3). For example, the 5-simplex is dual to itself, while the dual to the 4-cube is the so-called *cocube*, which has the symbol (3,3,4).

3.5.2. Theorem. *There are (up to homothety) six different regular polytopes in dimension 4; their symbols are*

$$(3,3,3), (4,3,3), (3,3,4), (3,4,3), (5,3,3), (3,3,5).$$

The reader who wishes to find a proof of this theorem is referred to Exercise

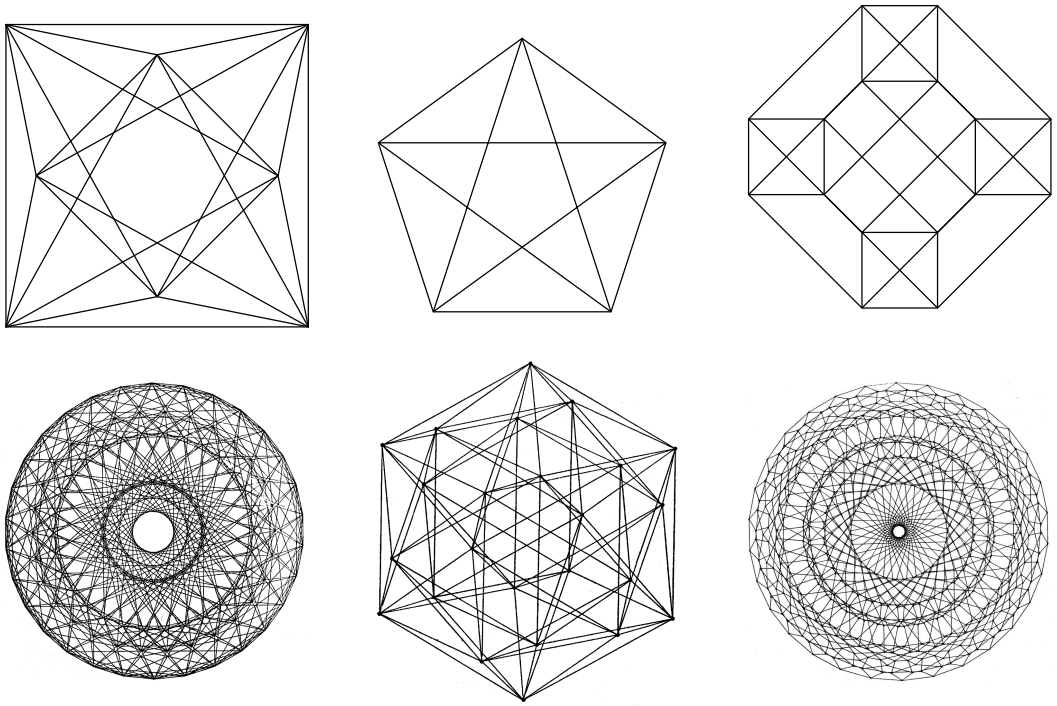


FIGURE 3.7. Regular 4-dimensional polyhedra

3.13, in which hints about the mysterious polytopes with symbols (3,4,3), (5,3,3), (3,3,5) appear.

3.5.3. Theorem. *In dimension $n \geq 5$ there are (up to homothety) three different regular polytopes: the n-simplex, the n-cube, and the n-cocube; their symbols are*

$$(3, 3, \ldots, 3, 3), (4, 3, \ldots, 3, 3), (3, 3, \ldots, 3, 4).$$

We omit the proof (see [2]); the reader is also referred to Exercise 3.14; an important formula used in the (inductive) proof appears as a hint in the Answers and Hints at the end of the book.

3.6. Problems

3.1. A regular pyramid of six lateral sides is inscribed in the sphere $\mathbb{S}^2$. Find its symmetry (i.e., isometry) group and its group of motions. How does your answer relate to the theorem on finite subgroups of $SO(3)$?

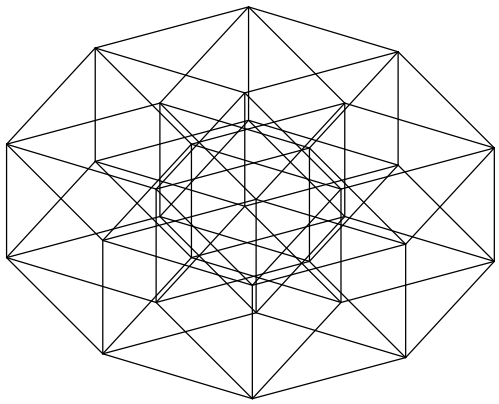


FIGURE 3.7. Projection of the edges of the 5-dimensional cube

3.2. Answer the same questions as in Problem 3.1 for

(a) the regular prism of six lateral sides;

(b) the regular truncated pyramid of five lateral sides;

(c) the double regular pyramid of six lateral sides (i.e., the union of two regular pyramids of six lateral sides with common base and vertices at the poles of the sphere);

3.3. Let G^+ be a finite subgroup of $SO(3)$ acting on the sphere $\mathbb{S}^2$ and F the set of all the points fixed by nontrivial elements of G^+; prove that F is invariant with respect to the action of G^+ and

$$|F| = |G^+| \cdot |A| - 2(|G^+| - 1),$$

where $A \subset F$ is a set containing exactly one point from each orbit of the action of G^+ on the set F.

3.4. Does the motion group of the cube have a subgroup isomorphic to the motion group of the regular tetrahedron?

3.5. Does the motion group of the dodecahedron have a subgroup isomorphic to the motion group of the cube?

3.6. In the motion group of the cube, find all groups isomorphic to $\mathbb{Z}_n$ and $\mathbb{D}_n$ for various values of n. Does it have any other subgroups?

3.7. Prove the existence of the dodecahedron in detail.

3.8. Given a cube inscribed in the sphere, let the set F consist of all the vertices of the cube, all the intersection points of the lines joining the centers

of its opposite faces, and of the lines joining the midpoints of opposite edges, and let G^+ be the motion group of the cube. Prove that G^+ acts on F, find all the orbits of this action and the stabilizers of all the points of F. Compare your findings with the proof of Theorem 3.1 in Case 4.

3.9. Given a regular tetrahedron inscribed in the sphere, let the set F consist of all its vertices and of the lines joining the midpoints of the edges, and let G^+ be the motion group of the tetrahedron. Prove that G^+ acts on F, find all the orbits of this action and the stabilizers of all the points of F. Compare your findings with the proof of Theorem 3.1 in Case 3.

3.10. Given a dodecahedron inscribed in the sphere, let the set F consist of all the vertices of the dodecahedron, all the intersection points of the lines joining the centers of its opposite faces and of the lines joining the midpoints of the edges, and let G^+ be the motion group of the dodecahedron. Prove that G^+ acts on F and complete the proof of Theorem 3.1 in Case 5.

3.11. Prove Theorem 3.1 in Case 4 by constructing an octahedron (instead of a cube) from the points of F. Compare with Fig.3.7.

3.12. Use your computer to produce a picture of the projection on an appropriately chosen two-dimensional plane of the five-dimensional cube.

3.13*. Prove the classification theorem for regular polyhedra in dimension four.

3.14*. Prove the classification theorem for regular polyhedra in dimension five.

Chapter 4

DISCRETE SUBGROUPS OF THE ISOMETRY GROUP OF THE PLANE AND TILINGS

This chapter, just as the previous one, deals with a classification of objects, the original interest in which was perhaps more aesthetic than scientific, and goes back many centuries ago. The objects in question are regular tilings (also called tessellations), i.e., configurations of identical figures that fill up the plane in a regular way. Each regular tiling is a geometry in the sense of Klein; it turns out that, up to isomorphism, there are 17 such geometries; their classification will be obtained by studying the corresponding transformation groups, which are discrete subgroups (see the definition in Section 4.3) of the isometry group of the Euclidean plane.

4.1. Tilings in architecture, art, and science

In architecture, regular tilings appear, in particular, as decorative mosaics (Fig.4.1) in the famous Alhambra palace (14th century Spain). According to M.Berger [2] and B.Grünbaum [7], part or all the 17 geometries mentioned above are realized by Alhambra mosaics. The reader can easily find beautiful color reproductions in the web by googling "Alhambra mosaics".

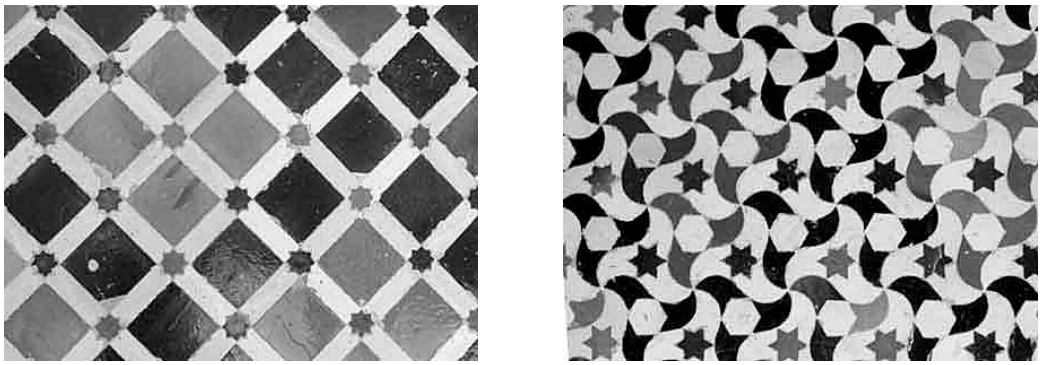


Figure 4.1. Two Alhambra mosaics

In art, the famous Dutch artist A.Escher, known for his "impossible" paintings, used regular tilings as the geometric basis of his wonderful "periodic" watercolors. Two of those are shown Fig.4.2.

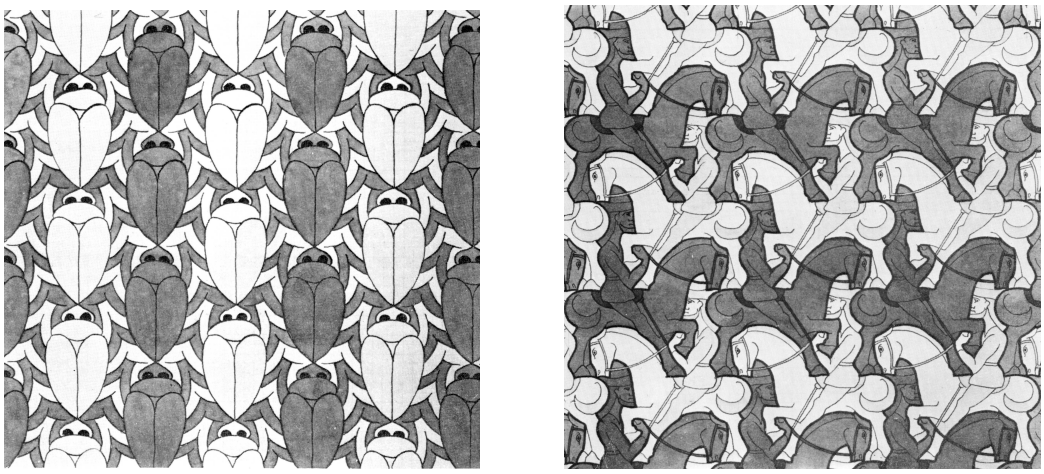


Fig.4.2. Two periodic watercolors by Escher

From the scientific viewpoint, not only regular tilings are important: it is possible to tile the plane by copies of one tile (or two) in an irregular

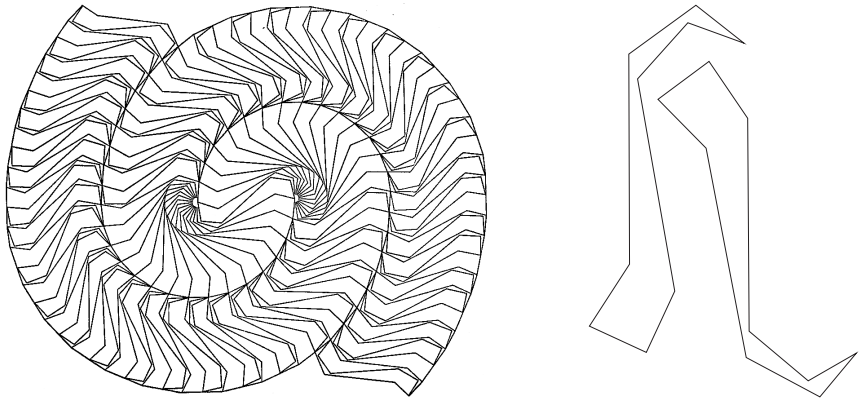


Fig.4.3. The Vorderberg tiling

(nonperiodic) way. It is easy to fill the plane with rectangular tiles of size say 10cm by 20cm in many nonperiodic ways. But the fact that $\mathbb{R}^2$ can be filled irregularly by nonconvex 9-gons is not obvious. Such an amazing construction, due to Vorderberg (1936), is shown in Fig.4.3. The figure indicates how to fill the plane by copies of two tiles (their enlarged copies are shown separately; they are actually mirror images of each other) by fitting them together to form two spiraling curved strips covering the whole plane.

Somewhat later, in the 1960ies, interest in irregular tilings was revived by the nonperiodic tilings due to the British mathematical physicist Roger Penrose, which are related to statistical models and the study of quasi-crystals. More recently, irregular tilings have attracted the attention of mathematicians, in particular that of the 2006 Fields medallist Andrey Okounkov in his work on three-dimensional Young diagrams.

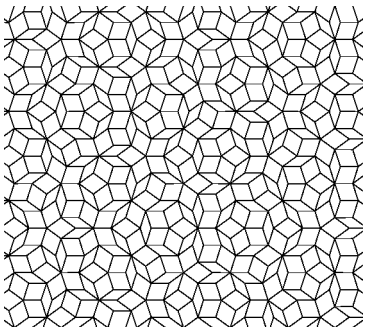


Figure 4.3. A Penrose tiling

4.2. Tilings and crystallography

The first proof of the classification theorem of regular tilings (defined below, see Sec.4.5.1) was obtained by the Russian crystallographer Fedorov in 1891. Mathematically, they are given by special discrete subgroups, called the *Fedorov groups*, of the isometry group $\mathrm{Sym}(\mathbb{R}^2)$ of the plane. As we mentioned above, there are 17 of them (up to isomorphism). The Fedorov groups act on the Euclidean plane, forming 17 different (i.e., nonisomorphic) geometries in the sense of Klein, which we call *tiling geometries*.

The proof given here, just as the one in the previous chapter, is group-theoretic, and is based on the study of discrete subgroups of the isometry group of the plane. In fact, the actual classification principle cannot be

stated without using transformation groups, and at first glance it is difficult to understand how it came about that the architects of the Alhambra palace, five centuries before the notion of group appeared in mathematics, actually found most or all the 17 regular tilings (in this connection, see the article [7] by B.Grünbaum). Actually, this is not surprising: a deep understanding of symmetry suffices to obtain answers to an intuitively clear question, even if one is unable to state the question in the terminology of modern mathematics.

Less visual, but more important for the applications (crystallography), is the generalization of the notion of regular tiling to three dimensions: configurations of identical polyhedra filling $\mathbb{R}^3$ in a regular way. Mathematically, they are also defined by means of discrete subgroups called *crystallographic groups* of the isometry group of $\mathbb{R}^3$ and have been classified: there are 230 of them. Their study is beyond the scope of this book.

We are concerned here with the two-dimensional situation, and accordingly we begin by recalling some facts from elementary plane geometry, namely facts concerning the structure of isometries of the plane $\mathbb{R}^2$.

4.3. Isometries of the plane

Recall that by $\mathrm{Sym}(\mathbb{R}^2)$ we denote the group of isometries (i.e., distance-preserving transformations) of the plane $\mathbb{R}^2$, and by $\mathrm{Sym}^+(\mathbb{R}^2)$ its group of motions (i.e., isometries preserving orientation). Examples of the latter are parallel translations and rotations, while reflections in a line are examples of isometries which are not motions (they reverse orientation).

(We consider an isometry orientation-reversing if it transforms a clockwise oriented circle into a counterclockwise oriented one. This is not a mathematical definition, since it appeals to the physical notion of "clockwise rotation", but there is a simple and rigorous mathematical definition of orientation-reversing (-preserving) isometry in linear algebra courses, based on the sign ($\pm$) of the determinant of the corresponding linear map.)

Below we list some well known facts about isometries of the plane; their proofs are relegated to exercises appearing at the end of the present chapter.

4.3.1. A classical theorem of elementary plane geometry says that any motion is either a *parallel translation* or a *rotation* (see Exercise 4.1).

4.3.2. A less popular but equally important fact is that any orientation-reversing isometry is a *glide reflection*, i.e., the composition of a reflection in some line and a parallel translation by a vector collinear to that line (Exercise 4.2).

4.3.3. The composition of two rotations is a rotation (except for the particular case in which the two angles of rotation are equal but opposite: then their composition is a parallel translation). In the general case, there is a simple construction that yields the center and angle of rotation of the composition of two rotations (see Exercise 4.3). This important fact plays the key role in the proof of the theorem on the classification of regular tilings.

4.3.4. The composition of a rotation and a parallel translation is a rotation by the same angle about a point obtained by shifting the center of the given rotation by the given translation vector (Exercise 4.4).

4.3.5. The composition of two reflections in lines l_1 and l_2 is a rotation about the intersection point of the lines l_1 and l_2 by an angle equal to twice the angle from l_1 to l_2 (Exercise 4.5).

4.4. Discrete groups and discrete geometries

The action of a group G on a space X is called *discrete* if none of its orbits possess accumulation points, i.e., there are no points of $x \in X$ such that any neighborhood of x contains infinitely many points belonging to one orbit. Here the word "space" can be understood as Euclidean space $\mathbb{R}^n$ (or as a subset of $\mathbb{R}^n$), but the definition remains valid for arbitrary metric and topological spaces.

A simple example of a discrete group acting on $\mathbb{R}^2$ is the group consisting of all translations of the form $k\,\vec{v}$, where v is a fixed nonzero vector and $k \in \mathbb{Z}$. The set of all rotations about the origin of $\mathbb{R}^2$ by angles which are integer multiples of $\sqrt{2}\,\pi$ is a group, but its action on $\mathbb{R}^2$ is not discrete (since $\sqrt{2}$ is irrational, orbits are dense subsets of circles centered at the origin).

4.5. The seventeen regular tilings

4.5.1 *Formal definition.* By definition, a *tiling* or *tessellation* of the plane $\mathbb{R}^2$ by a polygon T_0, the *tile*, is an infinite family $\{T_1, T_2, \dots\}$ of pairwise nonoverlapping (i.e., no two distinct tiles have common interior points) copies of T_0 filling the plane, i.e., $\mathbb{R}^2 = \bigcup_{i=1}^{\infty} T_i$.

For example, it is easy to tile the plane by any rectangle in different ways, e.g. as a rectangular lattice as well as in many irregular, nonperiodic ways. Another familiar tiling of the plane is the *honeycomb lattice*, where the plane is filled with identical copies of a regular hexagon.

A polygon $T_0 \subset \mathbb{R}^2$, called the *fundamental tile*, determines a *regular tiling* of the plane $\mathbb{R}^2$ if there is a subgroup G (called the *tiling group*) of the isometry group $\mathrm{Sym}(\mathbb{R}^2)$ of the plane such that

(i) G acts discretely on $\mathbb{R}^2$, i.e., all the orbits of G have no accumulation points;

(ii) the images of T_0 under the action of G fill the plane, i.e.,

$$\bigcup_{g\in G} g(T_0) = \mathbb{R}^2;$$

(iii) for $g, h \in G$ the images $g(T_0)$, $h(T_0)$ of the fundamental tile coincide if and only if $g = h$.

Actually, (ii) and (iii) imply (i), but we will not prove this (see the first volume of Berger's book, pp.37-38 of the French edition).

The action of a tiling group $G \subset \mathrm{Sym}(\mathbb{R}^2)$ on the plane $\mathbb{R}^2$ is, of course, a geometry in the sense of Klein that we call the *tiling geometry* (or *Fedorov geometry*) of the group G.

4.4.2. *Examples of regular tilings*

Six examples of regular tilings are shown in Fig.4.4.

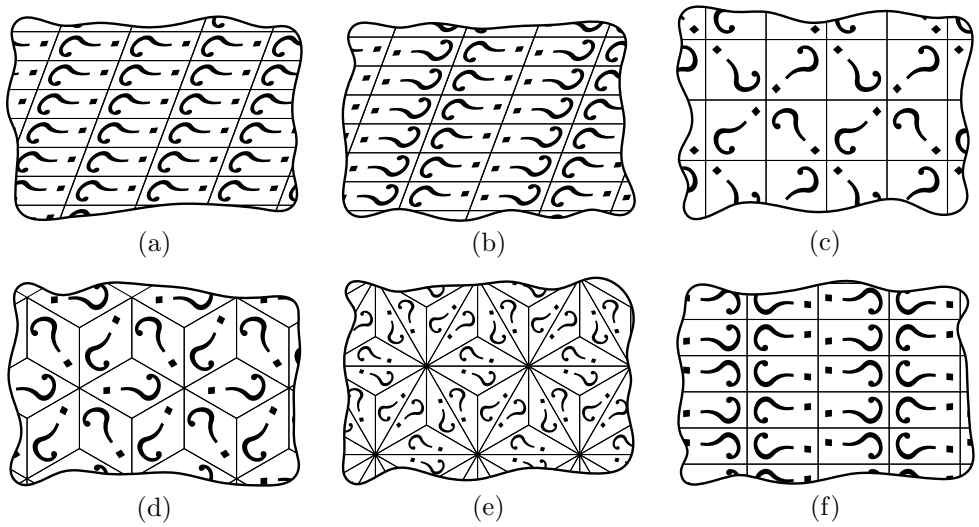


Figure 4.4. Six regular tilings of the plane

Given two tiles, there is one element of the transformation group that takes one to the other. The question marks show *how* the tiles are mapped to each other. (Without the question marks, the action of the transformation group would not be specified; see Exercise 4.16).

The first five tilings (a-e) are *positive*, i.e., they correspond to subgroups of the group $\mathrm{Sym}^+(\mathbb{R}^2)$ of motions (generated by all rotations and translations) of the plane (one-sided tiles slide along the plane). The sixth tiling (f) allows "turning over" the (two-sided) tiles.

Let us look at the corresponding tiling groups in more detail.

4.4.3. Theorem. [Fedorov, 1891]. *Up to isomorphism, there are exactly five different one-sided tiling geometries of the plane $\mathbb{R}^2$. They are shown in* Fig.4.4,a–e.

Proof. Let G be a group of positive tilings. Consider its subgroup G_T of all the parallel translations in G.

4.4.4. Lemma. *The subgroup G_T is generated by two noncollinear vectors v and u.*

Proof. Arguing by contradiction, suppose that G_T is trivial (there are no parallel translations except the identity). Let r, s be any two (nonidentical) rotations with different centers. Then $rsr^{-1}s^{-1}$ is a nonidentical translation (to prove this, draw a picture). A contradiction. □

Now suppose that all the elements of G_T are translations generated by (i.e., proportional to) one vector v. Then it is not difficult to obtain a contradiction with item (ii) of the definition of regular tilings. □

Now if G contains no rotations, i.e., $G = G_T$, then we get the tiling (a). Further, If G contains only rotations of order 2, then it is easy to see that we get the tiling (b).

4.4.5. Lemma. *If G contains a rotation of order $\alpha \geq 3$, then it contains two more rotations (of some some orders β and γ) such that*

$$\boxed{\frac{1}{\alpha}+\frac{1}{\beta}+\frac{1}{\gamma}=1.}$$

Sketch of the proof. Let A be the center of a rotation of order α. Let B and C be the nearest (from A) centers of rotation not obtainable from A by translations. Then the boxed formula follows from the fact that the sum of angles of triangle ABC is π. The detailed proof of this lemma is the topic of one of the exercises. □

Since the three rotations are of order greater or equal to 3, it follows from the boxed formula that only three cases are possible.

	$1/\alpha$	$1/\beta$	$1/\gamma$
case 1	1/3	1/3	1/3
case 2	1/2	1/4	1/4
case 3	1/2	1/3	1/6

Studying these cases one by one, it is easy to establish that they correspond to the tilings (d),(c),(e) of Fig.4.3, respectively.

This concludes the proof of Theorem 4.4.3. □

In the general case (all tilings, including those by two-sided tiles), there are exactly seventeen nonequivalent tilings. This was also proved by Fedorov. The 12 two-sided ones are shown on the next page.

We will not prove the second part of the classification theorem for regular plane tilings (it consists in finding the remaining 12 regular tilings, for which two-sided tiles are required). However, the reader can study some examples of these 12 tilings by doing some of the exercises. Note that there is a nice web site with beautiful examples of decorative patterns corresponding to the 17 regular tilings:

http://www2.spsu.edu/math/tile/symm/ident17.htm

One can also visit the Escher website.

4.5. The 230 crystallographic groups

The crystallographic groups are the analogs in $\mathbb{R}^3$ of the tiling groups in Euclidean space $\mathbb{R}^2$. The corresponding periodically repeated polyhedra are not only more beautiful than tilings, they are more important: the shapes of most of these polyhedra correspond to the shapes of real-life crystals. There are 230 crystallographic groups. The proof is very tedious: there are 230 cases to consider, in fact more, because many logically arising cases turn out to be geometrically impossible, and it lies, as we mentioned above, outside the scope of this book.

Those of you who would like to see some nontrivial examples of geometries corresponding to some of the crystallographic groups should look at Problem 4.5 and postpone their curiosity to the next chapter, where 4 examples of actual crystals will appear in the guise of Coxeter geometries. Another possibility is to consult the website http://webmineral.com/crystal.shtml or to google the words “crystallographic group”.

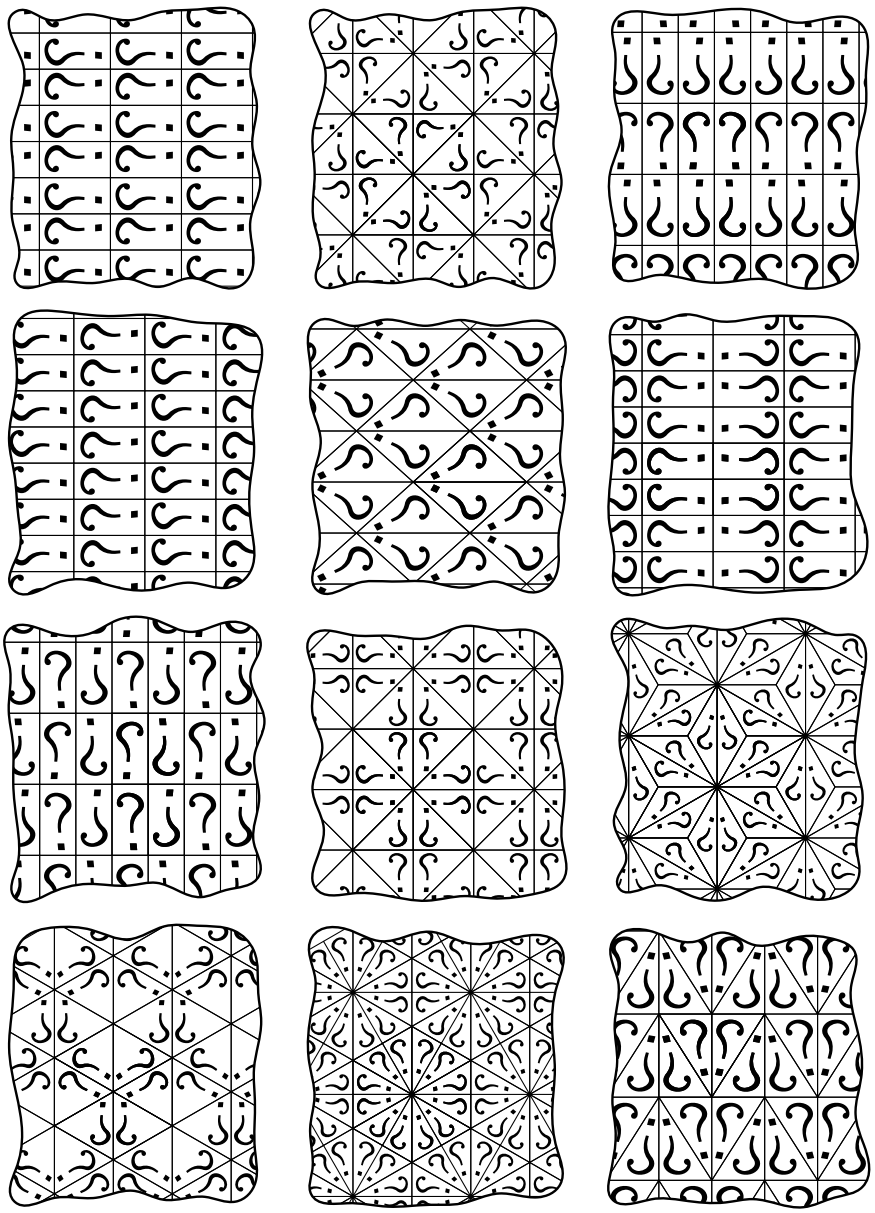


Figure 4.5. Two-sided regular tilings

4.6. Problems

4.1. Prove that any motion of the plane is either a translation by some vector v, $|v| \geq 0$, or a rotation r_A about some point A by a nonzero angle.

4.2. Prove that any orientation-reserving isometry of the plane is a glide reflection in some line L with glide vector u, $|u| \geq 0$, $u||L$.

4.3. Justify the following construction of the composition of two rotations $r = (a, \varphi)$ and (b, ψ). Join the points a and b, rotate the ray $[a, b\rangle$ around a by the angle $\varphi/2$, rotate the ray $[b, a\rangle$ around b by the angle $-\psi/2$, and denote by c the intersection point of the two obtained rays; then c is the center of rotation of the composition rs and its angle of rotation is $2(\pi - \varphi/2 - \psi/2)$. Show that this construction fails in the particular case in which the two angles of rotation are equal but opposite, and then their composition is a parallel translation).

4.4. Prove that the composition of a rotation and a parallel translation is a rotation by the same angle and find its center of rotation.

4.5. Prove that the composition of two reflections in lines l_1 and l_2 is a rotation about the intersection point of the lines l_1 and l_2 by an angle equal to twice the angle from l_1 to l_2.

4.6. Indicate a finite system of generators for the transformation groups corresponding to each of the tilings shown in Figure 4.4 a), b),...,f).

4.7. Is it true that the transformation group of the tiling shown on Figure 4.4 (b) is a subgroup of the one of Figure 4.4 (c)?

4.8. Indicate the points that are the centers of the rotation subgroups of the transformation group of the tiling shown in Figure 4.4(c).

4.9. Write out a presentation of the isometry group of the plane preserving

(a) the regular triangular lattice;

(b) the square lattice;

(c) the hexagonal (i.e., honeycomb) lattice.

4.10. For which of the five Platonic bodies can a (countable) collection of copies of the body fill Euclidean 3-space (without overlaps)?

4.11. For the two Escher pictures in Fig.4.2 indicate to which of the 17 Fedorov groups they correspond.

4.12. Exactly one of the 17 Fedorov groups contains a glide reflection but no reflections. Which one?

4.13. Which two of the 17 Fedorov groups contain rotations by $\pi/6$?

4.14. Which three of the 17 Fedorov groups contain rotations by $\pi/2$?

4.15. Which five of the 17 Fedorov groups contain rotations by π only?

4.16. Rearrange the question marks in the tiling (c) so as to make the corresponding geometry isomorphic that of the tiling (a).

Chapter 5
REFLECTION GROUPS AND COXETER GEOMETRIES

In this chapter, as in the previous one, we study geometries defined by certain discrete subgroups of the isometry group of the plane (and, more generally, of n-dimensional space), namely the subgroups generated by reflections (called Coxeter groups after the 20th century Canadian mathematician who invented them). These geometries are perhaps not as beautiful as those studied in the previous two chapters, but are more important in the applications (in algebra and topology). On the other hand, they do have an aesthetic origin: what one sees in a kaleidoscope (a child's toy very popular before the computer era) is an instance of such a geometry. Following E.B.Vinberg, we call these geometries (in the two-dimensional case) kaleidoscopes. We prove the classification theorem for them in dimension 2 and state its generalization to higher dimensions without proof (using the notion of Coxeter scheme).

5.1. An example: the kaleidoscope

The kaleidoscope is a children's toy: bright little pieces of glass are placed inside a regular triangular prism and are multiply reflected by three mirrors forming the lateral faces of the prism.

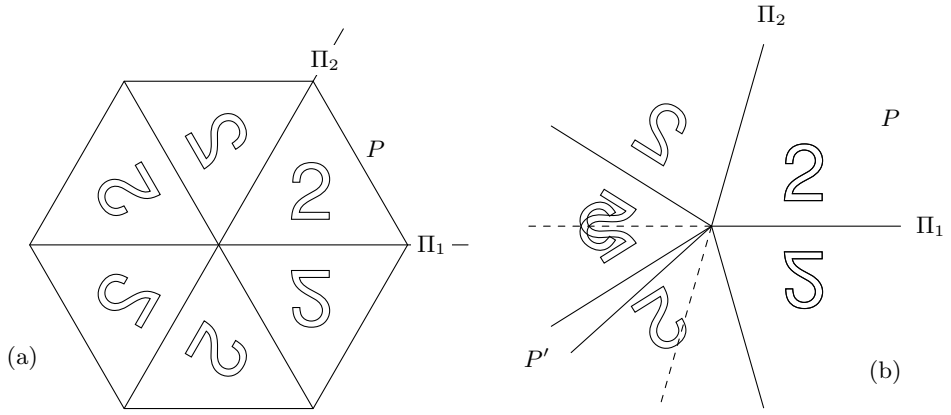


Fig.5.1. Geometry of the kaleidoscope

Looking into the prism, you see a colorful repeated pattern: the picture in the triangle and its mirror images alternate, forming a hexagon (the union

of six equilateral triangles), see Fig.5.1(a), surrounded by more equilateral triangles *ad infinitum*.

Mathematically, this is a two-dimensional phenomenon: the equilateral triangle forming the base of the prism is the fundamental domain of a discrete group acting on the plane of the base.

Now if the kaleidoscope is deformed (e.g., the angles between the faces of the prism are slightly changed), then the picture becomes fuzzy, no pattern can be seen. In such a situation, the images of the base triangle overlap infinitely many times (see Fig.5.1(b), the transformation group acting on the triangle is not discrete; we will not study this "bad" case: we only study the case of the "nice" kaleidoscope in dimension two and then generalize it to any dimension.

5.2. Coxeter polygons and polyhedra

Consider a dihedral angle $\alpha < \pi/2$ formed by two plane two-sided mirrors Π_1, Π_2. What will the observer O see? Any picture P inside the angle will be reflected by Π_1; its image P' will be in turn be reflected by the image of Π_1 by Π_2, and so on. At the same time, the picture P inside the angle will be reflected by Π_2; its image P'' will be in turn be reflected by the image of Π_2 by Π_1, etc. Two cases are possible: either the reflections coming from different sides will overlap (Fig.5.1,b) or the reflected pictures will coincide (Fig.5.1,a). Obviously, the pictures will coincide if (and only if) the angle α is of the form π/k, where $k = 2, 3, \ldots$.

Mathematically, this situation is the following. On the Euclidean plane, we take two straight lines forming the angle α and consider the group G of all transformations of the plane generated by the reflections in these two lines. Let F be the plane region bounded by the two rays forming the angle α. Obviously, no two regions $g(F)$ and $h(F)$, $g, h \in G$, $g \neq h$, overlap iff $\alpha = \pi/k$, where $k = 2, 3, \ldots$. In that case, G is the dihedral group $\mathbb{D}_k$.

Now suppose we are given a convex polygon F in the plane with vertex angles less than or equal to $\pi/2$. Consider the group G_F of transformations of the plane generated by reflections in the lines containing the sides of F. We say that G_F is *acts transitively on* F if the images $g(F)$, $g \in G_F$, never overlap. A necessary condition for the transitive action of G_F on F is that all the vertex angles of F be of the form π/k for various values of k; this follows from the argument in the previous paragraph. Obviously, this condition is not sufficient.

The previous arguments are the motivation for the following definition.

A convex polygon F is called a *Coxeter polygon* if all its vertex angles are of the form π/k for various values of $k = 2, 3, \ldots$ and it generates a transitive action of the group G_F. Coxeter polygons will be classified below – there are only four.

The above can be generalized to three-dimensional space. The corresponding definition is the following: a convex polyhedron is called a *Coxeter polyhedron* P if all its dihedral angles are of the form π/k for various values of $k = 2, 3, \ldots$ and it generates a transitive action of G_P, where G_P is the transformation group generated by the reflections in the planes containing the faces of P. Coxeter polyhedra will be classified below (there are seven).

5.3. Coxeter geometries on the plane

Let F be a Coxeter polygon in the plane $\mathbb{R}^2$. The *Coxeter geometry* with fundamental domain F is the geometry $(\mathbb{R}^2 : G_F)$, where G_F is the group of transformations of the plane generated by the reflections in the lines containing the sides of the polygon F. The goal of this section is to classify all Coxeter geometries on the plane.

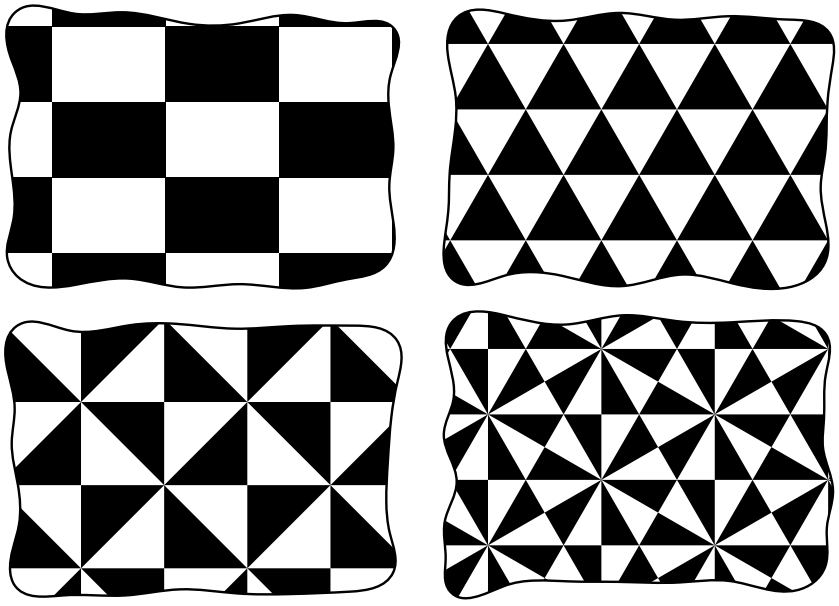


Figure 5.2. The four plane Coxeter geometries

5.3.1. Theorem. *Up to isomorphism, there are four Coxeter geometries in the plane; their fundamental polygons are the rectangle, the equilateral triangle, the isosceles right triangle, and the right triangle with angles* $\pi/3$ *and* $\pi/6$ (*see Fig.*5.2).

Proof. Let F be the fundamental polygon of a Coxeter geometry. If it has n sides, then the sum of its angles is $\pi(n-2)$ and so the average value of its angles is $\pi(1-2/n)$. Now n cannot be greater than 4, because F would then have an obtuse angle (and this contradicts the definition of Coxeter polygon). If $n=4$, then all angles of F are $\pi(1-2/4)=\pi/2$ and F is a rectangle. Finally, if $n=3$, and the angles of the fundamental triangle are $\pi/k, \pi/l, \pi/m$, then (since their sum is π), we obtain a Diophantine equation for k, l, m:

$$\boxed{\frac{1}{k}+\frac{1}{l}+\frac{1}{m}=1}.$$

This equation has three solutions: $(3,3,3)$, $(2,4,4)$, $(2,3,6)$. These solutions correspond to the three triangles listed in the theorem. □

5.4. Coxeter geometries in Euclidean space $\mathbb{R}^3$

5.4.1. In this section we study the Coxeter geometries in $\mathbb{R}^3$. A Coxeter polyhedron $F \subset \mathbb{R}^3$ is a convex polyhedron (i.e., the bounded intersection of a finite number of half-spaces in $\mathbb{R}^3$) with dihedral angles of the form π/k for various values of $k = 2, 3, \ldots$.

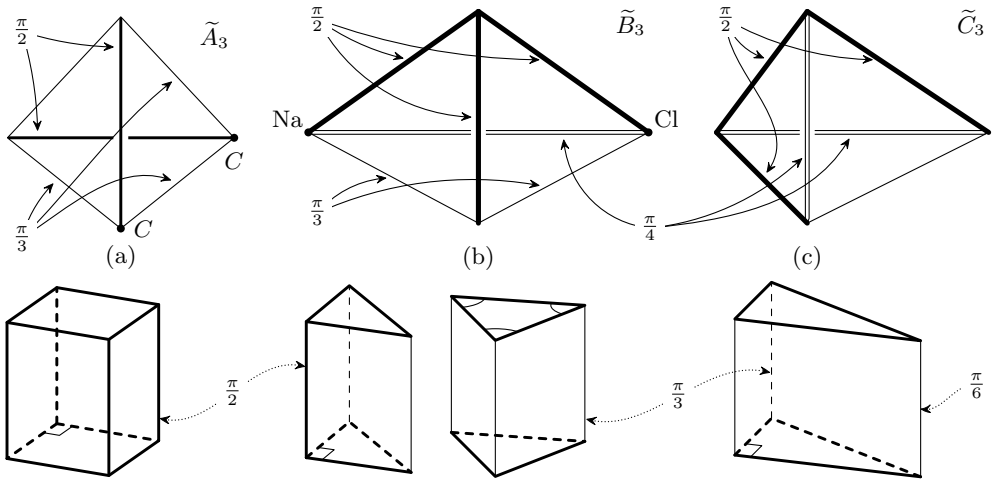


Figure 5.3. The seven Coxeter polygons in 3-space

A *Coxeter geometry* in $\mathbb{R}^d$ with fundamental polyhedron F is defined just as in the case $d = 2$ (see Section 5.3).

5.4.2. Theorem. *There are seven Coxeter geometries in three-dimensional space; their fundamental polyhedra are the four right prisms over the the rectangle, the equilateral triangle, the isosceles right triangle, and the right triangle with acute angles $\pi/3$ and $\pi/6$, and the three (nonregular) tetrahedra shown in Fig.*5.3.

It is not very difficult to prove that the seven polyhedra (listed in the theorem) indeed define Coxeter geometries. To prove that there are no other geometries, nontrivial information from linear algebra (in particular, the notion of Gramm matrix) is needed. Therefore, we omit the proof (see the book [4] or, for readers of Russian, a series of articles in *Matematicheskoye Prosveshchenie*, Ser.3, no. 7, 2003).

A remark about terminology. The term "Coxeter geometry" is not a standard term. E.B.Vinberg uses the term "kaleidoscope" instead. Also, we do *not* use the term "Coxeter group" for the transformation group of a Coxeter geometry. This is because "Coxeter group" is standardly used in a somewhat different sense than "transformation group of a Coxeter geometry".

Coxeter geometries are not only abstract mathematical objects, they are also important models in crystallography. For example, the polyhedron in Fig.5.3,b is the crystal of ordinary salt, while the polyhedron in Fig.5.3,a is a diamond crystal.

5.5. Coxeter schemes and the classification theorem

5.5.1. In this section we study the general case of a Coxeter geometry in $\mathbb{R}^d$ for an arbitrary positive integer d. A Coxeter polyhedron $F \subset \mathbb{R}^d$ is a convex polyhedron (i.e., the bounded intersection of a finite number of half-spaces in $\mathbb{R}^d$) with dihedral angles of the form π/k for various values of $k = 2, 3, \ldots$ such that the reflections in the $(d-1)$-dimensionsal hyperplanes containing its faces generate a transitively acting group G_F. (The definition of the measure of a dihedral angle in Euclidean space of arbitrary dimension d appears in the linear algebra course.) A *Coxeter geometry* in $\mathbb{R}^d$ with fundamental polyhedron F is defined exactly like in the cases $d = 2$ and $d = 3$ (see Sections 5.2 and 5.3).

5.5.2. A *Coxeter scheme* is a graph (with integer weights on the edges) encoding a Coxeter polyhedron (in particular, polygons) in any dimension d. The scheme of a given Coxeter polyhedron is constructed as follows:

its vertices correspond to the faces of the polyhedron, two vertices whose corresponding faces form an angle of π/m, $m \geq 3$, are joined by an edge with the weight $m-2$; if two faces are parallel, the corresponding vertices are joined by an edge with weight ∞. (Note that vertices corresponding to perpendicular edges are *not* joined by an edge.)

Graphically, instead of writing the weights 2,3,4 on the edges of a scheme, we draw double, triple, quadruple edges; instead of writing ∞ on an edge, we draw a very thick edge.

For example, the Coxeter scheme of the rectangle consists of two components, each of which has two vertices joined by an edge with weight ∞, while the scheme of an equilateral triangle has three vertices joined cyclically by three edges with weights 1.

5.5.3. Theorem. *The Coxeter geometries in all dimensions are classified by the connected components of their Coxeter schemes listed in Fig.*5.4.

We omit the proof (see the book [4] or, for readers of Russian, the articles in the issue of *Matematicheskoye Prosveshchenie* cited above).

5.6. Problems

5.1. Three planes P_1, P_2, P_3 passing through the z-axis of Euclidean space $\mathbb{R}^3$ are given. The angles between P_1 and P_2, P_2 and P_3 are α and β, respectively.

(a) Under what conditions on α and β will the group generated by reflections with respect to the three planes be finite?

(b) If these conditions are satisfied, how can one find the fundamental domain of this action?

5.2. Three straight lines L_1, L_2, L_3 in the Euclidean plane form a triangle with interior angles α, β, and γ.

(a) Under what conditions on α, β, γ will the group generated by reflections with respect to the three lines be discrete?

(b) If these conditions are satisfied, how can one find the fundamental domain of this action?

5.3. Consider the six lines $L_1, \ldots, L_6$ containing the six sides of a regular plane hexagon and denote by G the group generated by reflections with respect to these lines. Does this group determine a Coxeter geometry?

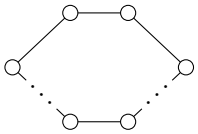
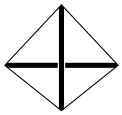
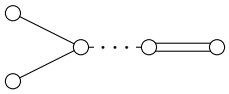
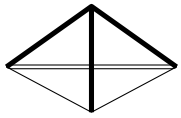
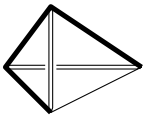
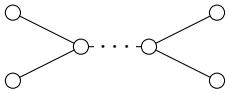
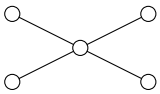
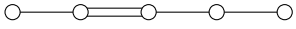
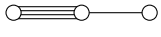
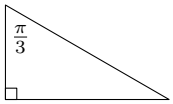
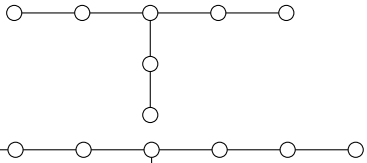
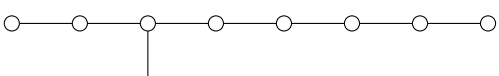
Name	Coxeter scheme	dim	#(faces)	view in $\mathbb{R}^3$
$\widetilde{A}_1$		1	2	
$\widetilde{A}_n$		$n-1$	n	
$\widetilde{B}_n$		$n-1$	n	
$\widetilde{C}_n$		$n-1$	n	
$\widetilde{D}_n$		$n-1$	$n \geqslant 5$	none!
$\widetilde{D}_4$		4	5	none!
$\widetilde{F}_4$		4	5	none!
$\widetilde{G}_2$		2	3	
$\widetilde{E}_6$				none!
$\widetilde{E}_7$				none!
$\widetilde{E}_8$				none!

Figure 5.4. Coxeter schemes for the Coxeter geometries

5.4. Let F be a Coxeter triangle, s_1, s_2, s_3 be the reflections with respect to its sides, and G_F the corresponding transformation group.

(a) Give a geometric description and a description by means of words in the alphabet s_1, s_2, s_3 of all the elements of G_F that leave a chosen vertex of F fixed.

(b) Give a geometric description and a description by means of words in the alphabet s_1, s_2, s_3 of all the elements of G_F which are parallel translations.

Consider the three cases of different Coxeter triangles separately.

5.5. Draw the Coxeter schemes of

(a) all the Coxeter triangles;

(b) all the three-dimensional Coxeter polyhedra.

5.6. Prove that all the edges at each vertex of any three-dimensional Coxeter polyhedron lie on three straight lines passing through that vertex.

5.7. Let $(F : G_F)$ be a Coxeter geometry of arbitrary dimension. Prove that

(a) if $s \in G_F$ is the reflection in a hyperplane P, then, for any $g \in G_F$, gsg^{-1} is the reflection in the hyperplane gP;

(b) any reflection from the group G_F is conjugate to the reflection in one of the faces of the polyhedron F,

5.8. Describe some four-dimensional Coxeter polyhedron other than the four-dimensional cube.

5.9. (a) Does the transformation group generated by the reflections in the faces of regular tetrahedron define a Coxeter geometry?

(b) Same question for the cube.

(c) Same question for the octahedron.

(d) Same question for the dodecahedron.

Chapter 6
SPHERICAL GEOMETRY

So far we have studied finite and discrete geometries, i.e., geometries in which the main transformation group is either finite or discrete. In this chapter, we begin our study of infinite continuous geometries with spherical geometry, the geometry ($\mathbb{S}^2$:O(3)) of the isometry group of the two-dimensional sphere, which is in fact the subgroup of all isometries of $\mathbb{R}^3$ that map the origin to itself; O(3) is called the *orthogonal group* in linear algebra courses.

But first we list the classical continuous geometries that will be studied in this course. Some of them may be familiar to the reader, others will be new.

6.1. A list of classical continuous geometries

Here we merely list, for future reference, several very classical geometries whose transformation groups are "continuous" rather than finite or discrete. We will not make the intuitively clear notion of continuous transformation group precise (this would involve defining the so-called *topological groups* or even *Lie groups*), because we will not study this notion in the general case: it is not needed in this introductory course. The material of this section is not used in the rest of the present chapter, so the reader who wants to learn about spherical geometry without delay can immediately go on to Sect. 6.3.

6.1.1. *Finite-dimensional vector spaces* over the field of real numbers are actually geometries in the sense of Klein (the main definition of Chapter 1). From that point of view, they can be written as

$$\boxed{(\mathbb{V}^n : \mathrm{GL}(n))},$$

where $\mathbb{V}^n$ denotes the n-dimensional vector space over $\mathbb{R}$ and $\mathrm{GL}(n)$ is the *general linear group*, i.e., the group of all nondegenerate linear transformations of $\mathbb{V}^n$ to itself.

The subgeometries of $(\mathbb{V}^n : \mathrm{GL}(n))$ obtained by replacing the group $\mathrm{GL}(n)$ by its subgroup $\mathrm{O}(n)$ (consisting of orthogonal transformations) is called the *n-dimensional orthonormed vector space* and denoted

$$\boxed{(\mathbb{V}^n : \mathrm{O}(n))}.$$

These "geometries" are rather algebraic and are usually studied in linear algebra courses. We assume that the reader has some background in linear algebra and remembers the first basic definitions and facts of the theory.

6.1.2. *Affine spaces* are, informally speaking, finite-dimensional vector spaces "without a fixed origin". This means that their transformation groups $\mathrm{Aff}(n)$ contain, besides $\mathrm{GL}(n)$, all parallel translations of the space (i.e., transformations of the space obtained by adding a fixed vector to all its elements). We denote the corresponding geometry by

$$\boxed{(\mathbb{V}^n : \mathrm{Aff}(n))} \quad \text{or} \quad \boxed{(\mathbb{R}^n : \mathrm{Aff}(n))},$$

the later notation indicating that the elements of the space are now regarded as *points*, i.e., the endpoints of the vectors (issuing from the origin) rather than the vectors themselves. This is a more geometric notion than that of vector space, but is also usually studied in linear algebra courses.

6.1.3. *Euclidean spaces* are geometries that we denote

$$\boxed{(\mathbb{R}^n : \mathrm{Sym}(\mathbb{R}^n))};$$

here $\mathrm{Sym}(\mathbb{R}^n)$ is the isometry group of Euclidean space $\mathbb{R}^n$, i.e., the group of distance-preserving transformations of $\mathbb{R}^n$. This group has, as a subgroup, the *orthogonal group* $\mathrm{O}(n)$ that consists of isometries leaving the origin fixed (the group $\mathrm{O}(n)$ should be familiar from the linear algebra course), but also contains the subgroup of parallel translations.

We assume that, for $n = 2, 3$, the reader knows Euclidean geometry from school (of course it was introduced differently, usually via some modification of Euclid's axioms) and is familiar with the structure of the isometry groups of Euclidean space for $n = 2, 3$.

The reader who feels uncomfortable with elementary Euclidean plane and space geometry can consult Appendix I. A rigorous axiomatic approach to Euclidean geometry in dimensions $d = 2, 3$ (based on Hilbert's axioms) appears in Appendix III.

Note that the transformation groups of these three geometries (vector spaces, affine and Euclidean spaces) act on the same space ($\mathbb{R}^n$ and $\mathbb{V}^n$ can be naturally identified), but the geometries that they determine are different, because the four groups $\mathrm{GL}(n)$, $\mathrm{O}(n)$, $\mathrm{Aff}(n)$, $\mathrm{Sym}(\mathbb{R}^3)$ are different. The corresponding geometries will not be studied in this course: traditionally,

this is done in linear algebra courses, and we have listed them here only to draw a complete picture of classical geometries.

Our list continues with three more classical geometries that we will study, at least in small dimensions (mostly in dimension 2).

6.1.4. *Hyperbolic spaces* $\mathbb{H}^n$ (called *Lobachevsky spaces* in Russia) are "spaces of constant negative curvature" (you will learn what this means much later, in differential geometry courses) with transformation group the isometry group of the hyperbolic space (i.e., the group of transformations preserving the "hyperbolic distance"). We will only study the hyperbolic space of dimension $n = 2$, i.e., the hyperbolic plane. Three models of $\mathbb{H}^2$ will be studied, in particular, the *Poincaré disk model*,

$$\boxed{(\mathbb{H}^2 : \mathcal{M})};$$

here $\mathbb{H}^2 := \{(x, y) \in \mathbb{R}^2 | x^2 + y^2 < 1\}$ is the open unit disk and $\mathcal{M}$ is the group of *Möbius transformations* (the definition appears in Chapter 7) that take the disk to itself.

We will also study two other models of hyperbolic plane geometry (the *half-plane model*, also due to Poincaré, and the *Cayley–Klein model*). A special chapter describes how attempts to prove Euclid's Fifth Postulate led to the appearance of hyperbolic plane geometry and the dramatic history of its creation by Gauss, Lobachevsky and Bolyai.

6.1.5. *Elliptic spaces* $\mathbb{E}\mathrm{ll}^n$ are "spaces of constant positive curvature" (what this means is explained in differential geometry courses). We will only study the two-dimensional case, i.e., the elliptic plane, in the present chapter after we are done with *spherical geometry*, which is the main topic of this chapter, but can also be regarded as the principal building block of elliptic plane geometry.

6.1.6. *Projective spaces* $\mathbb{R}P^n$ are obtained from affine spaces by "adding points at infinity" in a certain way, and taking, for the transformation group, a group of linear transformations on the so-called "homogeneous coordinates" of points $(x_1 : \cdots : x_n : x_{n+1}) \in \mathbb{R}P^n$. We can write this geometry as

$$\boxed{(\mathbb{R}P^n : \mathrm{Proj}(n))}.$$

For arbitrary n, projective geometry is usually studied in linear algebra courses. We will study the *projective plane* $\mathbb{R}P^2$ in this course, and only have a quick glance at projective space $\mathbb{R}P^3$ (see Chapter 12).

6.2. Some basic facts from Euclidean plane geometry

Here we list several fundamental facts of Euclidean plane geometry (including modern formulations of some of Euclid's postulates) in order to compare and contrast them with the corresponding facts of spherical, elliptic, and hyperbolic geometry.

I. *There exist a unique (straight) line passing through two given distinct points.*

II. *There exists a unique perpendicular to a given line passing through a given point.* (A *perpendicular* to a given line is a line forming four equal angles, called *right angles*, with the given one.)

III. *There exists a unique circle of given center and given radius.*

IV. *Given a point on a line and any positive number, there exist exactly two points on the line whose distance from the given point is equal to the given number.*

V. *There exists a unique parallel to a given line passing through a given point not on the given line.* (A *parallel* to a given line is a line without common points with the given one.) This is the modern version of Euclid's fifth postulate, sometimes described as the single most important and controversial scientific statement of all time.

VI. *The parameters of a triangle ABC, namely the angles α, β, γ at the vertices A, B, C and the sides a, b, c opposite to these vertices, satisfy the following formulas.*

(i) *Angle sum formula:* $\alpha + \beta + \gamma = \pi$.

(ii) *Sine formula:*

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

(ii) *Cosine formula:* $c^2 = a^2 + b^2 - 2ab\cos\gamma$.

6.3. Lines, distances, angles, polars, and perpendiculars on $\mathbb{S}^2$

Let $\mathbb{S}^2$ be the unit sphere in $\mathbb{R}^3$:

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

our present aim is to study the geometry $(\mathbb{S}^2 : \mathrm{O}(3))$, where $\mathrm{O}(3)$ is the orthogonal group (i.e., the group of isometries of $\mathbb{R}^3$ leaving the origin in place).

6.3.1. *Basic definitions.* By a *line* on the sphere we mean a great circle, i.e., the intersection of $\mathbb{S}^2$ with a plane passing through the sphere's center. For example, the equator of the sphere, as well as any meridian, is a line. The *angle* between two lines is defined as the dihedral angle (measured in radians) between the two planes containing the lines. For example, the angle between the equator and any meridian is $\pi/2$. The *distance* between two points A and B is defined as the measure (in radians) of the angle AOB. Thus the distance between the North and South Poles is π, the distance between the South Pole and any point on the equator is $\pi/2$.

Obviously, *the transformation group $O(3)$ preserves distances between points.* It can also be shown (we omit the proof) that, conversely, O(3) can be characterized as the group of distance-preserving transformations of the sphere (distance being understood in the spherical sense, i.e., as explained above).

6.3.2. *Poles, polars, perpendiculars, circles.* Let us look at the analogs in spherical geometry of the Euclidean postulates.

$\mathbf{I}_S$. *There exist a unique line passing through two given distinct points, except when the two points are antipodal, in which case there are infinitely many.* All the meridians joining the two poles give an example of the exceptional situation.

$\mathbf{II}_S$. *There exists a unique perpendicular to a given line passing through a given point, except when the point lies at the intersection of the perpendicular constructed from the center O of the sphere to the plane in which the line lies, in which case there are infinitely many such perpendiculars.* The exceptional

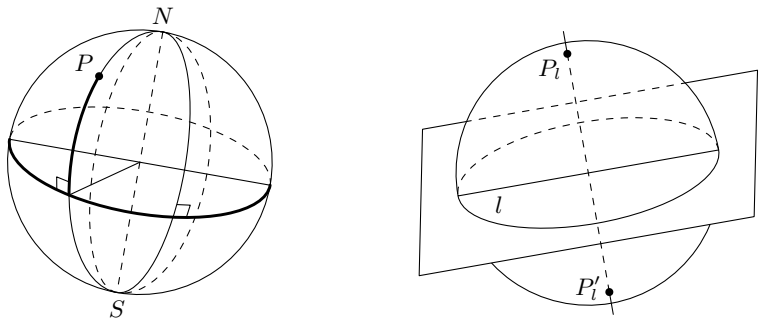


Figure 6.1. Perpendiculars, poles, and polars

situation is exemplified by the equator and, say, the North Pole: all the meridians (which all pass through the pole) are perpendicular to the equator.

More generally, the *polar* of a point P is the (spherical) line obtained by cutting the sphere by the plane passing through O and perpendicular to the (Euclidean!) straight line PO. Conversely, given a (spherical) line l, the *poles* of that line are the two antipodal points P_l and P_l' for which the (Euclidean) line P_lP_l' is perpendicular to the plane determined by l. The assertion II_S may now be restated as follows: *there exists a unique perpendicular to a given line passing through a given point, except when the point is a pole of that line, in which case all the lines passing through the pole are perpendicular to the given line.*

III$_S$. *There exists a unique circle of given center C and given radius ρ, provided $0 < \rho < \pi$.* It is defined as the set of points whose (spherical) distance from C is equal to ρ. It is easy to see that any (spherical) circle is actually a Euclidean circle, namely the one obtained as the intersection of the sphere with the plane perpendicular to the Euclidean line OC and passing through the point I on that line such that $OI = \cos\rho$. Note that the radius of the Euclidean circle will be less than ρ.

Given a spherical circle of center C and radius ρ, note that it can be regarded as the circle of radius $\pi - \rho$ and center C', where C' is the antipode of C. Further, note that the longest circle centered at C is the polar of the point C; its radius is $\pi/2$.

IV$_S$. Given a point on a line and any positive number, *there exist exactly two points on the line whose distance from the given point is equal to the given number, provided the number is less than π.*

V$_S$. Any two lines intersect in two *antipodal* points, i.e., in two points symmetric with respect to the center of the sphere $\mathbb{S}^2$. Therefore *there are no parallel lines in spherical geometry.* If two points A, B are not antipodal, then there is only one line joining them and one shortest line segment with endpoints at A and B. For opposite points, there is an infinity of lines joining them (for the North and South poles, these lines are the meridians).

6.3.3. *Lines as shortest paths.* It is proved in differential geometry courses that *spherical lines are geodesics*, i.e., they are the shortest paths between two points. To do this, one defines the length of a curve as a curvilinear integral and uses the calculus of variations to show that the curve (on the sphere) of minimal length joining two given points is indeed the arc of the great circle containing these points.

6.4. Biangles and triangles in $\mathbb{S}^2$

6.4.1. *Biangles.* Two lines l and m on the sphere intersect in two (antipodal) points P and P' and divide the sphere into four domains; each of them is called a *biangle*, it is bounded by two halves of the lines l and m, called its *sides*, and has two *vertices* (the points P and P'). The four domains form two congruent pairs; two biangles from a congruent pair touch each other at the common vertices P and P', and have the same angle at P and P'. The main parameter of a biangle is the measure α of the angle between the lines that determine it; if $\alpha \neq \pi/2$, the two biangles not congruent to the biangle of measure α are called *complementary*, their angle is $\pi - \alpha$. Note that the angle measure α determines the corresponding biangle up to an isometry of the sphere.

6.4.2. *Areas of figures in the sphere.* In order to correctly measure areas of figures on the plane, on the sphere, or on other surfaces, one must define what an area is, specify what figures are measurable (i.e., possess an area), and devise methods for computing areas. For the Euclidean plane, there are several approaches to area: many readers have probably heard of the theory of *Jordan measure*; more advanced readers may have studied *Lebesgue measure*; readers who have taken multivariable calculus courses know that areas may be computed by means of *double integrals*.

In this book, we will not develop a rigorous measure theory for the geometries that we study. In this subsection, we merely sketch an axiomatic approach for determining areas of spherical figures; this approach is similar to Jordan measure theory in the Euclidean plane. The theory says that there is a family of sets in $\mathbb{S}^2$, called *measurable*, satisfying the following axioms.

(i) *Invariance.* Two congruent measurable figures have the same area.

(ii) *Normalization.* The whole sphere is measurable and its area is 4π.

(iii) *Countable additivity.* If a measurable figure F is the union of a countable family of measurable figures $\{F_i\}$ without common interior points, then its area is equal to the sum of areas of the figures F_i.

An obvious consequence of these axioms is that the area of the North hemisphere is 2π, while each of the triangles obtained by dividing the hemisphere into four equal parts is of area $\pi/2$.

6.4.3. *Area of the biangle.* From the axioms formulated in the previous subsection, it is easy to deduce that the area $S_{\pi/2}$ of a biangle with angle measure $\pi/2$ is π. Indeed, the sphere is covered by four such non-overlapping

biangles, which are congruent to each other; they have the same area by (i), the sum of their areas is that of the sphere by (iii), and the latter is 4π by (ii), whence $S_{\pi/2} = (4\pi)/4 = \pi$.

For the case in which the angle measure α of a biangle is a rational multiple of π, a similar argument shows that

$$\boxed{S_\alpha = 2\alpha}. \tag{6.1}$$

This formula is actually true for any α, but for the case in which π/α is irrational, its proof requires a passage to the limit based on an additional "continuity axiom" that we have not explicitly stated. We therefore omit the proof, but will use the above formula for all values of α in what follows.

6.4.4. *Area of the triangle.* Let A, B, C be three distinct points of $\mathbb{S}^2$, no two of which are opposite. The union of the shortest line segments joining the points A and B, B and C, C and A is called the *triangle* ABC. For a triangle ABC, we always denote by α, β, γ the measure of the angles at A, B, C respectively and by a, b, c the lengths of the sides opposite to A, B, C (recall that the length a of BC is equal to the measure of the angle BOC in $\mathbb{R}^3$).

6.4.5. Theorem. *The area S_{ABC} of a spherical triangle with angles α, β, γ is equal to*

$$\boxed{S_{ABC} = \alpha + \beta + \gamma - \pi}.$$

Proof. There are 12 spherical biangles formed by pairs of lines AB, BC, CA. Choose six of them, namely those that contain triangle ABC or the antipodal triangle $A_1B_1C_1$ formed by the three points antipodal to A, B, C. Denote their areas by

$$S_I, S'_I, S_{II}, S'_{II}, S_{III}, S'_{III}.$$

Each point of the triangles ABC and $A_1B_1C_1$ is covered by exactly three of the chosen six biangles, while the other points of the sphere are covered by exactly one such biangle (we ignore the points on the lines). Therefore, using relation (6.1), we can write

$$\begin{aligned} 4\pi &= S_I + S'_I + S_{II} + S'_{II} + S_{III} + S'_{III} - 2S_{ABC} - 2S_{A_1B_1C_1} \\ &= 2\alpha + 2\beta + 2\gamma + 2\alpha + 2\beta + 2\gamma - 2S_{ABC} - 2S_{A_1B_1C_1} \\ &= 4(\alpha + \beta + \gamma) - 4S_{ABC}, \end{aligned}$$

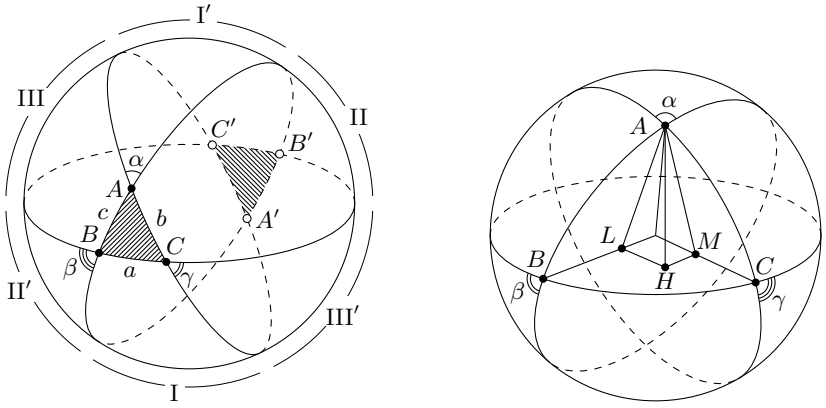


Figure 6.2. Area and sine theorem for the triangle

because the two triangles ABC $A_1B_1C_1$ have the same area (since they are congruent). Clearly, the previous formula implies the required equality. □

This theorem has the following fundamental consequence.

6.4.6. Corollary. *The sum of angles of any triangle is more than π.*

The analog of the sine formula for the Euclidean triangle is the following statement about spherical triangles.

6.4.7. Theorem. (The spherical sine theorem.)

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

In order to establish this formula, we will use the following statement, sometimes called the "theorem on the three perpendiculars".

6.4.8. Lemma. *Let $A \in \mathbb{R}^3$ be a point outside a plane $\mathcal{P}$, let K be its perpendicular projection on $\mathcal{P}$ and let L be its perpendicular projection on a line l contained in $\mathcal{P}$. Then KL is perpendicular to l.*

Proof of the lemma. The line l is perpendicular to the plane AKL because it is perpendicular to two nonparallel lines of AKL, namely to AL and AK (to the latter since AK is orthogonal to any line in $\mathcal{P}$). Therefore l is perpendicular to any line of the plane AKL, and in particular to LK. □

Proof of the theorem. Let H be the projection of A on the plane ABC, let L and M be the projections of A on the lines OB and OC. Then by

the lemma, L and M coincide with the projections of H on OB and OC. Therefore,

$$AH = LA\sin\beta = \sin c\sin\beta, \quad AH = MA\sin\gamma = \sin b\sin\gamma.$$

Thus, $\sin b : \sin\beta = \sin c : \sin\gamma$. Similarily, by projecting C on the plane AOB and arguing as above, we obtain $\sin b : \sin\beta = \sin a : \sin\alpha$. This immediately implies the required equality. $\square$

6.5. Other theorems about triangles.

In this section, we state a few more theorems about spherical triangles. Their proofs are relegated to the exercises appearing at the end of this chapter.

6.5.1. Theorem. (The first cosine theorem.)

$$\boxed{\cos a = \cos b\cos c + \sin b\sin c\cos\alpha}$$

6.5.2. Theorem. (The second cosine theorem.)

$$\boxed{\cos\alpha + \cos\beta\cos\gamma = \sin\beta\sin\gamma\cos c}$$

6.5.3. Corollary. (Analog of the Pythagoras theorem.) *If triangle ABC has a right angle at C, then*

$$\boxed{\cos c = \cos a\cos b}.$$

6.5.4. Theorem. *The medians of any triangle intersect at a single point.*

6.5.5. Theorem. *The altitudes of any triangle intersect at a single point.*

6.6. Coxeter triangles on the sphere $\mathbb{S}^2$

We will not develop the theory of tilings on the sphere $\mathbb{S}^2$ and Coxeter geometry on the sphere in full generality, but only consider *Coxeter triangles*, i.e., spherical triangles all of whose angles are of the form π/m, $m = 2, 3, \ldots$. It follows from Theorem 6.4.5 that any spherical Coxeter triangle $(\pi/p, \pi/q, \pi/r)$, N copies of which cover the sphere, must satisfy the Diophantine equation

$$N/p + N/q + N/r = N + 4.$$

The transformation group of the corresponding Coxeter geometry is finite, and so Theorem 3.2.6 tells us what group it has to be: it must be either one of the dihedral groups, or the tetrahedral, hexahedral, or dodecahedral group. The dihedral groups yield an obvious infinite series of tilings, one of which is shown in Figure 6.3.

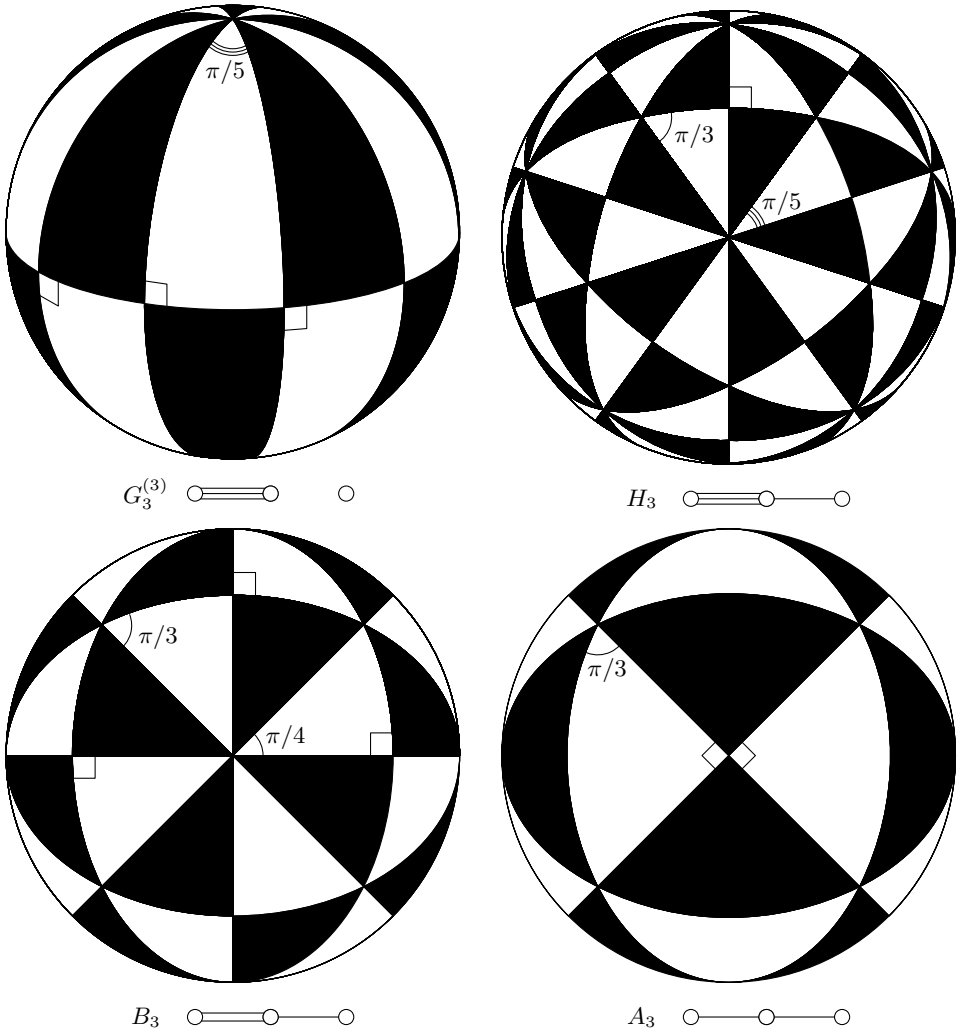


Figure 6.3. Four Coxeter tilings of the sphere

The three other groups yield three possibilities for N: $N = 24$, 48, 120, and we easily find the corresponding values of (p, q, r) in each of the three

cases. Finally, the solutions of our Diophantine equation are:

$$(2,3,3),(2,3,4),(2,3,5),(2,2,n) \quad \text{for} \quad n=2,3,\dots.$$

The corresponding tilings of the sphere (and their Coxeter schemes) are shown in Figure 6.3.

6.7. Two-dimensional elliptic geometry

6.7.1. Spherical geometry is closely related to the *elliptic geometry* invented by Riemann. Elliptic geometry is obtained from spherical geometry by "identifying opposite points of $\mathbb{S}^2$". The precise definition can be stated as follows. Consider the set $\mathbb{E}\mathrm{ll}^2$ whose elements are pairs of antipodal points $(x,-x)$ on the unit sphere $\mathbb{S}^2 \subset \mathbb{R}^3$. The group $O(3)$ acts on this set (because isometries of $\mathbb{S}^2$ take antipodal pairs of points to antipodal pairs), thus defining a geometry in the sense of Klein $(\mathbb{E}\mathrm{ll}^2 : \mathrm{O}(3))$, which we call *two-dimensional elliptic geometry*.

Lines in elliptic geometry are defined as great circles of the sphere $\mathbb{S}^2$, angles and distances are defined as in spherical geometry, and the trigonometry of triangles in elliptic geometry is the same as in spherical geometry. More generally, one can say that elliptic geometry is locally the same as spherical, but these geometries are drastically different globally. In particular, in elliptic geometry

- *one and only line passes through any two distinct points*;
- *for a given line and any given point* (*except one, called the pole of that line*) *there exists a unique perpendicular to that line passing through the point*.

The relationship between the two geometries is best expressed by the following statement, which yields simple proofs of the statements about elliptic geometry made above,

6.7.2. Theorem. *There exists a surjective morphism*

$$D : (\mathbb{S}^2 : \mathrm{O}(3)) \to (\mathbb{E}\mathrm{ll}^2 : \mathrm{O}(3)),$$

of spherical geometry onto elliptic geometry which is a local isomorphism (*in the sense that any domain contained in a half-sphere is mapped bijectively and isometrically onto its image*).

Proof. The map D is the obvious one: $D : x \mapsto (x,-x)$, while the homomorphism of the transformation groups is the identity isomorphism. All the assertions of the theorem are immediate. □

As we noted above, globally the two geometries are very different. Being metric spaces, they are topological spaces (in the metric topology) which are not even homeomorphic: one is a two-sided surface ($\mathbb{S}^2$), the other ($\mathbb{R}P^2$) is one-sided (it contains a Möbius strip).

6.8. Problems

In all the problems below a, b, c are the sides and α, β, γ are the opposite angles of a spherical triangle. The radius of the sphere is $R = 1$.

6.1. Prove the first cosine theorem on the sphere $\mathbb{S}^2$:

$$\cos a = \cos b \cos c + \sin b \sin c \cos \alpha.$$

6.2. Prove the second cosine theorem on the sphere $\mathbb{S}^2$:

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

6.3. Prove that $a + b + c < 2\pi$.

6.4. Does the Pythagorean theorem hold in spherical geometry? Prove the analogs of that theorem stated in Corollary 6.5.3.

6.5. Does the Moscow–New York flight fly over Spain? Over Greenland? Check your answer by stretching a thin string between Moscow and NY on a globe.

6.6. Find the infimum and the supremum of the sum of the angles of an equilateral triangle on the sphere.

6.7. The city A is located at the distance 1000km from the cities B and C, the trajectories of the flights from A to B and from A to C are perpendicular to each other. Estimate the distance between B and C. (You can take the radius of the Earth equal to 6400km)

6.8*. Find the area of the spherical disk of radius r (i.e., the domain bounded by a spherical circle of radius r).

6.9. Find fundamental domains for the actions of the isometry groups of the tetrahedron, the cube, the dodecahedron, and the icosahedron on the 2-sphere and indicate the number of their images under the corresponding group action.

6.10. Prove that any spherical triangle has a circumscribed and an inscribed circle.

6.11. Prove that the medians of a spherical triangle intersect at one point.

6.12. Prove that the altitudes of a spherical triangle always intersect at one point.

6.13. Suppose that the medians and the altitudes of a spherical triangle interest at the points M and A respectively. Can it happen that $M = A$?

Chapter 7

THE POINCARÉ DISK MODEL OF HYPERBOLIC GEOMETRY

In this chapter, we begin our study of the most popular of the non-Euclidean geometries – hyperbolic geometry, concentrating on the case of dimension two. We avoid the intricacies of the axiomatic approach (which will only be sketched in Chapter 10) and define hyperbolic plane geometry via the beautiful Poincaré disk model, which is the geometry of the disk determined by the action of a certain transformation group acting on the disk (namely, the group generated by reflections in circles orthogonal to the boundary of the disk).

In order to describe the model, we need some facts from Euclidean plane geometry, which should be studied in high school, but in most cases unfortunately aren't. So we begin by recalling some properties of inversion (which will be the main ingredient of the transformation group of our geometry) and some constructions related to orthogonal circles in the Euclidean plane. We then establish the basic facts of hyperbolic plane geometry and finally digress, following Poincaré's argumentation from his book *Science et Hypothèse* (for the English version, see [12]) about epistomological questions relating this geometry (and other geometries) to the physical world.

7.1. Inversion and orthogonal circles

7.1.1. *Inversion and its properties.* The main tool that we will need in this chapter is inversion, a classical transformation from elementary plane geometry. Denote by $\mathcal{R}$ the plane $\mathbb{R}^2$ with an added extra point (called the *point at infinity* and denoted by ∞). The set $\mathcal{R} := \mathbb{R}^2 \cup \infty$ can also be interpreted as the complex numbers $\mathbb{C}$ with the "point at infinity" added; it is then called the *Riemann sphere* and denoted by $\overline{\mathbb{C}}$.

An *inversion* of center $O \in \mathbb{R}^2$ and radius $r > 0$ is the transformation of $\mathcal{R}$ that maps each point M to the point N on the ray OM so that

$$\boxed{|OM| \cdot |ON| = r^2} \tag{7.1}$$

and interchanges the points O and ∞. Sometimes inversions are called *reflections* with respect to the *circle of inversion*, i.e., the circle of radius r centered at O.

There is a simple geometric way of constructing the image of a point M under an inversion of center O and radius r: draw the circle of inversion, lower the perpendicular to OM from M to its intersection point T with the circle and construct the tangent to the circle at T to its intersection point N with the ray OM; then N will be the image of M under the given inversion. Indeed, the two right triangles OMT and OTN are similar (they have a common acute angle at O), and therefore

$$\frac{|OM|}{|OT|} = \frac{|OT|}{|ON|},$$

and since $|OT| = r$, we obtain (7.1).

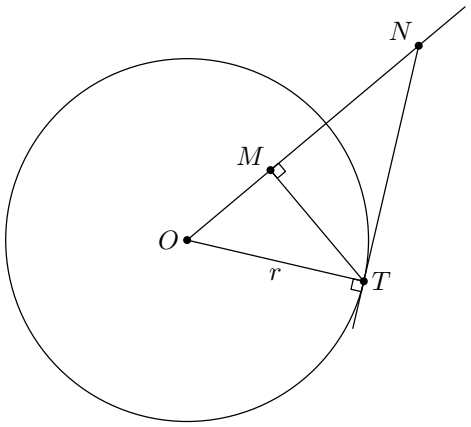


Figure 7.1. Inversion $|OM| \cdot |ON| = r^2$

If the extended plane $\mathcal{R}$ is interpreted as the Riemann sphere $\overline{\mathbb{C}}$, then an example of an inversion (of center O and radius 1) is the map $z \mapsto 1/\overline{z}$, where the bar over z denotes complex conjugation.

It follows immediately from the definition that inversions are bijections of $\mathcal{R} = \overline{\mathbb{C}}$ that leave the points of the circle of inversion in place, "turn the circle inside out" in the sense that points inside the circle are taken to points outside it (and vice versa), and are *involutions* (i.e., the composition of an inversion with itself is the identity). Further, inversions possess the following important properties.

(i) *Inversions map any circle or straight line into a circle or straight line.* In particular, lines passing through the center of inversion are mapped to

themselves (but are "turned inside out" in the sense that O goes to ∞ and vice versa, while the part of the line inside the circle of inversion goes to the outside part and vice versa); circles passing through the center of inversion are taken to straight lines, while straight lines not passing through the center of inversion are taken to circles passing through that center (see Fig.7.2).

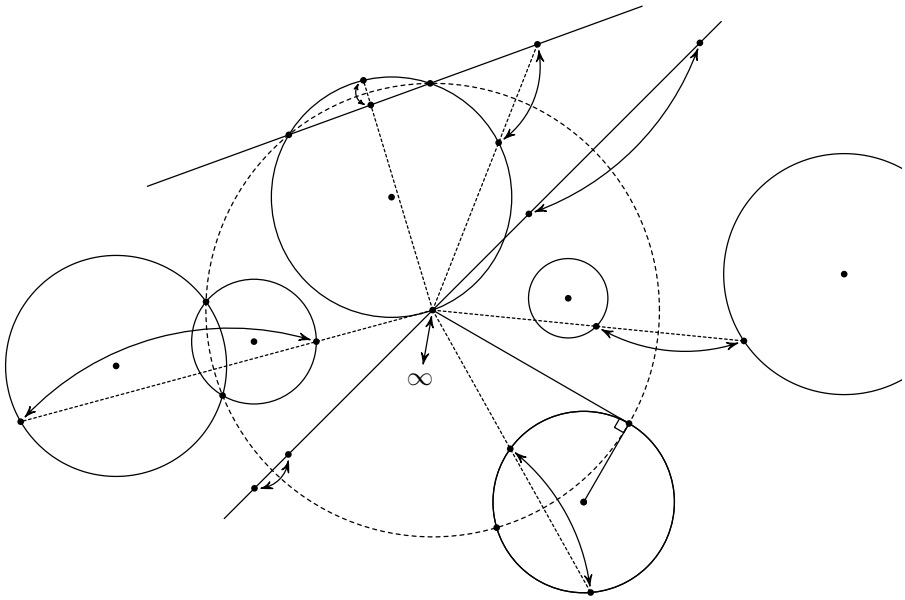


Figure 7.2. Images of circles and lines under inversion

(ii) *Inversions preserve* (*the measure of*) *angles;* here by the measure of an angle formed by two intersecting curves we mean the ordinary (Euclidean) measure of the angle formed by their tangents at the intersection point.

(iii) *Inversions map any circle or straight line orthogonal to the circle of inversion into itself.* Look at Fig.7.3, which shows two orthogonal circles $\mathcal{C}_O$ and $\mathcal{C}_I$ of centers O and I, respectively.

It follows from the definition of orthogonality that the tangent from the center O of $\mathcal{C}_O$ to the other circle $\mathcal{C}_I$ passes through the intersection point T of the two circles. Now let us consider the inversion of center O and radius $r = |OT|$. According to property (iii) above, it takes the circle $\mathcal{C}_I$ to itself; in particular, the point M is mapped to N, the point T (as well as the other intersection point of the two circles) stays in place, and the two arcs of $\mathcal{C}_I$ cut out by $\mathcal{C}_O$ are interchanged. Note further that, vice versa, the inversion

in the circle $\mathcal{C}_I$ transforms $\mathcal{C}_O$ in an analogous way.

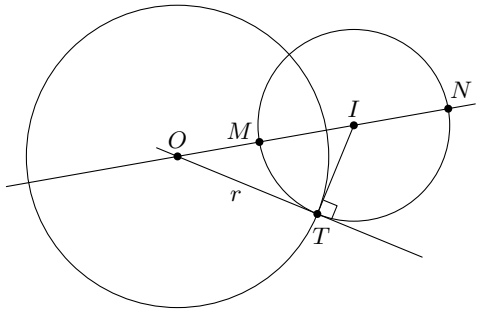


Figure 7.3. Orthogonal circles

The (elementary) proofs of properties (i)–(iii) are left to the reader (see Exercises 7.1–7.3).

7.1.2. *Construction of orthogonal circles.* We have already noted the important role that orthogonal circles play in inversion (see 7.1.1.(iii)). Here we will describe several constructions of orthogonal circles that will be used in subsequent sections.

7.1.3. Lemma. *Let A be a point inside a circle $\mathcal{C}$ centered at some point O; then there exists a circle orthogonal to $\mathcal{C}$ such that the reflection in this circle takes A to O.*

Proof. From A draw the perpendicular to line OA to its intersection T with the circle $\mathcal{C}$ (see Fig.7.4).

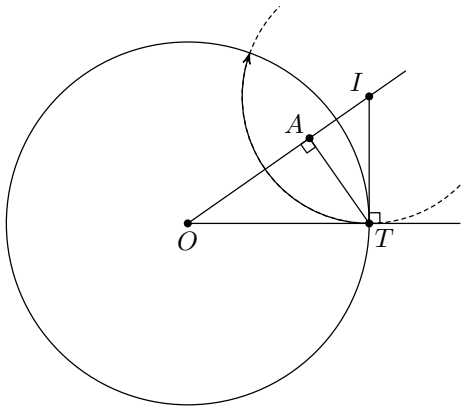


Figure 7.4. Inversion taking an arbitrary point A to O

Draw the tangent to $\mathcal{C}$ at T to its intersection at I with OA. Then the circle of radius IT centered at I is the one we need. Indeed, the similar right triangles IAT and ITO yield $|IA|/|IT| = |IT|/|IO|$, whence we obtain $|IA| \cdot |IO| = |IT|^2$, which means that O is the reflection of A in the circle of radius $|IT|$ centered at I, as required. $\square$

7.1.4. Corollary. (i) *Let A and B be points inside a circle $\mathcal{C}_0$ not lying on the same diameter; then there exists a unique circle orthogonal to $\mathcal{C}_0$ and passing through A and B.*

(ii) *Let A be a point inside a circle $\mathcal{C}_0$ and P a point on $\mathcal{C}_0$, with A and P not lying on the same diameter; then there exists a unique circle orthogonal to $\mathcal{C}_0$ passing through A and P.*

(iii) *Let P and Q be points on a circle $\mathcal{C}_0$ of center O such that PQ is not a diameter; then there exists a unique circle $\mathcal{C}$ orthogonal to $\mathcal{C}_0$ and passing through P and Q.*

(iv) *Let A be a point inside a circle $\mathcal{C}_0$ of center O and $\mathcal{D}$ be a circle orthogonal to $\mathcal{C}_0$; then there exists a unique circle $\mathcal{C}$ orthogonal to both $\mathcal{C}_0$ and $\mathcal{D}$ and passing through A.*

Proof. To prove (i), we describe an effective step-by-step construction, which can be carried out by ruler and compass, yielding the required circle. The construction is shown on Figure 7.5, with the numbers in parentheses near each point indicating at which step the point was obtained.

First, we apply Lemma 7.1.3, to define an inversion φ taking A to the center O of the given circle; to do this, we lower a perpendicular from A to OA to its intersection T (1) with $\mathcal{C}$, then draw the perpendicular to OT from T to its intersection I (2) with OA; the required inversion is centered at I and is of radius $|IT|$. Joining B and I, we construct the tangent BS (3) to the circle of the inversion φ and find the image B' (4) of B under φ by dropping a perpendicular from S to IB.

Next, we draw the line $B'O$ and obtain the intersection points M, N of this line with the circle of the inversion φ. Finally, we draw the circle $\mathcal{C}$ passing through the points M, N, I. Then $\mathcal{C}$ "miraculously" passes through A and B and is orthogonal to $\mathcal{C}_0$! Of course, there is no miracle in this: $\mathcal{C}$ passes through A and B because it is the inverse image under φ of the line OB' (see 7.1.1(i)), it is orthogonal to $\mathcal{C}_0$ since so is OB' (see 7.1.1(ii)).

Uniqueness is obvious in the case $A = O$ and follows in the general case by 7.1.1(i)-(ii). $\square$

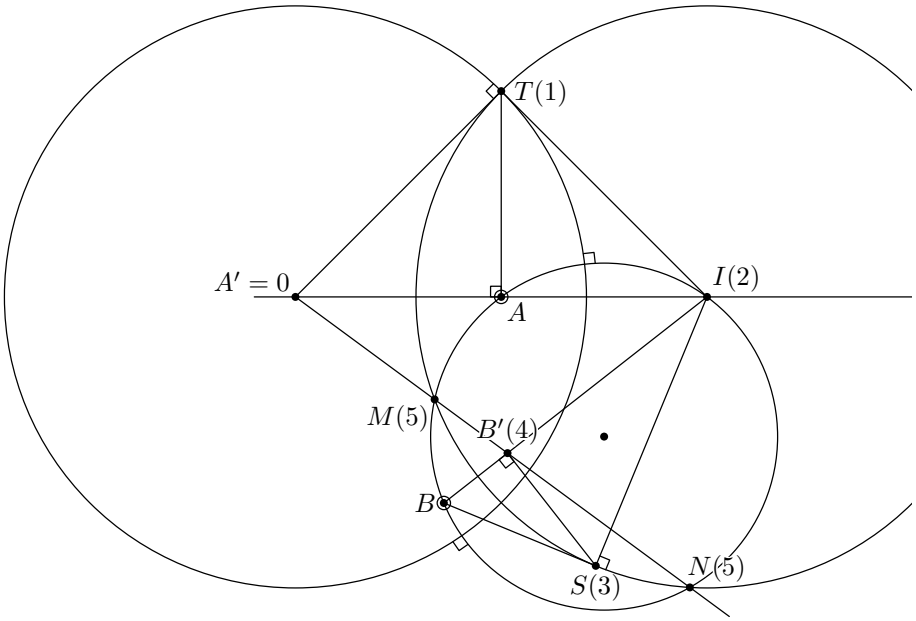


Figure 7.5. Circle orthogonal to $\mathcal{C}_0$ containing A, B

The proof of (ii) is analogous: we send A to O by an inversion φ, join O and $\varphi(P)$ and continue the argument as above. □

To prove (iii), construct lines OP and OQ, draw perpendiculars to these lines from P and Q respectively and denote by I their intersection point. Then the circle of radius $|IP|$ centered at I is the required one. Its uniqueness is easily proved by contradiction. □

To prove (iv), we again use Lemma 7.1.3 to construct an inversion φ that takes $\mathcal{C}_0$ to itself and sends A to O. From the point O, we draw the (unique) ray $\mathcal{R}$ orthogonal to $\varphi(\mathcal{L})$. Then the circle $\varphi^{-1}(\mathcal{R})$ is the required one. □

7.2. Definition of the disk model

7.2.1. The disk model of the *hyperbolic plane* is the geometry $(\mathbb{H}^2 : \mathcal{M})$ whose points are the points of the open disk

$$\mathbb{H}^2 := \big\{(x, y) \in \mathbb{R}^2 \,|\, x^2 + y^2 < 1\big\},$$

and whose transformation group $\mathcal{M}$ is the group generated by reflections in all the circles orthogonal to the boundary circle $\mathbb{A} := \{(x, y) : x^2 + y^2 = 1\}$ of $\mathbb{H}^2$, and by reflections in all the diameters of the circle $\mathbb{A}$. Now $\mathcal{M}$ is indeed

a transformation group of $\mathbb{H}^2$: the discussion in 7.1.1 implies that a reflection of the type considered takes points of $\mathbb{H}^2$ to points of $\mathbb{H}^2$ and, being its own inverse, we have the implication $\varphi \in \mathcal{M} \Longrightarrow \varphi^{-1} \in \mathcal{M}$.

We will often call $\mathbb{H}^2$ the *hyperbolic plane*. The boundary circle $\mathbb{A}$ (which is not part of the hyperbolic plane) is called the *absolute*.

7.2.2. We will see later that $\mathcal{M}$ is actually the isometry group of hyperbolic geometry with respect to the *hyperbolic distance*, which will be defined in the next chapter. We will see that although the Euclidean distance between points of $\mathbb{H}^2$ is always less than 2, the hyperbolic plane is unbounded with respect to the hyperbolic distance. Endpoints of a short segment (in the Euclidean sense!) near the absolute are very far away from each other in the sense of hyperbolic distance.

Figure 7.6 gives an idea of what an isometric transformation (the simplest one – a reflection in a line) does to a picture. Note that from our Euclidean point of view, the reflection changes the size and the shape of the picture, whereas from the hyperbolic point of view, the size and shape of the image is exactly the same as that of the original. It should also be clear that *hyperbolic reflections reverse orientation*.

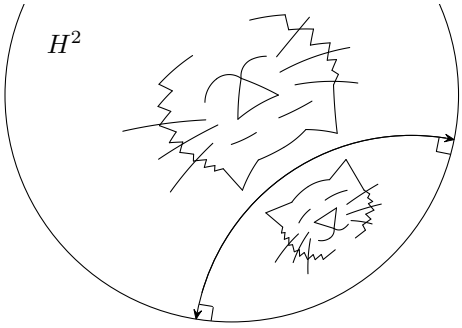


Figure 7.6. An isometry in the hyperbolic plane

7.3. Points and lines in the hyperbolic plane

7.3.1. First we define *points of the hyperbolic plane* simply as points of the open disk $\mathbb{H}^2$. We then define the *lines* on the hyperbolic plane as the intersections with $\mathbb{H}^2$ of the (Euclidean) circles orthogonal to the absolute as well as the diameters (without endpoints) of the absolute (see Fig.7.7).

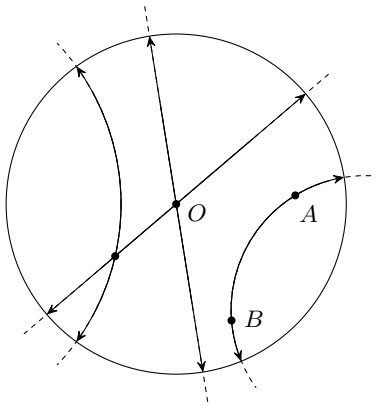


Figure 7.7. Lines on the hyperbolic plane

Note that the endpoints of the arcs and the diameters do not belong to the hyperbolic plane: they lie in the absolute, whose points are not points of our geometry.

Figure 7.7 shows that some lines intersect in one point, others have no common points, and none have two common points (unlike lines in spherical geometry). This is not surprising, because we have the following statement.

7.3.2. Theorem. *One and only one line passes through any pair of distinct points of the hyperbolic plane.*

Proof. The theorem immediately follows from Corollary 7.1.4, (i). □

7.4. Perpendiculars

7.4.1. Two lines in $\mathbb{H}^2$ are called *perpendicular* if they are orthogonal in the sense of elementary Euclidean geometry. When both are diameters, they are perpendicular in the usual sense, when both are arcs of circles, they have perpendicular tangents at the intersection point, when one is an arc and the other a diameter, then the diameter is perpendicular to the tangent to the arc at the intersection point.

7.4.2. Theorem. *There is one and only one line passing through a given point and perpendicular to a given line.*

Proof. The theorem immediately follows from Corollary 7.1.4, (iv). □

7.5. Parallels and nonintersecting lines

7.5.1. Let l be a line and P be a point of the hyperbolic plane $\mathbb{H}^2$ not contained in the line l. Denote by A and B the points at which l intersects

the absolute. Consider the lines $k = PA$ and $m = PB$ and denote their second intersection points with the absolute by A' and B'. Clearly, the lines k and m do not intersect l. Moreover, any line passing through P between k and m (i.e., any line containing P and joining the arcs AA' and BB') does not intersect l. The lines APA' and BPB' are called *parallels* to l through P, and the lines between them are called *nonintersecting lines* with l.

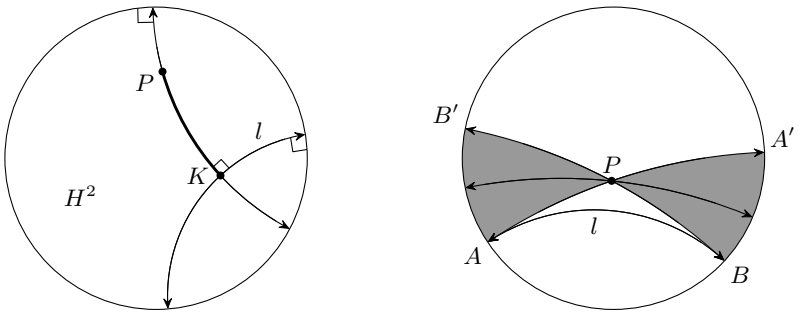


Figure 7.8. Perpendiculars and parallels

We have proved the following statement.

7.5.2. Theorem. *There are infinitely many lines passing through a given point P not intersecting a given line l if $P \notin l$. These lines are all located between the two parallels to l.* □

This theorem contradicts Euclid's famous *Fifth Postulate*, which, in its modern formulation, says that one and only one parallel to a given line passes through a given point. For more than two thousand years, many attempts to prove that the Fifth Postulate follows from Euclid's other postulates (which, unlike the Fifth Postulate, were simple and intuitively obvious) were made by mathematicians and philosophers. Had such a proof been found, Euclidean geometry could have been declared to be an absolute truth both from the physical and the philosophical point of view, it would have been an example of facts that the German philosopher Kant included in the category of *synthetic apriori*. For two thousand years, the naive belief among scientists in the absolute truth of Euclidean geometry made it difficult for the would be discoverers of other geometries to realize that they had found something worthwhile. Thus the appearance of a consistent geometry in which the Fifth Postulate does not hold was not only a crucial development in the history of mathematics, but one of the turning points in the philosophy of science. In this connection, see the discussion in Chapter 11.

7.6. Sum of the angles of a triangle

7.6.1. Consider three points A, B, C not on one line. The three segments AB, BC, CA (called *sides*) form a *triangle* with *vertices* A, B, C. The *angles* of the triangle, measured in radians, are defined as equal to the (Euclidean measure of the) angles between the tangents to the sides at the vertices.

7.6.2. Theorem. *The sum of the angles α, β, γ of a triangle ABC is less than two right angles:*

$$\boxed{\alpha + \beta + \gamma < \pi}.$$

Proof. In view of Lemma 7.1.3, we can assume without loss of generality that A is O (the center of $\mathbb{H}^2$). But then if we compare the hyperbolic triangle OBC with the Euclidean triangle OBC, we see that they have the same angle at O, but the Euclidean angles at B and C are larger than their hyperbolic counterparts (look at Fig.7.9), which implies the claim of the theorem. □

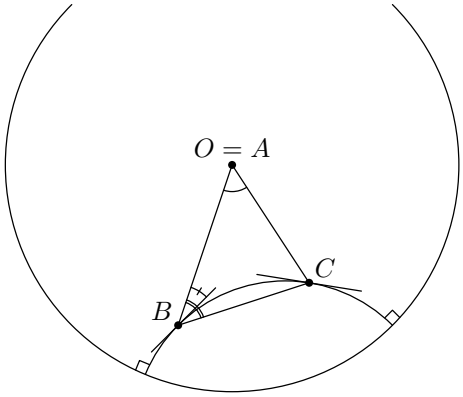


Figure 7.9. Sum of the angles of a hyperbolic triangle

It is easy to see that very small triangles have angles sums very close to π, in fact *the least upper bound of the angle sum of hyperbolic triangles is exactly π*. Further, *the greatest lower bound of these sums is 0*. To see this, divide the absolute into three equal arcs by three points P, Q, R and construct three circles orthogonal to the absolute passing through the pairs of points P and Q, Q and R, R and P. These circles exist by Corollary 7.1.2, item (iii). Then all the angles of the "triangle" PQR are zero, so its angle sum is zero. Of course, PQR is not a real triangle in our geometry (its vertices, being

on the absolute, are not points of $\mathbb{H}^2$), but if we take three points P', Q', R' close enough to P, Q, R, then the angle sum of triangle $P'Q'R'$ will be less than any prescribed $\varepsilon > 0$.

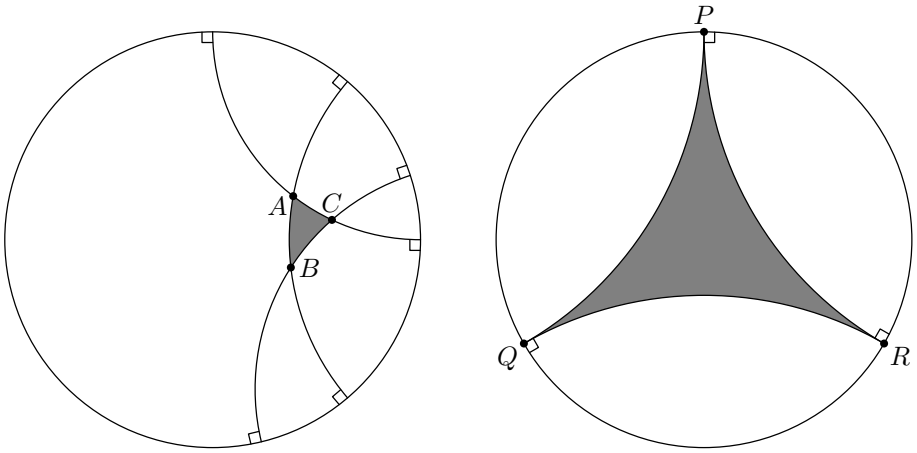


Figure 7.10. Ordinary triangle and "triangle" with angle sum 0

7.7. Rotations and circles in the hyperbolic plane

We mentioned previously that distance between points of the hyperbolic plane will be defined later. Recall that the hyperbolic plane is the geometry $(\mathbb{H}^2 : \mathcal{M})$, in which, by definition, $\mathcal{M}$ is the transformation group generated by all reflections in all the lines of $\mathbb{H}^2$. If we take the composition of two reflections in two intersecting lines, then what we get should be a "rotation", but we can't assert that at this point, because we don't have any definition of rotation: the usual (Euclidean) definition of a rotation or even that of a circle cannot be given until distance is defined.

But the notions of rotation and of circle *can* be defined without appealing to distance in the following natural way: a *rotation* about a point $P \in \mathbb{H}^2$ is, by definition, the composition of any two reflections in lines passing through P. If I and A are distinct points of $\mathbb{H}^2$, then the (hyperbolic) *circle* of center I and radius IA is the set of images of A under all rotations about I.

7.7.1. Theorem. *A (hyperbolic) circle in the Poincaré disk model is a Euclidean circle, and vice versa, any Euclidean circle inside $\mathbb{H}^2$ is a hyperbolic circle in the geometry $(\mathbb{H}^2 : \mathcal{M})$.*

Proof. Let $\mathcal{C}$ be a circle of center I and radius IA in the geometry $(\mathbb{H}^2 : \mathcal{M})$. Using Lemma 7.1.3, we can send I to the center O of $\mathbb{H}^2$ by a

reflection φ. Let ρ be a rotation about I determined by two lines l_1 and l_2. Then the lines $d_1 := \varphi(l_1)$ and $d_2 := \varphi(l_2)$ are diameters of the absolute and the composition of reflections in these diameters is a Euclidean rotation about O (and simultaneously a hyperbolic one). This rotation takes the point $\varphi(A)$ to a point on the circle $\mathcal{C}'$ of center O and radius $O\varphi(A)$, which is simultaneously a hyperbolic and Euclidean circle. Now by Corollary 7.1.4 item (i), the inverse image of $\varphi^{-1}(\mathcal{C}')$ will be a (Euclidean!) circle. But $\varphi^{-1}(\mathcal{C}')$ coincides with $\mathcal{C}$) by construction, so $\mathcal{C}$) is indeed a Euclidean circle in our model.

The proof of the converse assertion is similar and is left to the reader (see Exercise 7.7).

7.8. Hyperbolic geometry and the physical world

In his famous book *Science et Hypothèse*, Henri Poincaré describes the physics of a small "universe" and the physical theories that its inhabitants would create. The universe considered by Poincaré is Euclidean, plane (two-dimensional), has the form of an open unit disk. Its temperature is 100° Farenheit at the center of the disk and decreases linearly to absolute zero at its boundary. The lengths of objects (including living creatures) are proportional to temperature.

How will a little flat creature endowed with reason and living in this disk describe the main physical laws of his universe? The first question he/she may ask could be: Is the world bounded or infinite? To answer this question, an expedition is organized; but as the expedition moves towards the boundary of the disk, the legs of the explorers become smaller, their steps shorter – they will never reach the boundary, and conclude that the world is infinite.

The next question may be: Does the temperature in the universe vary? Having constructed a thermometer (based on different expansion coefficients of various materials), scientists carry it around the universe and take measurements. However, since the lengths of all objects change similarly with temperature, the thermometer gives the same measurement all over the universe – the scientists conclude that the temperature is constant.

Then the scientists might study straight lines, i.e., investigate what is the shortest path between two points. They will discover that the shortest path is what we perceive to be the arc of the circle containing the two points and orthogonal to the boundary disk (this is because such a circular path brings the investigator nearer to the center of the disk, and thus increases the length of his steps). Further, they will find that the shortest path is unique

and regard such paths as "straight lines".

Continuing to develop geometry, the inhabitants of Poincaré's little flat universe will decide that there is more than one parallel to a given line passing through a given point, the sum of angles of triangles is less than π, and obtain other statements of hyperbolic geometry.

Thus they will come to the conclusion that they live in an infinite flat universe with constant temperature governed by the laws of hyperbolic geometry. But this not true – their universe is a finite disk, its temperature is variable (tends to zero towards the boundary) and the underlying geometry is Euclidean, not hyperbolic!

The philosophical conclusion of Poincaré's argument is not agnosticism – he goes further. The physical model described above, according to Poincaré, shows not only that the truth about the universe cannot be discovered, but that it makes no sense to speak of any "truth" or approximation of truth in science – pragmatically, the inhabitants of his physical model are perfectly right to use hyperbolic geometry as the foundation of their physics because it is convenient, and it is counterproductive to search for any abstract Truth which has no practical meaning anyway.

This conclusion has been challenged by other thinkers, but we will not get involved in this philosophical discussion.

7.8. Problems

7.1. Prove that inversion maps circles and straight lines to circles or straight lines.

7.2. Prove that inversion maps any circle orthogonal to the circle of inversion into itself.

7.3. Prove that inversion is conformal (i.e., it preserves the measure of angles).

7.4. Prove that if P is point lying outside a circle γ and A, B are the intersection points with the circle of a line l passing through P, then the product $|PA| \cdot |PB|$ (often called the *power of P with respect to γ*) does not depend on the choice of l.

7.5. Prove that if P is point lying inside a circle γ and A, B are the intersection points with the circle of a line l passing through P, then the product $|PA| \cdot |PB|$ (often called the *power of P with respect to γ*) does not depend on the choice of l.

7.6. Prove that inversion with respect to a circle orthogonal to a given circle $\mathcal{C}$ maps the disk bounded by $\mathcal{C}$ bijectively onto itself.

7.7. Prove that any Euclidean circle inside the disk model is also a hyperbolic circle. Does the ordinary (Euclidean) center coincide with its "hyperbolic center"?

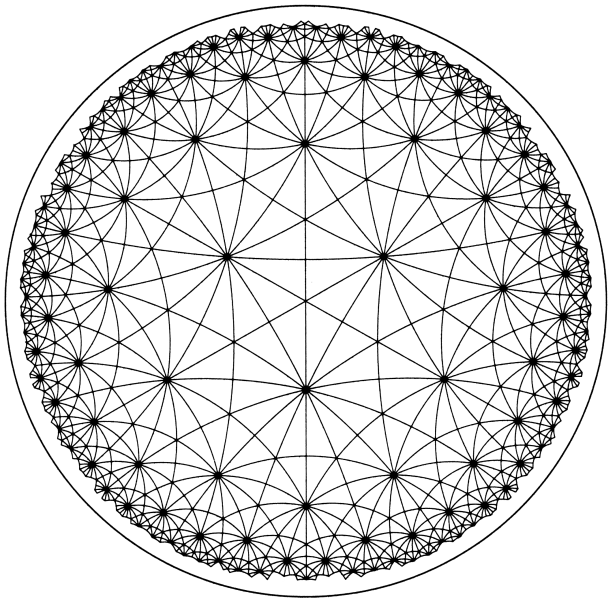


Figure 7.11. A pattern of lines in $\mathbb{H}^2$

7.8. Study Figure 7.11. Does it demonstrate any tilings of $\mathbb{H}^2$ by regular polygons? Of how many sides? Do you discern a Coxeter geometry in this picture with "hyperbolic Coxeter triangles" as fundamental domains? What are their angles?

7.9. Prove that any inversion of $\overline{\mathbb{C}}$ preserves the cross ratio of four points:

$$\langle z_1, z_2, z_3, z_4 \rangle := \frac{z_3 - z_1}{z_3 - z_2} \ : \ \frac{z_4 - z_1}{z_4 - z_2}.$$

7.10*. Using complex numbers, invent a formula for the distance between points on the Poincaré disk model and prove that "symetry with respect to straight lines" (i.e., inversion) preserves this distance.

7.11. Prove that hyperbolic geometry is homogeneous in the sense that for any two flags (i.e., half planes with a marked point on the boundary) there exists an isometry taking one flag to the other.

7.12. Prove that the hyperbolic plane (as defined via the Poincaré disk model) can be tiled by regular pentagons.

7.13. Define inversion (together with the center and the sphere of inversion) in Euclidean space $\mathbb{R}^3$, state and prove its main properties: inversion takes planes and spheres to planes or spheres, any sphere orthogonal to the sphere of inversion to itself, any plane passing through the center of inversion to itself.

7.14. Using the previous exercise, prove that any inversion in $\mathbb{R}^3$ takes circles and straight lines to circles or straight lines.

7.15. Prove that any inversion in $\mathbb{R}^3$ is conformal (preserves the measure of angles).

7.16. Construct a model of hyperbolic space geometry on the open unit ball (use Exercise 7.13).

7.17. Prove that there is a unique common perpendicular joining any two nonintersecting lines.

7.18. Let $A_\infty P$ and $A_\infty P'$ be two parallel lines (with A_∞ a point on the absolute). Given a point M on $A_\infty P$, we say that $M' \in A_\infty P'$ is the *corresponding point* to M if the angles $A_\infty MM'$ and $A_\infty M'M$ are equal. Prove that any point $M \in A_\infty P$ has a unique corresponding point on the line $A_\infty P'$.

7.19. The locus of all points corresponding to a point M on $A_\infty P$ and lying on all the parallels to $A_\infty P$ is known as a *horocycle*. What do horocycles look like in the Poincaré disk model?

Chapter 8
THE POINCARÉ HALF PLANE MODEL

In this chapter, we will present another model of the hyperbolic plane, also due to Poincaré. This model is also a geometry in the sense of Klein, and we will learn in subsequent chapters that it is actually isomorphic (as a geometry) to the disk model studied in Chapter 7.

The points of the half plane model are simply complex numbers with positive imaginary part (the part of the complex numbers that lies "above" the real axis). Such a configuration of points does not appear to be as symmetric as that of the disk, but the half plane model has the advantage that the elements of its transformation group (which is a concrete subgroup of the Möbius group of linear fractional transformations, see the definition below) are defined by simple explicit formulas and there is a neat formula for the distance between two points.

It will turn out that the isometry group with respect to this distance is actually the transformation group of the model, so that this model shows that hyperbolic geometry is a geometry in the traditional sense: its structure is defined by a distance function. This will allow us to study "hyperbolic trigonometry", and understand the meaning of certain mysterious "absolute constants" that arise in hyperbolic plane geometry.

In order to define the half plane-model, we will need to specify certain transformation groups acting on the Riemann sphere $\overline{\mathbb{C}} = \mathbb{C} \cup \infty$, and we begin this chapter by studying these transformations.

8.1. Affine and linear-fractional transformations of $\overline{\mathbb{C}}$

In this section, we will be studying various linear-fractional groups acting on the Riemann sphere $\overline{\mathbb{C}}$. An efficient tool in our constructions will be the notion of cross ratio, with which we begin.

8.1.1. *Cross ratio of four complex numbers.* The *cross-ratio* of four complex numbers $z_1, z_2, z_3, z_4 \in \mathbb{C}$ is defined as the number

$$\boxed{\langle z_1, z_2, z_3, z_4 \rangle := \frac{z_3 - z_1}{z_3 - z_2} : \frac{z_4 - z_1}{z_4 - z_2}}. \tag{8.1}$$

The cross ratio $\langle z_1, z_2, z_3, z_4 \rangle$ possesses the following properties.

8.1.2. *Affine transformations.* A transformation of $\overline{\mathbb{C}}$ onto itself of the form $z \mapsto az + b$, $\infty \mapsto \infty$, where $a, b \in \mathbb{C}$ and $a \neq 0$, is called *affine*. In particular, if $a = 1$, the corresponding affine transformation is a parallel translation (by the vector OB, where B is the point of the complex plane corresponding to the complex number b).

8.1.3. Theorem. *Affine transformations take straight lines to straight lines, circles to circles, and preserve angles and cross ratios.*

Proof. Denoting $a = re^{i\varphi}$, $r > 0$, we can write

$$z \mapsto e^{i\varphi}z \mapsto r(e^{i\varphi}z) \mapsto (re^{i\varphi}z) + b = az + b,$$

which shows that any affine transformation is the composition of a rotation (by the angle φ), a homothety (with coefficient r), and a parallel translation (by the vector b). This implies the theorem, because rotations, homotheties, and translations obviously possess all four of the properties asserted by the theorem. The least obvious of these facts is that homotheties preserve cross ratio, but this follows immediately from the fact that homothety in the plane of the complex variable is multiplication by a real number (which will cancel out in each of the fractions of the cross ratio). □

8.1.4. *Linear-fractional transformations.* A transformation of $\overline{\mathbb{C}}$ given on $\mathbb{C} \setminus \{-d/c\}$ by

$$\boxed{z \mapsto \frac{az + b}{cz + d}\,, \quad \text{where} \quad ac - bd \neq 0}, \tag{8.2}$$

which takes the point $-d/c$ to ∞ and ∞ to a/c is called *linear-fractional*.

The set of all linear-fractional transformations form a group, called the *Möbius group* and denoted by Möb.

Indeed, the fact that the composition of two linear-fractional transformations is a linear-fractional transformation can be shown as follows: substitute $(a_1z + b_1)/(c_1z + d_1)$ for z in the expression $(az + b)/(cz + d)$, which yields (after some manipulations)

$$\frac{(aa_1 + bc_1)z + (ab_1 + bd_1)}{(ca_1 + dc_1)z + (cb_1 + dd_1)}; \tag{8.3}$$

but this expression is of the same form as (8.2), so the composition is indeed linear-fractional.

The fact that the inverse of any linear-fractional transformation is a linear-fractional transformation is also easy to prove. To do that, it suffices to find values of a_1, b_1, c_1, d_1 (in terms of a, b, c, d) so that the expression (8.3) reduces to $(1 \cdot z + 0)/(0 \cdot z + 1)$; such values must satisfy the system of four linear equations in four unknowns

$$aa_1 + bc_1 = 1, \quad ab_1 + bd_1 = 0, \quad ca_1 + dc_1 = 0, \quad cb_1 + dd_1 = 1,$$

but this system obviously has a nonzero solution.

The following property of linear-fractional transformations gives an insight in the geometric meaning of this class of transformations and turns out to be extremely useful in constructing and analyzing them.

8.1.4. Lemma. *Let z_1, z_2, z_3 and w_1, w_2, w_3 be two triplets of distinct points of the Riemann sphere. Then there exists a unique linear-fractional transformation taking z_i to w_i, $i = 1, 2, 3$.*

8.1.5. Theorem. *Linear-fractional transformations take straight lines and circles to straight lines or circles, and preserve angles and cross ratios.*

Proof. As can be easily checked, the image of the point z under the linear-fractional transformation (8.1) may be rewritten as

$$\frac{az + b}{cz + d} = \frac{a}{c} + \frac{bc - ad}{c(cz + d)},$$

and therefore can be regarded as the composition

$$z \mapsto cz + d =: z_1 \mapsto cz_1 =: z_2 \mapsto 1/z_2 =: z_3 \mapsto (bc - ad)z_3 =: z_4 \mapsto$$

$$\mapsto \frac{a}{c} + z_4 = \frac{a}{c} + \frac{bc - ad}{c(cz + d)} = \frac{az + b}{cz + d}$$

of an affine transformation, a homothety, a transformation taking z to $1/z$, another homothety, and a parallel translation. About all these transformations, except $z \mapsto 1/z$, we know that they take straight lines to straight lines, circles to circles, and preserve angles and cross ratios.

Concerning the transformation $z \mapsto 1/z$, a straightforward if somewhat tedious calculation shows that it preserves cross ratios (one replaces z_i by $1/z_i$, $i = 1, 2, 3, 4$, and the obtained rather cumbersome fractions, after cancellations, reacquire the exact form of the original ratio). Further, since $\overline{1/z} = 1/\overline{z}$, the transformation $z \mapsto 1/z$ is the composition of a reflection, an

inversion, and another reflection. But we know that inversion takes straight lines or circles to straight lines or circles and preserves angles (see 7.1.1, items (i)–(iii)), which proves the theorem. □

8.1.6. *Two examples of linear-fractional transformations.* Linear-fractional transformations are the subject matter of an important chapter of the theory of a complex variable; in it, one studies what types of domains can be mapped into each other by linear-fractional transformations. We will not need the general theory of this study, but the two following examples of linear-fractional transformations will be very important for what follows.

Example 8.1. *The linear-fractional transformation*

$$\Omega : z \mapsto i \cdot \frac{1+z}{1-z}$$

maps the unit disk $\mathbb{D}^2 := \{z \in \mathbb{C} : |z|^2 \leq 1\}$ *to the upper half plane* $\mathbb{C}_+ := \{z \in \mathbb{C} : Im\, z > 0\}$. Indeed, it is easy to verify that the points $-1, i, 1$ are mapped to $0, -1, \infty$, respectively, which means (by Theorem 8.1.3) that the boundary circle of the disk $\mathbb{D}^2$ is mapped to the real axis. A simple computation shows that $|z| < 1$ implies that $\mathrm{Im}(\Omega(z)) > 0$, as required.

Example 8.2. *The linear-fractional transformations*

$$z \mapsto \frac{az+b}{cz+d} \quad \text{and} \quad z \mapsto \frac{a(-\overline{z})+b}{c(-\overline{z})+d}, \tag{8.3}$$

where $a, b, c, d \in \mathbb{R}$ *and* $ac - bd > 0$ *take the upper half plane to itself, the first of them preserving, the second, reversing the orientation of the half plane.*

For the first of these formulas, it is obvious that points of the real axis are taken to points of the real axis; further, if z, $\mathrm{Im} z > 0$, is any point in the upper half-plane, then

$$\mathrm{Im}\,\frac{az+b}{cz+d} = \mathrm{Im}\,\frac{(az+b)(c\overline{z}+d)}{|cz+d|^2} = \mathrm{Im}\,\frac{adz+bc\overline{z}}{|cz+d|^2} = \frac{(ad-bc)\mathrm{Im}\, z}{|cz+d|^2},$$

which is positive iff $ac - bd > 0$.

The second formula differs from the first by a transformation of the form $z \mapsto -\overline{z}$, which obviously takes the upper half plane to itself, but reverses the orientation.

The set of all linear-fractional transformations (8.3) *constitute a group under composition*, which we denote by $\mathbb{R}$Möb. Indeed, this follows from the

fact that the set of all linear-fractional transformations of the form (8.2) is a group and a composition of transformations taking the half plane to itself take the half plane to itself. The group $\mathbb{R}$Möb will be the transformation group of the half plane model.

8.2. The Poincaré half-plane model

The *Poincaré half-plane model* is the geometry consisting of the points $z \in \mathbb{C}$ such that $Im\, z > 0$, supplied with the transformation group $\mathbb{R}$Möb. In this geometry, *straight lines* are defined either as open half circles (in the upper half-plane) perpendicular to the line $Im\, z = 0$ (which is called the *absolute*) or as the open rays $\{z \in \mathbb{C} : Re z = x_0 \in \mathbb{R}, Im\, z > 0\}$.

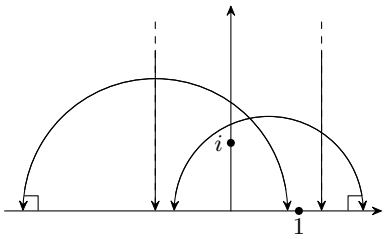


Figure 8.1. "Straight lines" in the half-plane model

8.3. Perpendiculars and parallels

The situation with perpendiculars and parallels in the half-plane model is quite similar to that for the disk model, except that the corresponding pictures look very different.

8.3.1. Theorem *Given a point P and a line l in the half plane model, there exists a unique perpendicular to l passing through P.*

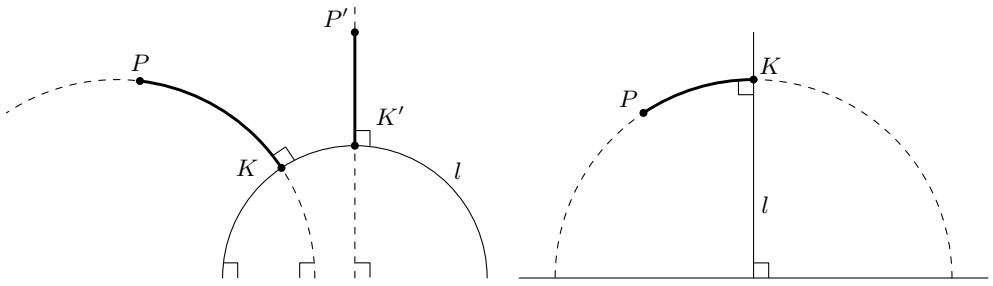


Figure 8.2. Perpendiculars in the half-plane model

Proof. There are two cases to consider (depending on whether l is a half-line or a half-circle), see Figure 8.2. In the first, the theorem follows from Exercise 8.2, in the second, the proof is obvious. □

8.3.2. Theorem. *Given a point P and a line l in the half plane model, there exist infinitely many lines passing through P and not intersecting l. All these nonintersecting lines lie between the two parallels to l from P.*

Proof. There are two cases to consider (depending on whether l is a half-line or a half-circle); see Figure 8.3. In the first, the theorem follows from the obvious fact there is exactly one half circle centered on the real axis passing through a point X on the real axis and a point P outside it, in the second, the proof is immediate. □

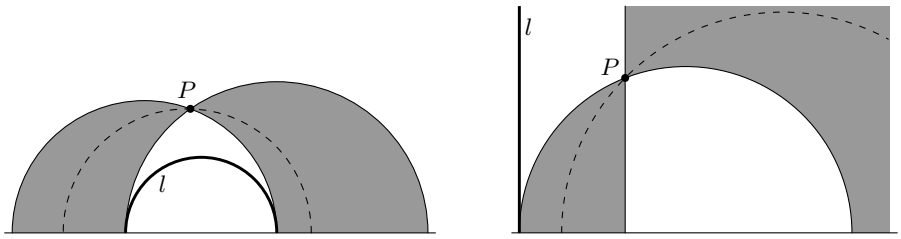


Figure 8.3. Parallels in the half-plane model

8.4. Isometries w.r.t. Möbius distance

Let us define the *Möbius distance* $\mu(A, B)$ between two points A, B of the upper half-plane by setting

$$\boxed{\mu(A, B) := |\ln(\langle A, B, X, Y\rangle)|,}$$

where X and Y are the intersection points of the line (AB) with the absolute if the points A, B have different real parts (note that $\langle A, B, X, Y\rangle \in \mathbb{R}$ because the four points lie on a circle, so that the log is well defined); if $Re(A) = Re(B) = x_0$, we set

$$\boxed{\mu(A, B) := |\ln(\langle A, B, \infty, X\rangle)|,}$$

where X is the point with coordinates $(x_0, 0)$.

8.4.1. Theorem. *The isometry group of the upper half plane with respect to the distance μ coincides with the group $\mathbb{R}M\ddot{o}b$ described in Example* 8.1.

The proof is a tedious verification that we omit. □

8.5. Problems

8.1. Prove that

(a) linear-fractional transformations preserve the cross-ratio of four points on the Riemann sphere $\overline{\mathbb{C}}$;

(b) a linear-fractional transformation is uniquely determined by three points and their images.

8.2. Let l be a straight line in the Euclidean plane, γ a circle with center O on l, P a point not on l and not on the perpendicular to l from O. Prove that there exists a unique circle passing through P, orthogonal to γ, and centered on l.

8.3. Let l be a straight line in the Euclidean plane, γ a circle with diameter AB on l, P a point not on l and not in γ. Prove that there exists a unique circle passing through P and A with center on l, and a unique circle passing through P and B with center on l.

8.4. Prove that all motions (i.e., orientation-preserving isometries) of the Poincaré disk model are of the form

$$z \mapsto \frac{az+b}{\overline{b}z+\overline{a}},$$

where a and b are complex numbers such that $|a|^2 = |b|^2 = 1$.

8.5. Show that there exists an isometry of the half-plane model that takes any flag to any other flag (a flag is a triple consisting of a line in the hyperbolic plane, one of the two half-planes that the line bounds, and a point on that line).

8.6*. Find a formula for the area of a triangle in hyperbolic geometry.

Chapter 9
THE CAYLEY–KLEIN MODEL

In this chapter, we study one more model of hyperbolic plane geometry – the Cayley–Klein model. Its set of points consists of all the points of the open disk (just as in the case of the Poincaré disk model) and its transformation group is isomorphic to $\mathcal{M}$ (the transformation group of the Poincaré model), but the action of $\mathcal{M}$ in the two models is not the same. As a result, the lines in the two models look very different: instead of arcs of circles as in the Poincaré model, in the second model lines are open chords of the disk.

Another essential difference between our study of the two models is in the approach to the definition of the Cayley–Klein model as a geometry (in the sense of Klein), i.e., the definition of its transformation group. This is done in a more traditional way: we will begin by defining the distance between points and then introduce the transformation group of the geometry as the isometry group of this distance, i.e., the group of all distance-preserving bijections of its set of points.

9.1. Isometry and the Cayley–Klein model

9.1.1. *The distance function.* Let $\mathbb{H}^2$ be the interior of the unit disk on the Euclidean plane and let A and B be points of $\mathbb{H}^2$. Suppose the (Euclidean) line AB intersects the boundary of the disk $\overline{\mathbb{H}}^2$ at the points X and Y, the points Y, A, B, X appearing on the line AB in that order (see Fig.9.1).

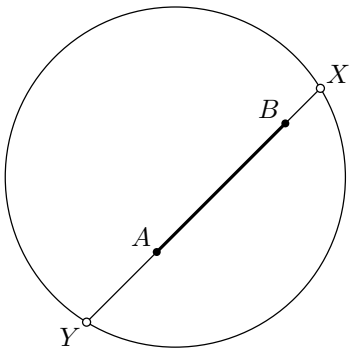


Figure 9.1. Line in the Cayley–Klein model

Then the *distance* d between the points A and B is defined as

$$\boxed{d(A,B) := \frac{1}{2}\left|\log \frac{|AX|}{|BX|}\cdot\frac{|AY|}{|BY|}\right|}. \tag{9.1}$$

The coefficient $1/2$ in the right-hand side of (9.1) can be replaced by any other positive real number c – all such distances define the same geometry (up to isomorphism, but not up to isometry). The reason for this strange choice ($c = 1/2$ rather than the more natural $c = 1$) is that the coefficient $c = 1/2$ leads to more elegant formulas than $c = 1$ and gives the same metric as in the Poincaré model.

Note that if the points Y, A, B, X are ordered on the line AB as shown in the figure (and $A \neq B$), then the expression under the logarithm sign is greater than 1 and therefore the distance between A and B is positive. Note further that if we introduce coordinates on the line AB, placing the origin "to the left" of Y and assigning the real numbers y, a, b, x to the points Y, A, B, X, respectively, then the expression under the logarithm sign can rewritten as the following cross ratio

$$\frac{x-a}{x-b} : \frac{y-a}{y-b} = \langle a, b, x, y\rangle.$$

This cross ratio looks very similar to the one we used to define the distance in the half plane model, but it should be stressed that here we are dealing with real numbers rather than complex ones.

9.1.2. *Properties of the distance function. The distance function d given by* (9.1) *defines a metric on the open disk $\mathbb{H}^2$, i.e.,*

(i) $d(A,B) \geq 0$, *and* $d(A,B) = 0$ *if and only if* $A = B$.

(ii) $d(A,B) = d(B,A)$.

(iii) $d(A,B) + d(B,C) \geq d(A,C)$.

Proof. Item (i) obviously holds: the distance $d(A,B)$ between distinct points A and B is positive (as we have shown above), while if $A = B$, then the denominators in (9.1) cancel, leaving us with log(1)=0.

Item (ii) follows from the obvious formula

$$\frac{x-a}{x-b} : \frac{y-a}{y-b} = \Big(\frac{x-b}{x-a} : \frac{y-b}{y-a}\Big)^{-1}.$$

Finally, item (iii) can be proved by using projective transformations. Since we won't be using (iii) in what follows, we postpone its proof to Chapter 12 (see Exercise 12.13).

9.1.3. *Definition of the Cayley–Klein model.* As explained above, we will define the geometry (in the sense of definition 1.4.1) of the Cayley–Klein model by taking for its transformation group the isometry group of the distance d, i.e., the group of all distance-preserving bijections of $\mathbb{H}^2$, which we denote by $\mathcal{N}$. (We will prove later that $\mathcal{N}$ is isomorphic to $\mathcal{M}$, the transformation group of the Poincaré disk model, but this fact does not concern us now.)

Thus we define the *Cayley–Klein model* of the hyperbolic plane as the geometry $\big(\mathbb{H}^2 : \mathcal{N}\big)$, where $\mathcal{N}$ is the isometry group of the open unit disk $\mathbb{H}^2$ with respect to the distance (9.1).

9.1.4. *Lines and points in the Cayley–Klein model.* The *points* of the Cayley–Klein model, as explained above, are simply the points of the open unit disk $\mathbb{H}^2$ in $\mathbb{R}^2$. The boundary of the disk is traditionally called the *absolute*, and its points do *not* belong to our geometry.

The *lines* of our geometry are defined as the chords of the absolute (without their endpoints). This definition immediately implies the fundamental facts that *one and only one line passes through any two distinct points* and that *two noncoinciding lines either don't intersect or have exactly one common point*.

In the two following sections, just as in the corresponding sections in the previous two chapters, we shall derive the basic facts of hyperbolic geometry in the case of the model under consideration.

9.2. Parallels in the Cayley–Klein model.

The situation with parallelism in this model is similar to that in the Poincaré disk model, except that the picture looks slightly different (rectilinear chords instead of arcs of circles).

9.2.1. *Definitions.* Given a line $l = AB$ and a point P not on this line, it is easy to describe the lines that pass through P and do not intersect l. Indeed, denoting by l and m the lines passing through P and through the intersection points X, Y of line l with the absolute, we see that any line passing through P and lying between k and m does not intersect line l; these lines are called *nonintersecting* lines w.r.t. AB, while the lines k and m are the *parallels* to AB passing through P (see Figure 9.2).

More generally, two lines (i.e., open chords of the disk) are *parallel* if they have no common points in $\mathbb{H}^2$ and one common point on the absolute; if two lines (chords) have no common points at all (in the closed disk $\overline{\mathbb{H}^2}$), then they are called *nonintersecting*.

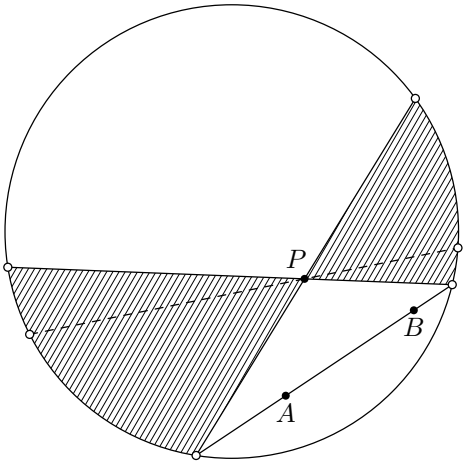


Figure 9.2. Parallels and nonintersecting lines

We have shown that *there are infinitely many lines passing through a given point P not intersecting a given line $l = AB$ if $P \notin l$; these lines are all located between the two parallels to l passing through P.*

9.2.2. *Remark.* Note that the set of of all lines passing through a fixed point of the absolute

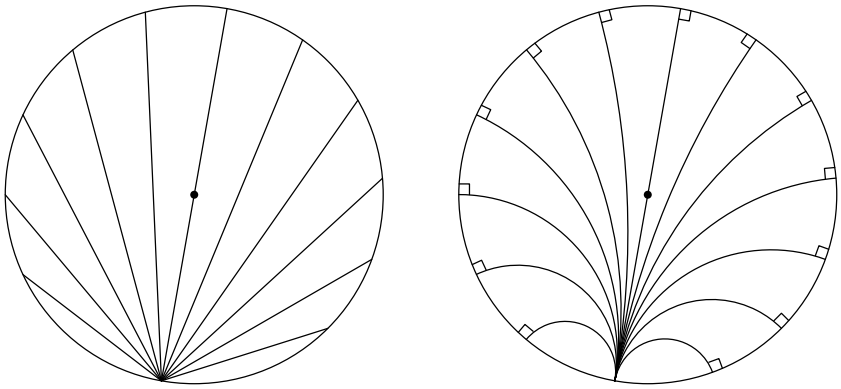


Figure 9.3. Parallels filling the hyperbolic plane

point of the absolute fills the entire hyperbolic plane $\mathbb{H}^2$ (see Figure 9.3, where both disk models are pictured).

This means that, by using the metric on each of these lines, we can try to define the notion of "parallel translation" and therefore that of a "free vector" of sorts in hyperbolic geometry. This might lead one to think that one can associate a linear space with our geometry. Unfortunately, this is not the case (see the discussion in 9.4.1 and in Exercise 9.7).

9.3. Perpendiculars in the Cayley–Klein model.

9.3.1. *What they look like.* Unlike perpendiculars in the Poincaré disk model, perpendicular lines in the Cayley–Klein model do not form right angles in the Euclidean sense. An exactly constructed example is shown in Fig.9.3 (the nontrivial geometric construction by means of which this "hyperbolically perpendicular straight line" was drawn does not appear on the figure, and will be discussed in the next chapter, in 10.1.4).

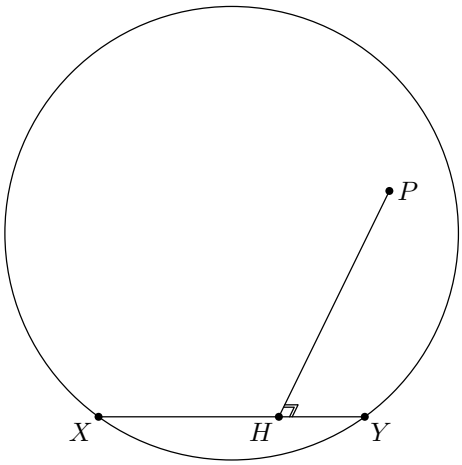


Figure 9.4. Strange looking perpendicular

9.3.2. *Definitions.* Before discussing perpendicularity, we must *define* what perpendicular lines are. To do that, we first define a *reflection with respect to a given line* as the nonidentical isometry of $\mathbb{H}^2$ that takes each point of the given line to itself. Now we can define two lines as *perpendicular* if the reflection with respect to one of them takes the other line to itself. It is true that *there exists one and only one perpendicular to a given line passing*

through a given point, but the proof of this fact directly in the Cayley–Klein model is quite difficult and is omitted.

9.3.3. *Remark.* It should be stressed that in our model the "hyperbolic measure" of angles is in general *not* equal to their Euclidean measure. In particular, triangles in the Cayley–Klein model, which look like rectilinear Euclidean triangles, have angle sums *less* than π (although visually this is does not seem to be the case).

9.4. The hyperbolic line and relativity

In this section, we digress about the distance function on hyperbolic straight lines and point out a remarkable relationship between the composition of shifts on such a line and the additivity of velocities in the Special Relativity Theory of Einstein. But we begin with a general remark concerning vectors in hyperbolic geometry.

9.4.1. *Remark about free vectors.* The notion of free vector in Euclidean geometry, defined as an equivalence class of equal fixed vectors, allows to associate to the Euclidean plane a two-dimensional real vector space whose elements are precisely the free vectors of the Euclidean plane. Any free vector also defines parallel shifts of the entire plane in a natural way. All this is possible because at each point of the Euclidean plane there is one and only one (fixed) vector pointing in the same direction and having the same length as a given (fixed) vector. On the hyperbolic plane supplied with a metric, we can say when two vectors have the same length, but the expression "point in the same direction" is meaningless (compare with Remark 9.2.2), so that *there is no well-defined notion of parallel shift*. However, the notion of parallel shift *along a fixed hyperbolic straight line* makes sense, and we discuss it in the next subsection.

9.4.2. *Adding shifts and velocities.* Let us distinguish some hyperbolic straight line in the Cayley–Klein model (i.e., an open chord of the open disk $\mathbb{H}^2$) and parametrize it by an appropriate Euclidean parameter x so that it is isometric to the open interval $(-1, 1)$. Let v be a real number of absolute value less than 1. Consider the map

$$T_v : [-1, 1] \to [-1, 1], \quad x \mapsto \frac{x + v}{xv + 1}.$$

It is easy to prove that T_v is a bijection of the closed interval $[-1, 1]$ to itself leaving its endpoints in place and its restriction to the open interval $(-1, 1)$

is an isometry with respect to the hyperbolic distance (for the details, see Exercise 9.9). This isometry can therefore be regarded as *the parallel shift along the given hyperbolic line by the vector* v.

Let us calculate the composition of two parallel shifts by the vectors v_1 and v_2:

$$x \mapsto \frac{x+v_1}{xv_1+1} \mapsto \Big(\frac{x+v_2}{xv_2+1}+v_1\Big)\Big/\Big(\frac{x+v_2}{xv_2+1}v_1+1\Big) =$$
$$= \Big(x+\frac{v_1+v_2}{1+v_1v_2}\Big)\Big/\Big(x\frac{v_1+v_2}{1+v_1v_2}+1\Big);$$

we see that the composition $T_{v_2}\circ T_{v_1}$ is exactly the parallel shift T_v, where v is defined by the formula

$$v := \frac{v_1+v_2}{1+v_1v_2}. \tag{9.1}$$

Thus we have proved that *the composition of two parallel shifts by vectors* v_1 *and* v_2 *is a parallel shift by the vector* v *given by formula* (9.1).

The reader will surely have noticed that this formula is the analog of the famous Einstein formula for the addition of velocities:

$$v := \frac{v_1+v_2}{c+v_1v_2},$$

where c is the speed of light. The two formulas differ only in the choice of the scale of velocity, in our hyperbolic scale the "speed of light" is set equal to 1. Note that in both situations, if the "velocity vectors" v_1 and v_2 are very small as compared to the constant c (or 1 in our case), then v is approximately equal to v_1+v_2.

The above observation is an argument in favor of our universe being hyperbolic rather than Euclidean. (Actually, most physicists believe it is neither.)

9.5. Problems

9.1. Prove that for three points A, B, C on one line, where B is between A and C, one has $d(A,B)+d(B,C)=d(A,C)$.

9.2. Prove that the equality $d(A,B)+d(B,C)=d(A,C)$ implies that the points A, B, C lie on one line and B is between A and C.

9.3. Prove that the reflection in a line in the Cayley-Klein model is an involution.

9.4. Show that the notion of perpendicular lines in the Cayley–Klein model (as introduced in 9.3.2) is well defined (i.e., does not depend on the order of the two lines).

9.5. Prove that the four angles formed at the intersection point of two perpendiculars are congruent.

9.6*. Prove that the sum of angles of a triangle in the Cayley–Klein model is less than π directly from the definitions pertaining to the model.

9.7. Having defined the notion of free vector in hyperbolic geometry as suggested in 9.2.2, try to define the sum of two vectors and investigate the possibility of associating a two-dimensional vector space with hyperbolic plane geometry.

9.8. Construct a triangle in the Cayley–Klein model with angle sum less than a given positive ε.

9.9. Prove that the parallel shift T_v defined in 9.4.2 does take $(-1, 1)$ to itself and find the appropriate hyperbolic distance for which it is an isometry.

Chapter 10

HYPERBOLIC PLANE TRIGONOMETRY AND ABSOLUTE CONSTANTS

We begin this chapter by showing that the three models of the hyperbolic plane are, in fact, isomorphic geometries. In continuing and concluding our study of hyperbolic plane geometry, we will then feel free to use whichever model is more convenient in the given context. This study includes the main formulas of hyperbolic trigonometry, which we obtain after having recalled the definitions of the hyperbolic functions, usually studied in complex analysis. In conclusion of the chapter, we learn that in hyperbolic geometry, unlike Euclidean geometry, there are inherent absolute constants.

10.1. Isomorphism between the two disk models

As we mentioned in the previous chapter, the Cayley–Klein model and the Poincaré disk model are isomorphic. This means that that there is a bijection between their sets of points and an isomorphism of their transformation groups which are compatible in the sense specified in 1.4.4. To prove this, we will need a classical construction from Euclidean space geometry.

10.1.1. *Stereographic projection.* Let $\mathbb{S}^2$ be the unit sphere, let Π be the equatorial plane of the sphere, and N be its North Pole. The *stereographic projection* $\sigma : \Pi \to \mathbb{S}^2$ is the map that takes each point $M \in \mathbb{S}^2 \setminus N$ to the intersection point M' of the ray NM with Π.

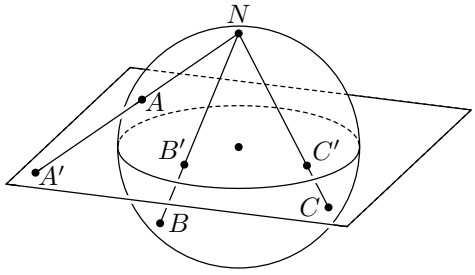


Figure 10.1. Stereographic projection

Obviously, σ is a bijection of $\mathbb{S}^2 \setminus \{N\}$ onto Π. It is also not hard to prove that *stereographic projection is conformal* (see Exercise 10.1).

10.1.2. *Bijection between the sets of points of the two disk models.* We regard the intersection of the open unit ball with the equatorial plane Π as the set $\mathbb{H}^2$ of points of both disk models. In order to prove that the two models are isomorphic, we begin by establishing a bijection β between their point sets. This bijection is *not the identity map*, and can be described as follows.

Let A be an arbitrary point of $\mathbb{H}^2$ and let XY be the chord (of the absolute) perpendicular to the radius OA (Fig.10.2). Consider the vertical plane containing XY; it intersects the unit sphere along a circle. Denote by A_1 the intersection of the downward vertical ray passing through A with this circle. Now join the points A_1 and N and denote by A' the intersection of A_1N and the equatorial plane. The correspondence $A \mapsto A'$ defines a map from $\mathbb{H}^2$ to $\mathbb{H}^2$ that we denote by β.

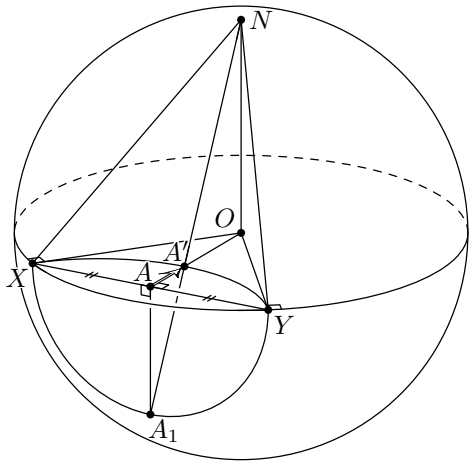


Figure 10.2. Bijection between the two disk models

It is not hard to prove that *the map β is a bijection of $\mathbb{H}^2$ onto itself* (for the details, see Exercise 10.2).

10.1.3. *Isomorphism between the transformation groups.* The next step in the proof of the fact that the two disk models are isomorphic geometries is the construction of an isomorphism between their transformation groups $\mathcal{N}$ and $\mathcal{M}$ that would be compatible with β. But that construction is in a sense automatic, because, as we shall see, the compatibility condition actually prescribes the choice of isomorphism.

Our aim is to construct an isomorphism $\varphi : \mathcal{N} \to \mathcal{M}$, where $\mathcal{N}$ and $\mathcal{M}$ are the transformation groups of the Cayley–Klein and the Poincaré disk models, respectively. Let $g \in \mathcal{N}$ be an arbitrary element and A be an arbitrary point of the Poincaré disk. We define the element $\varphi(g)$ by setting

$$\big(\varphi(g)\big)(A) := \beta\big(g\big((\beta^{-1}(A)\big)\big),$$

where β is the bijection defined in the previous subsection. This formula says that in order to obtain the image $B := \big(\varphi(g)\big)(A)$ under $\varphi(g)$ of an arbitrary point A, we perform the only possible natural actions: pull back the point A from the Poincaré disk model to the Cayley–Klein disk via β^{-1}, obtaining $A' := \beta^{-1}(A)$, act on A' by g, and return the obtained point $g\big((\beta^{-1}(A)\big)$ to the Poincaré disk via β (look at Figure 10.3).

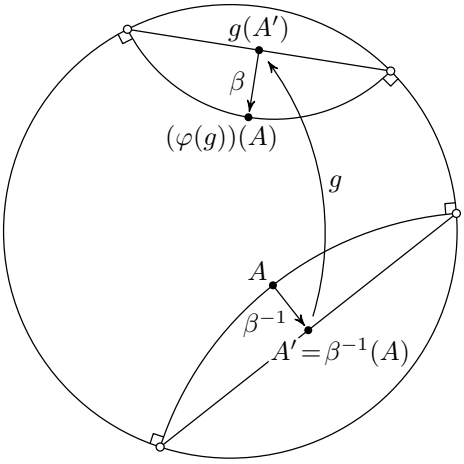


Figure 10.3. Isomorphism of the two disk models

The fact that φ) is a group homomorphism is obvious by construction, the fact that it is bijective is also easy to prove (see Exercise 10.3), while the fact that the pair $(\beta, \varphi$ is an isomorphism of geometries is also immediate from the construction. We have proved the following theorem.

10.1.4. Theorem. *The map β from* 10.1.3 *defines an isomorphism of the geometry $(\mathbb{H}^2 : \mathcal{N})$ (the Cayley–Klein model) and the geometry $(\mathbb{H}^2 : \mathcal{M})$ (the Poincaré disk model) if we define the corresponding isomorphism (which we denote by φ) of the groups $\mathcal{N}$ and $\mathcal{M}$ by setting*

$$\big(\varphi(g)\big)(A) := \beta\big(g\big((\beta^{-1}(A)\big)\big),$$

where A is any point of the Poincaré disk and $g \in \mathcal{N}$.

10.1.4. *Construction of perpendiculars in the Cayley–Klein model.* The fact that we have a concrete isomorphism between the two disk models can be used to construct the "strange looking perpendiculars" (look at Fig. 9.4 again) in the Cayley–Klein model model. To do that, we use the bijection β from 10.1.2 to pass to the Poincaré disk model, where we know how to construct perpendiculars (see Theorem 7.4.1) and, having performed that construction, we return to the Cayley–Klein model via β^{-1}, obtaining the required perpendiculars.

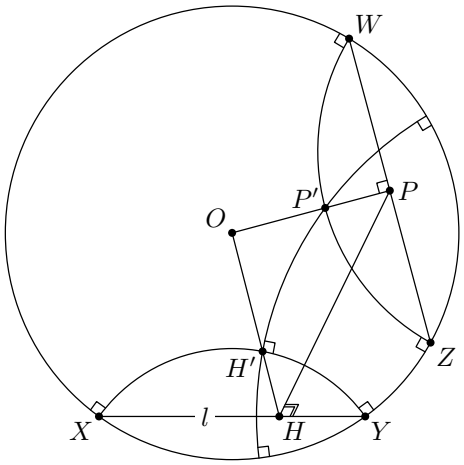


Figure 10.4. Constructing perpendiculars in the Cayley–Klein model

In more detail, the construction is as follows (Fig.10.4). We are given a line $l = XY$ and a point P in the Cayley–Klein model $\mathbb{H}^2$. First we construct the chord WZ containing P and perpendicular to the radius OP. Next, we construct the two arcs of circles perpendicular to the absolute and passing through the points X, Y and W, Z and denote by P' the intersection point of the arc subtending WZ with the radius OP. Note that the two arcs are the images of the Cayley–Klein lines XY and WZ under the bijection β (see 10.1.2) and are therefore lines in the Poincaré disk model.

From the point P', we draw the arc orthogonal to the absolute and orthogonal to the arc l' subtending XY (see 7.4.1) and denote by H' the intersection point of these two arcs. Note that H' is the foot of the perpendicular lowered from P' to l' in the sense of the Poincaré disk model. Now if we construct

the ray OH', its intersection point H with the line l is the foot of the required perpendicular lowered from P to l, because the map β^{-1} transforms the Poincaré perpendicular $P'H'$ to the Cayley–Klein perpendicular PH.

10.2. Isomorphism between the two Poincaré models

In this section we show that the Poincaré disk model (Chapter 7) is isomorphic to the half-plane model studied in Chapter 8. To do that, we will need the linear-fractional transformation Ω defined (in Example 8.1) by the formula

$$\Omega : z \mapsto i \cdot \frac{1+z}{1-z};$$

Ω maps the unit disk $\mathbb{D}^2 := \{z \in \mathbb{C} : |z|^2 \le 1\}$ to the upper half plane $\mathbb{C}_+ := \{z \in \mathbb{C} : Im\, z > 0\}$. The transformation Ω, together with the compatibility (equivariance) condition determines the isomorphism between the two geometries. More precisely, we have the following result:

10.2.1. Theorem. *The map Ω from Example 8.1 defines an isomorphism of the geometry $(\mathbb{H}^2 : \mathcal{M})$ (the Poincaré disk model from Chapter 7) and the geometry $(\mathbb{C}_+ : \mathbb{R}M\ddot{o}b)$ (the Poincaré half-plane model) if we define the corresponding isomorphism (which we denote by Δ) of the groups $\mathbb{R}$Möb and $\mathcal{M}$ by setting*

$$M \ni g \mapsto \Omega \circ g \circ \Omega^{-1} \in \mathbb{R}\text{Mob}.$$

Proof. The map Ω is one-to-one because it has the obvious inverse given by the rule $w \mapsto (i-w)/(i+w)$. The isomorphism Δ is compatible with the group actions by definition. $\square$

Now let us define the *Lobachevsky distance* λ between two points A, B of the open disk $\mathbb{H}^2$ (in the framework of the Poincaré disk model) by setting

$$\boxed{\lambda(A,B) := |\ln(\langle A, B, X, Y\rangle)|,}$$

where X and Y are the intersection points of the line (AB) with the absolute.

Now Theorems 8.1.3, 8.1.5, and 10.2.1 immediately imply the following result:

10.2.2. Corollary. *The group of isometric transformations of the disk with respect to the distance λ coincides with the group $\mathcal{M}$ generated by all reflections in the "straight lines" of the disk model.*

Since isomorphism of geometries is a transitive relation, we have the following

10.2.3. Corollary. *The three models of hyperbolic geometry, namely the Poincaré disk and half plane models and the Cayley–Klein model, are isomorphic as geometries in the sense of Klein.*

10.3. Hyperbolic functions

The complex exponent e^z, $z \in \mathbb{C}$, is related to the ordinary trigonometric functions by the beautiful *Euler formula*:

$$e^{i\varphi} \;=\; \cos\varphi + i\sin\varphi,$$

whose proof is obvious if we consider the unit circle centered at the origin of the plane $\mathbb{C}$. The real exponent e^x, $x \in \mathbb{R}$, is related to the "trigonometric functions" of hyperbolic geometry, known as the *hyperbolic functions* $\operatorname{sh}, \operatorname{ch}, \operatorname{th}, \operatorname{cth}$ (*hyperbolic sine, cosine, tangent, cotangent*, respectively) and defined by the formulas

$$\operatorname{sh} x := \frac{e^x - e^{-x}}{2} \qquad \operatorname{ch} x := \frac{e^x + e^{-x}}{2}$$

$$\operatorname{th} x := \frac{e^x - e^{-x}}{e^x + e^{-x}} \qquad \operatorname{cth} x := \frac{e^x + e^{-x}}{e^x - e^{-x}}$$

These functions satisfy formulas similar to the main formulas for ordinary trigonometric functions. Here are some examples:

$$\operatorname{ch}^2 x - \operatorname{sh}^2 x = 1, \quad \operatorname{th} x \operatorname{cth} x = 1, \quad \operatorname{sh}(x \pm y) = \operatorname{sh} x \operatorname{ch} y \pm \operatorname{ch} x \operatorname{sh} y,$$

$$\operatorname{sh} 2x = 2\operatorname{sh} x \operatorname{ch} x, \quad \operatorname{ch} 2x = \operatorname{sh}^2 x + \operatorname{ch}^2 x, \quad \operatorname{ch}(x \pm y) = \operatorname{ch} x \operatorname{ch} y \pm \operatorname{sh} x \operatorname{sh} y.$$

The proofs are obtained by plugging in the definitions in the formulas and performing simple calculations.

10.4. Trigonometry on the hyperbolic plane

Because of Corollary 10.2.3, the elementary trigonometric formulas for hyperbolic triangles are exactly the same for the half-plane and the disk model. Their proof is quite straightforward (perhaps a little simpler in the case of the half-plane) and are relegated to the exercises. We state them in the form of theorems. Below ABC is a triangle, α, β, γ are the angles

opposite to A, B, C, respectively, and a, b, c the sides opposite to A, B, C, respectively.

10.4.1. Theorem. (Hyperbolic sine theorem)

$$\boxed{\frac{\operatorname{sh} a}{\sin \alpha} = \frac{\operatorname{sh} b}{\sin \beta} = \frac{\operatorname{sh} c}{\sin \gamma}.}$$

10.4.2. Theorem. (Hyperbolic cosine theorem)

$$\boxed{\operatorname{ch} a = \operatorname{ch} b \operatorname{ch} c - \operatorname{sh} c \operatorname{sh} b \cos \alpha.}$$

10.5. The angle of parallelism and the Schweikart constant

10.5.1. Let (AB) be a line in hyperbolic geometry (we can use either one of the two models here) and C be a point not on (AB); let X and Y be the intersection points of the line (AB) with the absolute, so that the rays $[CX)$ and $[CY)$ are the parallels to (AB) passing through C; let $[CH]$, $H \in (AB)$, be the perpendicular lowered from C to (AB); let $d := \lambda(C, H)$ be the Lobachevsky distance between C and H; finally, let α be the measure of the angle XCH (or, which is the same, of YCH).

Then it is not difficult to prove that α depends only on d (see Exercise 10.11); α is called the *angle of parallelism*.

10.5.2. Theorem. *The angle of parallelism α is given by the formula:*

$$\boxed{\operatorname{tgh} d \;=\; \cos \alpha}.$$

For the proof, see Exercise 10.9.

This formula shows, in particular, that when d is very small, the angle of parallelism is close to $\pi/2$, while for large values of d, α becomes very small.

10.5.3. Now let O be the center of the disk model and let $[OA)$ and $[OB)$ be perpendicular rays issuing from O; let X and Y be the intersection points of the rays $[OA)$ and $[OB)$ with the absolute; let (CD) be the line intersecting the absolute at X and Y; let $[OH]$, $H \in (CD)$, be the perpendicular lowered from O to (CD); let $\sigma := \lambda(O, H)$ be the hyperbolic distance between O and H.

The number σ is called the *Schweikart constant*; it is an absolute constant of the hyperbolic plane. If we think of hyperbolic geometry as a model of

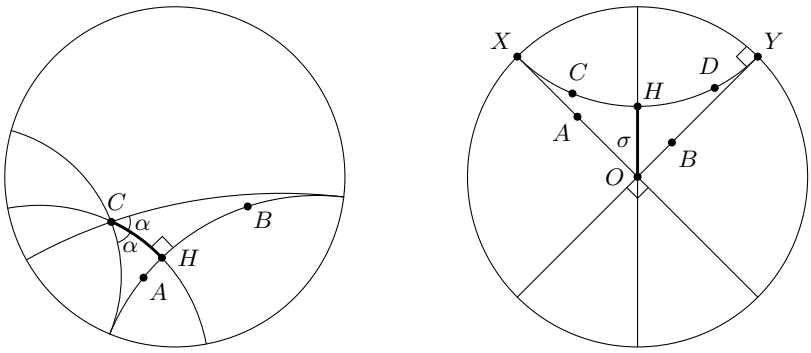


Figure 10.5. The angle of parallelism and the Schweikart constant

physical reality, then we must conclude that there is an absolute unit of length in our universe (no such unit appears in the Euclidean model of space).

10.5.3. Another absolute constant of hyperbolic geometry comes from the measure of a standard area, namely that of a special infinite "triangle". To construct this triangle, consider three rays issuing from the center (actually, any other point will do) of the disk model and forming angles of $2\pi/3$. Denote by X, Y, Z their intersection points with the absolute, and consider the lines XY, , YZ , ZX. They form an "infinite equilateral triangle" with all three angles equal to zero. Then its area can be computed by the formula for the area of a triangle in hyperbolic geometry

$$S = \pi - \alpha - \beta - \gamma \Longrightarrow S = \pi$$

(see Chapter 8 and Exercise 8.6).

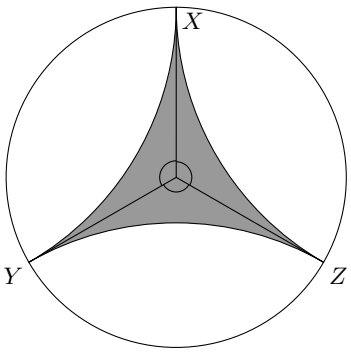


Figure 10.6. Infinite triangle

The above argument was not very rigorous, since the formula used is applicable only to finite triangles, but it can be made rigorous by approximating triangle XYZ by finite triangles and passing to the limit.

Thus we have obtained a third absolute constant, namely π, the area of the figure bounded by three lines joining three points of the absolute.

10.5.4. Remark. We noted above (see Section 9.4) that the formula for adding vectors on the hyperbolic line is very similar to Einstein's formula for adding the velocities of inertial frames. In this section, we have obtained three absolute constants – this is another trait of hyperbolic geometry that is similar to the properties of Einstein's theory of the physical world, in which absolute constants (e.g. the speed of light) appear. In this connection, one should not be misled by the word "relativity": Einstein's theory doesn't say that "everything is relative", on the contrary, it supplies us with physically meaningful absolute constants, something that a Euclidean model of the universe cannot do. On the other hand, a physical model entirely based on hyperbolic space geometry and an independent "time axis" is not viable either: our universe is more complicated than that, time and space are not independent, according to Einstein, they "mingle together" in a certain sense.

10.5. Problems

10.1. Prove that stereographic projection is conformal.

10.2. Prove that the map β constructed in 10.1.2 is bijective and show that any chord of $\mathbb{H}^2$ (i.e., any line in the Cayley–Klein model) is taken by β to the arc of the circle passing through X and Y and orthogonal to the absolute (i.e., to a line in the Poincaré disk model).

10.3. Prove the main relations between the hyperbolic functions indicated in Section 10.3.

10.4. Prove the hyperbolic sine theorem.

10.5. Prove the hyperbolic cosine theorem.

10.6. Prove that two triangles with equal sides are congruent in hyperbolic geometry.

10.7. Prove that in hyperbolic geometry two triangles having an equal angle and equal sides forming this angle are congruent.

10.8. Show that homothety is not conformal in hyperbolic geometry.

10.9. (a) Prove the formula for the angle of parallelism α for a point A and a line l:

$$\tanh(d) = \cos(\alpha)$$

where d is the distance from A to l (thereby showing that the angle of parallelism depends only on the distance from the point to the line).

(b) Prove that the previous formula is equivalent to the following one (obtained independently by Bolyai and Lobachevsky):

$$\tan\frac{\alpha}{2} = e^{-d}$$

10.10. Prove that in a triangle with right angle γ the sides a, b, c and their opposite angles $\alpha, \beta, \gamma = \pi/2$ satisfy the following relations:

$$\operatorname{sh} a = \operatorname{sh} c \sin\alpha;\ \operatorname{tgh} b = \operatorname{tgh} c \cos\alpha;\ \operatorname{ctg}\alpha \operatorname{ctg}\beta = \operatorname{ch} c;\ \cos\alpha = \operatorname{ch} a \sin\beta.$$

What do these relations tend to as a, b, c become very small?

10.11. Prove that the sides a, b, c and opposite angles α, β, γ of any triangle on the hyperbolic plane satisfy the following relations:

$$\text{(a)}\quad \operatorname{ch} a \sin\beta = \operatorname{ch} b \sin\alpha \cos\beta + \cos\alpha \sin\gamma;$$

$$\text{(b)}\quad \operatorname{ch} a = \frac{\cos\alpha + \cos\beta\cos\gamma}{\sin\beta\sin\gamma}.$$

10.12. Prove that if the corresponding angles of two triangles are equal, then the triangles are congruent.

10.13. Prove that all the points of the (Euclidean) straight line $y = kx$ that lie in the upper half plane $y > 0$ are equidistant from the (hyperbolic) straight line Oy.

10.14. (a) Prove that any hyperbolic circle contained in any one of the Poincaré models of hyperbolic geometry is actually a Euclidean circle.

(b) For the Poincaré upper half plane model, find the Euclidean center and radius of the hyperbolic circle of radius r centered at the point (a, b).

(c) For the Poincaré model in the unit disk D, find the relationship between the radii of the Euclidean and the hyperbolic circles centered at the center of D.

10.15. Prove the triangle inequality for the distance in the Poincaré half plane model.

10.16. Prove that the three (a) bissectors (b) medians (c) altitudes of any hyperbolic triangle intersect at one point.

10.17. (*The hyperbolic Menelaus Theorem.*) The line l intersects the lines BC, CA, AB (containing the sides) of triangle ABC at the points A_1, B_1, C_1 respectively; then

$$\frac{\operatorname{sh} AC_1}{\operatorname{sh} C_1B}\frac{\operatorname{sh} BA_1}{\operatorname{sh} A_1C}\frac{\operatorname{sh} CB_1}{\operatorname{sh} B_1A} = 1.$$

10.18. (*The hyperbolic Ceva Theorem.*) The points A_1, B_1, C_1 are chosen on the sides BC, CA, AB of triangle ABC. Prove that the segments AA_1, BB_1, CC_1 intersect at one point if and only if one of the following two equivalent conditions hold:

$$\frac{\sin ACC_1}{\sin C_1CB}\cdot\frac{\sin BAA_1}{\sin A_1AC}\cdot\frac{\sin CBB_1}{\sin B_1BA} = 1, \quad \frac{\operatorname{sh} AC_1}{\operatorname{sh} C_1B}\cdot\frac{\operatorname{sh} BA_1}{\operatorname{sh} A_1C}\cdot\frac{\operatorname{sh} CB_1}{\operatorname{sh} B_1A} = 1.$$

Chapter 11
HISTORY OF NON-EUCLIDEAN GEOMETRY

In this chapter, we will retrace the history of the creation of non-Euclidean geometry by Gauss, Lobachevsky, and Bolyai (and their predecessors and followers) and discuss the traditional axiomatic approach to the foundations of geometry. The story begins with Euclid's *Elements*, the brilliant first attempt to construct mathematics as a deductive science (see [8]).

11.1. Euclid's fifth postulate

The Ancient Greeks realized that, in a deductive science, in order to deduce (prove) facts from other facts by logical reasoning, it is necessary to start from some facts which are not proved. Euclid called these facts *postulates* (we call them *axioms*) and explicitly formulated five of them. He also used several other axioms implicitly (without formulating them). Apparently, Euclid (and other Greek mathematicians) thought that the postulates should be self-evident (simple and so obvious that no doubt about their truth could arise).

Euclid's last axiom, the *fifth postulate*, however, is not simple and not obvious. Its modern equivalent it can be stated as follows:

(V+) *For any straight line and any point not on this line there is a unique parallel to this line passing through the given point.*

Here by a *parallel* to a given line one means a straight line that has no common points with the given line. In Euclid's formulation, the statement was more complicated and less obvious.

(V) *If a straight line falling on two straight lines makes the sum of the interior angles on one side less than two right angles, then the two straight lines, if extended indefinitely, meet on that side on which are the angles with sum less than two right angles.*

Presumably, Greek mathematicians (perhaps Euclid himself) tried to deduce the fifth postulate from the other axioms. In any case, in Euclid's *Elements*, the application of the fifth postulate is postponed as much as possible: it occurs for the first time in the proof of Proposition 27 of Book 1 (there are 48 propositions, i.e., theorems in our terminology, in that book). The interested reader may want to look at the postulates and theorems in Book 1 of Euclid's *Elements*: they appear in Appendix II of the present book.

After Euclid, for more than two thousand years, many scientists tried to prove the fifth postulate, and many "succeeded", usually by proving statements equivalent to (V) by means of arguments based on additional axioms which were not explicitly formulated.

11.2. Statements equivalent to the fifth postulate

We have already mentioned one such statement, namely (V+). Here are some more (in square brackets [ ], we indicate the mathematician who used this approach to "prove" the fifth postulate).

(1) *The sum of the three angles of any triangle is equal to* π (*to two right angles,* in Euclid's terminology). [This statement appears in Euclid's *Elements* as Proposition 32, and was proved by using the fifth postulate; Legendre gave a "proof" in 1805 without the fifth postulate.]

(2) *A line intersecting one of two parallel lines intersects the other.* [Proclus, 5th century]

(3) *Similar but not congruent triangles exist.* [John Wallis, 1663]

(4) *The fourth angle of a quadrilateral with three right angles is also a right angle.* [Nasiraddin, 13th century, Saccheri, 1679, Lambert, 1776] . Such a quadrilateral was later called a *Saccheri quadrilateral.*

Trying to prove the fifth postulate, most mathematicians (including those mentioned above) argued by contradiction. As a rule, they considered two cases, assuming that the sum of angles of a triangle is (a) more than π or (b) less than π (equivalently, that the fourth angle of the Saccheri quadrilateral is more (less) than $\pi/2$, or that there are no parallels, respectively more than one parallel, through a given point to a given line). In the first case, it is possible to correctly obtain a contradiction using the Euclidean axioms. In the second case, a contradiction does not follow, but the desire to prove the fifth postulate was so strong that the mathematicians working on the problem usually produced what they claimed to be a proof, but which was actually flawed.

11.3. Gauss

Carl Friedrich Gauss (1777–1855) first began working on the fifth postulate in 1796, at the age of nineteen, and argued by contradiction, like his predecessors, but went much further in developing the theory in case (b). It is not clear when he came to the conclusion that no contradiction would arise. In a famous letter (1824) to his friend F.A.Taurinus, he explained that in the

case $\alpha + \beta + \gamma < \pi$ one obtains a "thoroughly consistent curious geometry", which he called "non-Euclidean". He concluded his letter by asking Taurinus not to tell anyone about his "private communication", which he was thinking of publishing at "some future time" .

Later, in 1832, he learned from his friend Farkas Bolyai that the latter's son, Janos, had arrived at the same conclusions. Later, in 1841, he found out that Lobachevsky had done the same. Gauss even learned Russian (to read Lobachevsky's early work?), but never directly communicated with either Janos Bolyai or Lobachevsky about these questions.

Portrait: google wikipedia Gauss in cyrillic

Carl Friedrich Gauss

The most amazing thing, however, is that Gauss, when he was not thinking about number theory or the fifth postulate, had constructed the differential geometry of surfaces, including surfaces of constant negative curvature, which are, in fact, a model (at least locally) of hyperbolic geometry. All these years, he had this model before his eyes, but never made the obvious connection with non-Euclidean geometry. He died without suspecting that a proof of the consistency of hyperbolic geometry was at his finger tips!

11.4. Lobachevsky

Nikolay Ivanovich Lobachevsky (1793–1856), like everybody else, tried to prove the fifth postulate by contradiction. As he progressed further in the case $\alpha + \beta + \gamma < \pi$, he became convinced that the theory was consistent.

In an unpublished textbook, written in 1823, he mentions that all attempts to prove the fifth postulate were erroneous. In 1826, Lobachevsky read a paper in Kazan about a "new geometry" (which he later called imaginary), and published (in Russian) a memoir about it in the Kazan Bulletin, which was unnoticed abroad. Trying to gain recognition, he published his work in German (*Geometrische Untersuchugen*, 1840) and in French (*Pangéométrie*, 1855), but without success (for an English translation of his work, see [11]).

Portrait in Wikipedia: google "Lobachevsky" in cyrillic

Nikolay Ivanovich Lobachevsky

N.I.Lobachevsky was not only the President (Rector, in the Russian terminology) of Kazan University, but also its Head Librarian. The Kazan library received many scientific periodicals, including the most famous mathematical journal of the time, *Crelle's Journal*. Library cards (which have come down to us) show that Lobachevsky read every issue of Crelle's Journal that reached Kazan, except two successive issues in the 1830ies. These two issues contained two papers by Mindling, in which the latter obtained, on surfaces of constant negative curvature, trigonometric formulas identical to the trigonometric formulas previously obtained by Lobachevsky on the hyperbolic plane. Had Lobachevsky seen one of these papers, he would have immediately observed that they constituted a proof of the consistency of hyperbolic geometry!

11.5. Bolyai

Janos Bolyai (1802–1860) was the son of a mathematician, Farkas Bolyai, who had "proved" the fifth postulate (his friend Gauss had pointed out his error). Janos first followed in his father's footsteps by trying to prove the fifth postulate by contradiction, but soon realized that he was obtaining a consistent geometry. In 1823 he wrote to his father: "Out of nothing I have created a strange new universe". But it was only in 1832 (three years after Lobachevsky) that his investigations were published in an Appendix to his father's book *Tentamen* (both were written in latin; for the German translation, see [15], the English translation of the Appendix appears in [17] and in [14], p.375).

Farkas sent the book to Gauss, asking to comment on the Appendix. Instead of praising and encouraging Janos, Gauss wrote that this would be "praising myself", since he had discovered the same things thirty years before, and the Appendix "spared him the effort" of writing up his discovery. Discouraged, Janos Bolyai stopped working for several years, but then started working on a book that would contain a detailed exposition of his results.

Portrait: google "Janos Bolyai Wikipedia"

Janos Bolyai

When Gauss had learned about Lobachevsky's results, he "kindly" communicated this fact to Janos Bolyai via the latter's father. For a while, Janos thought that Lobachevsky did not exist, that he was a creation of Gauss, who

used "Lobachevsky" as a pen name to publish results stolen from J.Bolyai's famous Appendix! Fortunately, Janos Bolyai finally understood that this was not the case, but he never finished his book, in fact published nothing more. He died fairly young, unrecognized by his contemporaries...

11.6. Beltrami, Helmholtz, Lie, Cayley, Klein, and Poincaré

The first proof of the consistency of hyperbolic geometry is attributed to Beltrami, who showed (1868) that its axioms and theorems hold (at least locally) on surfaces of constant negative curvature. The physicist Helmholtz was probably the first to understand how one can prove the consistency of hyperbolic geometry, but his arguments were regarded as unsufficiently rigorous by mathematicians. Sophus Lie improved the arguments of Helmhotz and was the first to stress the role of transformation groups in mathematics. Klein gave the definition of geometry that we introduced in Chapter 1, and, simultaneously with Cayley (but independently of him), gave an elementary global model of hyperbolic geometry; he also coined the terms *hyperbolic, parabolic, elliptic* for the three geometries. Poincaré constructed the two models of hyperbolic geometry that we discussed in Chapters 7 and 8.

10.7. Hilbert

David Hilbert made the first successful attempt to give an axiomatic exposition of Euclidean (space) geometry, rigorous in the modern sense of the word. It consists of 21 axioms, three undefined concepts (*point, line, plane*), and several undefined relations. Hilbert's axioms for plane geometry are presented and discussed in Appendix III of the present book.

The axiomatic approach is rarely used in teaching geometry in our time, because Euclidean geometry can be introduced in a much simpler way: it can easily be constructed as a branch of linear algebra over the real numbers (based on the fact that the straight line is "isomorphic" to the real numbers $\mathbb{R}$). This fact can be deduced from Hilbert's axioms by using the axiomatic definition of the real numbers and checking that these algebraic axioms are satisfied by the points of any line, provided the product and sum operation are appropriately defined on it.

Chapter 12
PROJECTIVE GEOMETRY

In this chapter, we introduce the main ideas of projective geometry for the particular case of $\mathbb{R}P^2$, the projective plane, and only have a brief look at projective space $\mathbb{R}P^3$. The general theory of d-dimensional projective spaces ($\mathbb{R}P^d$, $d \geq 1$) is traditionally studied in linear algebra courses by means of the so-called homogeneous coordinate model, but we do not go beyond the dimension $d = 3$. We use a more geometric approach, which may seem strange at first, because in our model "points" of $\mathbb{R}P^2$ will be lines in Euclidean space $\mathbb{R}^3$, but ultimately we will appeal to the homogeneous coordinate model.

12.1. The projective plane as a geometry

12.1.1. *Main definition.* The *projective plane* $\mathbb{R}P^2$ is defined as the geometry $(\mathbb{R}P^2 : \mathrm{Proj}(2))$, whose elements (called *projective points*) are straight lines in $\mathbb{R}^3$ passing through the origin O and whose transformation group $\mathrm{Proj}(2)$ is defined as follows. We start with the general linear group $GL(3)$ and identify any two linear transformations of $\mathbb{R}^3$ whose matrices can be obtained from each other by multiplication by nonzero constants; the composition of matrices is well defined on such equivalence classes of transformations, and $\mathrm{Proj}(2)$ is defined as the group whose elements are these classes and the group operation is composition (i.e., multiplication of matrices).

12.1.2. *Points and lines.* The elements of $\mathbb{R}P^2$ (projective points) are Euclidean lines; nevertheless, we will often simply call them *points* (of our geometry). The *straight lines* (of our geometry) are defined as the (Euclidean) planes passing through the origin. These definitions immediately imply the two following assertions.

I. *One and only one "line" passes through any two distinct "points".*

II. *Any two distinct "lines" intersect in one and only one "point".*

Thus there are no parallel lines in our geometry, just as in spherical geometry. But we will see that the two geometries are very different; in particular, there is no natural metric in projective geometry (and hence no measure of angles, no perpendiculars, no areas, and so on). Unlike spherical geometry, in which "straight lines" intersect in two points, in projective geometry lines intersect in one point, not two.

12.1.3. *Intuitive description.* You can imagine the projective plane as a Euclidean plane to which a "line at infinity" Λ_∞ has been added. When you move along a Euclidean line L to infinity in some direction, you intersect the line at infinity at some point $P = L \cap \Lambda_\infty$; if you move along L in the opposite direction, you will reach Λ and intersect it *at the same point* P. Parallels (in the Euclidean sense) intersect on the infinite line. Thus lines in $\mathbb{R}P^2$ are some kind of cycles (like "infinite circles"). The line at infinity, however, should not be regarded as a "special" line, because most projective transformations transform it into an "ordinary" line. The informal description of $\mathbb{R}P^2$ given here will be made rigorous in Subsection 12.2.4. below.

12.2. Homogeneous coordinates

12.2.1. Returning to our geometry $(\mathbb{R}P^2 : \mathrm{Proj}(2))$, let us introduce coordinates for our points. Each point L (i.e., each Euclidean line passing through the origin) is uniquely determined by its *direction vector*, i.e., by three coordinates (x_1, x_2, x_3), in the standard basis of $\mathbb{R}^3$, namely in the basis

$$e_1 = (1, 0, 0), \quad e_2 = (0, 1, 0), \quad e_3 = (0, 0, 1).$$

Conversely, however, points *do not* uniquely determine the coordinates: if λ is a nonzero real number, then $(\lambda x_1, \lambda x_2, \lambda x_3)$ determines the same point as (x_1, x_2, x_3). In this situation, we call the two sets of coordinates *equivalent*, denote the corresponding equivalence class by $(x_1 : x_2 : x_3)$, and refer to $\chi(L) = (x_1 : x_2 : x_3)$ as the *homogeneous coordinates* of the point L.

12.2.2. Homogeneous coordinates make the computation of the action of elements $g \in \mathrm{Proj}(2)$ on points $L \in \mathbb{R}P^2$ very easy: the transformation g is given by a 3×3 matrix $A_g \in \mathrm{GL}(3)$ (defined up to a constant), and

$$g(L) = A_g((x_1 : x_2 : x_3)) = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}.$$

The geometric meaning of the transformation with matrix A_g is that its column vectors are the images of the standard basis vectors under that transformation, but since A_g is defined up to a nonzero scalar, these images are also defined up to a nonzero scalar multiple.

12.2.3. *Projective spaces of higher dimensions.* In linear algebra courses, the *projective space* $\mathbb{R}P^d$, for any value of d, is defined in a similar way: its elements are homogeneous coordinates $(x_0 : x_1 : \cdots : x_d)$, i.e., equivalence

classes of $(d+1)$-tuples $(x_0 : x_1 : \cdots : x_d)$ of real numbers (not all equal to zero) up to multiplication by a nonzero constant. The group $\mathrm{Proj}(d+1)$ acts on each element by multiplication by $(d+1)\times(d+1)$ matrices corresponding to linear operators in $\mathbb{R}^{d+1}$ (defined up to a constant). We will not study higher-dimensional projective spaces $\mathbb{R}P^d$, $d > 3$, in this course. A detailed account can be found in most linear algebra courses. However, we will look at projective space $\mathbb{R}P^3$ briefly in Section 12.8 below.

12.2.4. Now let us describe a rigorous model of $\mathbb{R}P^2$ that will explain why $\mathbb{R}P^2$ is called the projective *plane*. In $\mathbb{R}^3$ consider the plane Π given by the equation $x_3 = 1$. Points of this plane have coordinates of the form $(x_1, x_2, 1)$. To the plane Π add the *line at infinity* Λ_∞ whose *points* are equivalence classes of Euclidean points $(x_1, x_2, 0)$ up to multiplication by a non-zero constant (notation $(x_1 : x_2 : 0)$). The set $\Pi \cup \Lambda_\infty$ is the set of points of the projective plane.

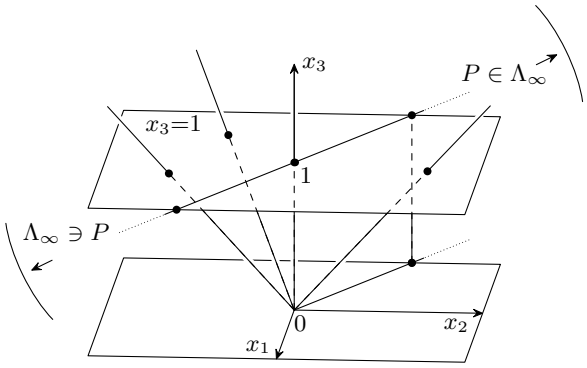


Figure 12.1. The projective plane

Note that the "points at infinity" $(x_1 : x_2 : 0) \in \Lambda_\infty$ determine Euclidean straight lines in the plane $x_3 = 0$. Intuitively, you should think of these lines as "pointing to infinity" in a certain direction, so that the set Λ_∞ "surrounds" the plane Π. More precisely, these lines are not rays, they are ordinary "two-sided" lines, and so they point to infinity in two opposite directions, but they intersect the projective line Λ_∞ at only one point (you should think of this point as being the identification of two diametrically opposite points at infinity). The reader familiar with elementary topology should recognize the classical topological model of $\mathbb{R}P^2$ obtained by identifying diametrically opposite points of the boundary of the unit disk $\mathbb{D}^2$.

The *lines* in this model of $\mathbb{R}P^2$ are the ordinary (Euclidean) lines in Π plus the "line" Λ_∞. There is an obvious bijection between the points and lines of $\mathbb{R}P^2$ (as defined in the previous section) and those in the model $\Pi \cup \Lambda_\infty$; in particular, the line Λ_∞ corresponds to the (Euclidean) plane $x_3 = 0$. Using this bijection, it is easy to define the action of Proj(2) in this model.

12.3. Projective transformations

12.3.1. One may want to ask: Why is our geometry called "projective", when it is defined by a group of *linear* operators in $\mathbb{R}^3$? Let us try to answer this question. Let Π_1 and Π_2 be two planes in $\mathbb{R}^3$ and let $P \in \mathbb{R}^3$ be a point. The *projection* of Π_1 to Π_2 from P is the map π that to each point $A \in \Pi_1$ assigns the point $A' \in \Pi_2$ at which the line PA intersects Π_2. This assignment is not necessarily bijective: π will be undefined at some points X (if PX is parallel to Π_2) and not onto (some points of Π_2 will not be covered), see Fig.12.2.

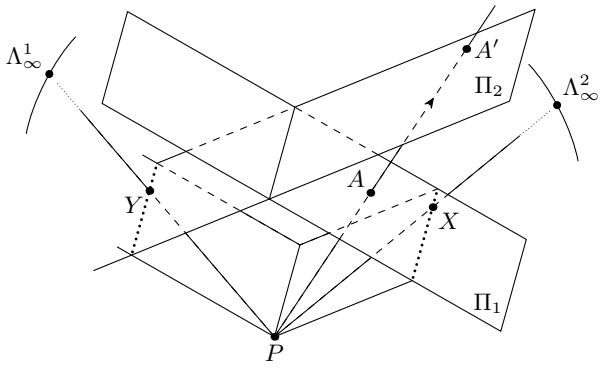


Figure 12.2. Projective transformations of planes

However, if we supply Π_1 and Π_2 with lines at infinity Λ^1_∞ and Λ^2_∞, and appropriately define the projection, then we obtain a bijection between the projective planes $\Pi_1 \cup \Lambda^1_\infty$ and $\Pi_2 \cup \Lambda^1_\infty$. The details are left to the reader.

12.3.2. A set of points $A_1, \dots, A_n$, $n \geq 3$, of the projective plane (interpreted as the model described in 12.2.4) are said to be *in general position* if for any three of them A_k, A_l, A_m, the vectors $\overrightarrow{OA}_k, \overrightarrow{OA}_l, \overrightarrow{OA}_m$ constitute a basis of $\mathbb{R}^3$. If one of the points, say A_i, lies on the line at infinity, the vector $\overrightarrow{OA}_1$ is well defined, in coordinates it has the form $(a : b : 0)$. If three points or more from our collection lie on the infinite line, then, of course, the collection will not be in general position.

Another way of defining a collection of points in general position is to say that no three of them lie on the same line.

12.3.3. Theorem. *There exists one and only one projective transformation that takes four points $A, B, C, D \in \mathbb{R}P^2$ in general position to four other points $A', B', C', D' \in \mathbb{R}P^2$ in general position.*

Proof. In accordance with our model of the projective plane, we can think of the points A, B, C and A', B', C as lying in the plane $x_3 = 1$. By assumption, the vectors $\overrightarrow{OA}, \overrightarrow{OB}, \overrightarrow{OC}$ constitute a basis of $\mathbb{R}^3$. Let (a_1, a_2, a_3), (b_1, b_2, b_3), (c_1, c_2, c_3) be the coordinates of the vectors $\overrightarrow{OA'}, \overrightarrow{OB'}, \overrightarrow{OC'}$ in that basis. Then the matrix

$$M = \begin{pmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{pmatrix}$$

can be regarded as a linear transformation of $\mathbb{R}^3$ taking A, B, C to A', B', C. Now let us multiply the columns of this matrix by scalar constants, obtaining the matrix

$$A_g = \begin{pmatrix} \lambda a_1 & \mu b_1 & \nu c_1 \\ \lambda a_2 & \mu b_2 & \nu c_2 \\ \lambda a_3 & \mu b_3 & \nu c_3 \end{pmatrix}$$

which we now regard as defining an element g of Proj(2). Clearly, A_g takes the points $A, B, C \in \mathbb{R}P^2$ to the points $A', B', C' \in \mathbb{R}P^2$, although the same matrix regarded as acting in $\mathbb{R}^3$ *does not* take $A, B, C \in \mathbb{R}^3$ to $A', B', C' \in \mathbb{R}^3$ (when not all three of the scalars λ, μ, ν are equal to 1).

Now let us denote by (d_1, d_2, d_3) the coordinates of the point D in the basis $\overrightarrow{OA}, \overrightarrow{OB}, \overrightarrow{OC}$ and by (d'_1, d'_2, d'_3) the coordinates of the point D' in the same basis. We claim that it is possible to choose the scalar parameters λ, μ, ν so that A_g will take $D \in \mathbb{R}P^2$ to $D' \in \mathbb{R}P^2$.

Indeed, this will be case if the matrix A_g applied to the vector (d_1, d_2, d_3) will give the vector (d'_1, d'_2, d'_3), or, which is the same thing, the system of equations

$$\begin{cases} a_1 d_1 \lambda + b_1 d_2 \mu + c_1 d_3 \nu = d'_1 \\ a_2 d_1 \lambda + b_2 d_2 \mu + c_2 d_3 \nu = d'_1 \\ a_3 d_1 \lambda + b_3 d_2 \mu + c_3 d_3 \nu = d'_1 \end{cases}$$

in the unknowns λ, μ, ν will have a solution. But the determinant Δ of this system can be expressed as $\Delta = d_1 d_2 d_3 \det(M)$ and so is nonzero. Hence our

system of equations has a nonzero solution in λ, μ, ν. Thus we have shown that $A_g(D) = D'$ (if we choose for the values of λ, μ, ν the solution of our system) and proved the existence of the required projective transformation.

Its uniqueness follows by working out the construction of A_g in reverse order, which will bring us back to the same matrix (up to multiplication by a scalar). $\square$

12.4. Cross-ratio of collinear points

12.4.1. *Main definitions.* We mentioned above that there is no natural metric on the projective plane, and no affine structure (the ratio of the two segments determined by *three* collinear points of $\mathbb{R}P^2$ is not well defined). Nevertheless, the affine structure in $\mathbb{R}^3$ allows us to define the cross ratio of any *four* ordered collinear points of $\mathbb{R}P^2$.

The definition is the following. Let k, l, m, n be collinear points in $\mathbb{R}P^2$, i.e., four coplanar lines of $\mathbb{R}^3$ passing through the origin; suppose a line s cuts our four lines at the points A, B, C, D, respectively. Then the vectors $\overrightarrow{AC}$ and $\overrightarrow{BC}$ are proportional, i.e., $\overrightarrow{AC} = \lambda \overrightarrow{BC}$; the real number λ (which may be negative) is denoted by $\langle A, B, C\rangle$; the number $\langle A, B, D\rangle$ is defined similarity. We now put

$$\langle A, B, C, D\rangle := \frac{\langle A, B, C\rangle}{\langle A, B, D\rangle};$$

the number thus obtained is called the *cross-ratio* of the points A, B, C, D. It is not difficult to show that it is well defined, i.e., does not depend on the choice of the secant line s. Now if one of the points, say B, lies on the infinite line Λ_∞, then we put $\langle A, B, C, D\rangle := \langle C, D, A\rangle$ (similarly for the others).

12.4.2. *Coordinate expressions.* The cross ratio is easy to compute in coordinates. To this end, we return to the model

$$\Pi = \{(x, y, z) \in \mathbb{R}^3 | z = 1\} \subset \mathbb{R}P^2 = \Pi \cup \Lambda_\infty$$

and suppose that the collinear points A, B, C, D have the coordinates:

$$(x_A, y_A, 1),\ (x_B, y_B, 1),\ (x_C, y_C, 1),\ (x_D, y_D, 1).$$

Then, obviously,

$$\langle A, B, C\rangle = \frac{x_C - x_A}{x_C - x_B} = \frac{y_C - y_A}{y_C - y_B},\ \langle A, B, D\rangle = \frac{x_D - x_A}{x_D - x_B} = \frac{y_D - y_A}{y_D - y_B}$$

and therefore

$$\langle A, B, C, D\rangle = \frac{x_C - x_A}{x_C - x_B} : \frac{x_D - x_A}{x_D - x_B} = \frac{y_C - y_A}{y_C - y_B} : \frac{y_D - y_A}{y_D - y_B}.$$

If one of the points, say B, is on the infinite line (at its intersection with the line containing the points A, C, D), then the cross ratio reduces to the ordinary ratio What happens in this case may be described by saying that "the infinities cancel":

$$\frac{x_C - x_A}{x_C - \infty} : \frac{x_D - x_A}{x_D - \infty} = \frac{x_C - x_A}{x_D - x_A} = \langle C, D, A\rangle.$$

In the case when all four points A, B, C, D lie on the infinite line, their cross ratio is also a well defined real number. Its calculation is the object of Exercise 12.3.

12.4.3. Theorem. *The cross-ratio of four collinear points is invariant under projective transformations.*

Proof. The proof is a problem in linear algebra; see Exercise 12.4. □

12.5. Projective duality

12.5.1. Points and lines on the projective plane ($\mathbb{R}P^2$: Proj(2)) play, in a certain sense, symmetric roles. This will be easier to see if we introduce the notion of *incidence*: we will say that two lines a and b are *incident at the point* P if P is the intersection point of the lines a and b, and that the two points P and Q are *incident at the line* a if a passes through P and Q. Also, together with the standard term *collinear* (used for points all lying on one line) we will use the term *copunctal* for lines all passing through one and the same point.

Given an assertion of projective geometry formulated in this terminology, we can translate it into another statement, called *dual*, by replacing the word "line" by the word "point" (and "collinear" by "'copunctal") and vice versa. For example, statement I from Section 12.1 can be expressed as: "One and only one line is incident to two distinct points"; its translation (i.e., the dual statement) will be "One and only one point is incident to two distinct lines", which is exactly the assertion of II (see Section 12.1). Another example: "Any projective transformation takes collinear points to collinear points" translates to "Any projective transformation takes copunctal lines to copunctal lines".

What is remarkable is that this kind of translation always translates true statements to true statements. To prove this, we will define the *dual geometry*

to the geometry of $\mathbb{R}P^2$: it is the geometry $(D\mathbb{R}P^2 : \mathrm{Proj}(2))$ whose *points* are planes of $\mathbb{R}^3$ passing through the origin under the action of the group of linear nondegenerate transformations of $\mathbb{R}^3$. In $D\mathbb{R}P^2$, the intersections of two points (i.e., Euclidean planes) will be called the *line passing through the points* (it is actually a Euclidean line in Euclidean 3-space).

12.5.2. Theorem. *The two geometries $(D\mathbb{R}P^2 : Proj(2))$ and $(\mathbb{R}P^2 : Proj(2))$ are isomorphic: there is a bijection, called duality and denoted by D, between the sets of points of the two geometries compatible with an isomorphism of $GL(3)$ onto itself.*

Proof. To each "point" Π of $D\mathbb{R}P^2$, i.e., to each plane of $\mathbb{R}^3$ given by the equation $a_1x_1+a_2x_2+a_3x_3 = 0$, we assign the point of $\mathbb{R}P^2$ with homogeneous coordinates $(a_1 : a_2 : a_3)$ (which is of course the Euclidean line passing through the origin and perpendicular to the plane). If an element $g \in \mathrm{Proj}(2)$ takes the point $(a_1 : a_2 : a_3)$ to some point $(b_1 : b_2 : b_3)$, then the same element will take the plane Π to the plane given by $b_1x_1 + b_2x_2 + b_3x_3 = 0$. Thus the duality map $D : \mathbb{R}P^2 \to D\mathbb{R}P^2$ (which is obviously bijective) is compatible with the action of Proj(2), so that we have constructed the required isomorphism. □

Note that the duality correspondence is an *involution*, i.e., $D \circ D$ identically maps $\mathbb{R}P^2$ onto itself. Further, note that the isomorphism constructed above preserves incidence: if two points A, B of $\mathbb{R}P^2$ (i.e., two Euclidean lines passing through the origin O of $\mathbb{R}^3$) are incident to the line l (i.e., are contained in a Euclidean plane Π_l), then the two lines $D(A), D(B)$ in $D\mathbb{R}P^2$ intersect in the point (of $D\mathbb{R}P^2$) $D(l) = \Pi_l$. Thus we have the following statement.

12.5.3. Corollary: Duality Principle. *There is a bijection between the set of lines and the set of points of $\mathbb{R}P^2$ that preserves incidence and takes any theorem of projective geometry to a theorem of projective geometry.*

12.6. Conics in $\mathbb{R}P^2$

The nondegenerate conic sections (or *conics* for short) in the Euclidean plane are, as is well known, the ellipse, the hyperbola and the parabola. In $\mathbb{R}P^2$, these three curves are projectively equivalent, so that *there exists only one nondegenerate conic in $\mathbb{R}P^2$ (up to projective equivalence)*.

A conic in $\mathbb{R}P^2$ can be defined as any set of points obtained from the curve C given by $(x_1)^2 + (x_2)^2 = 1$ (in the plane-with-line-at-infinity model

described in Section 12.2, this curve is the Euclidean circle) by a projective transformation. Any projective transformation under which the image of C does not intersect the line at infinity Λ_∞ transforms C into an ellipse; a projective transformation that takes one point of C to Λ_∞ transforms C into a parabola, and a projective transformation that takes two points of C to Λ_∞ transforms C into a hyperbola.

12.7. The Pappus, Desargues, and Pascal theorems

We conclude our study of $\mathbb{R}P^2$ with three beautiful classical theorems. All three can be regarded as theorems about points and lines either in the projective plane or in the affine (in particular Euclidean) plane.

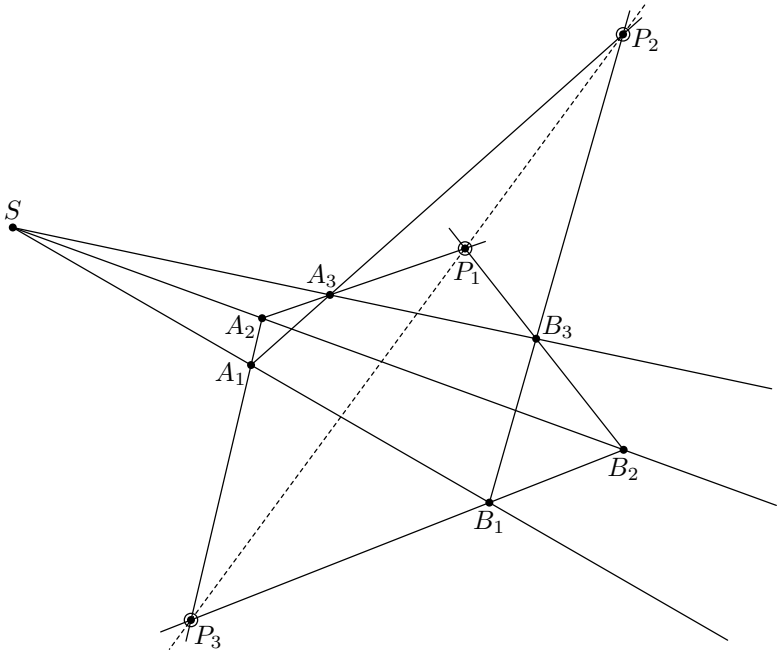


Figure 12.3. Desargues' theorem

12.7.1. Desargues' Theorem. *Suppose that the lines joining the corresponding vertices of triangles $A_1A_2A_3$ and $B_1B_2B_3$ intersect at one point S. Then the intersection points P_1, P_2, P_3 of the lines A_2A_3 and B_2B_3, A_3A_1 and B_3B_1, A_1A_2 and B_1B_2, respectively, are collinear.*

Proof. We begin by passing from the plane to 3-space and prove the three-dimensional analog of Desargue's theorem. (The proof of the 3-D theorem

turns out to be unexpectedly simple, but the argument used in it doesn't work in the plane!). We then use the 3-D theorem to prove Desargue's theorem in the plane by means of a continuous deformation of the spatial picture to the planar one.

Suppose we are given two triangles $A_1\widehat{A}_2A_3$ and $B_1\widehat{B}_2B_3$ in Euclidean space $\mathbb{R}^2$ such that the three lines A_1B_1, $\widehat{A}_2\widehat{B}_2$, A_3B_3 intersect at one point S. (The reader should think of the points A_1, B_1, A_3, B_3, S as being the same as in the planar version of the theorem, while the points A_2, B_2 have been "lifted out" of the plane.) Then the lines SB_1, $S\widehat{B}_2$, SB_3 define a trihedral angle in $\mathbb{R}^3$ (see Fig.12.4).

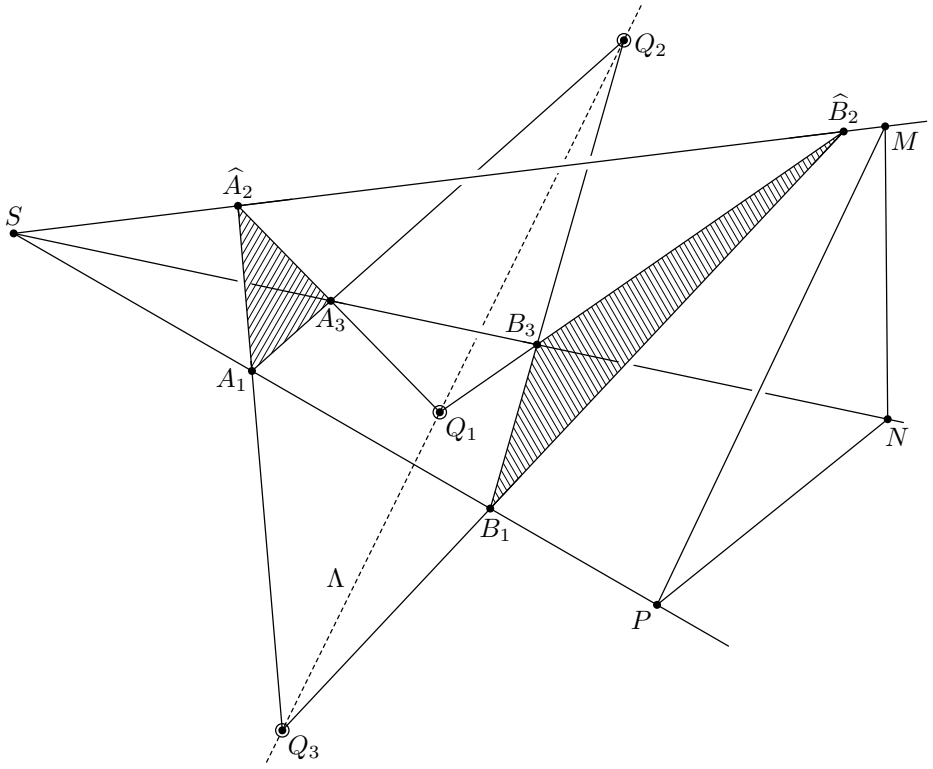


Figure 12.4. Desargues' theorem in space

Consider the three pairs of lines $\widehat{A}_2A_3$ and $\widehat{B}_2B_3$, $\widehat{A}_2A_1$ and $\widehat{B}_2B_1$, A_1A_3 and B_1B_3. We claim that *each of these pairs has a common point (in space!) and these three points are collinear.*

Indeed, the (Euclidean) planes $\Pi_1 := (A_1\widehat{A}_2A_3)$ and $\Pi_2 := (B_1\widehat{B}_2B_3)$

intersect in a line Λ. Obviously, the lines $\widehat{A}_2A_3$ and $\widehat{B}_2B_3$ intersect at a point (denoted Q_1) of Λ, and so do the lines $\widehat{A}_2A_1$ and $\widehat{B}_2B_1$ (the intersection point is denoted by Q_3) as well as the lines A_1A_3 and B_1B_3 (at Q_2). Since the points Q_1, Q_2, Q_3 all lie on Λ, they are collinear, as claimed.

Let us pass to the proof of the planar version of the theorem.

Consider the plane B_1SB_3 (which we think of as being "horizontal"), construct a plane perpendicular to it through the line SB_2, in that plane choose a point O "below" the horizontal plane, and choose points $\widehat{A}_2$ and $\widehat{B}_2$ so that S, $\widehat{A}_2$, $\widehat{B}_2$ are collinear by projecting the points A_2, B_2 from O (see Fig.12.5).

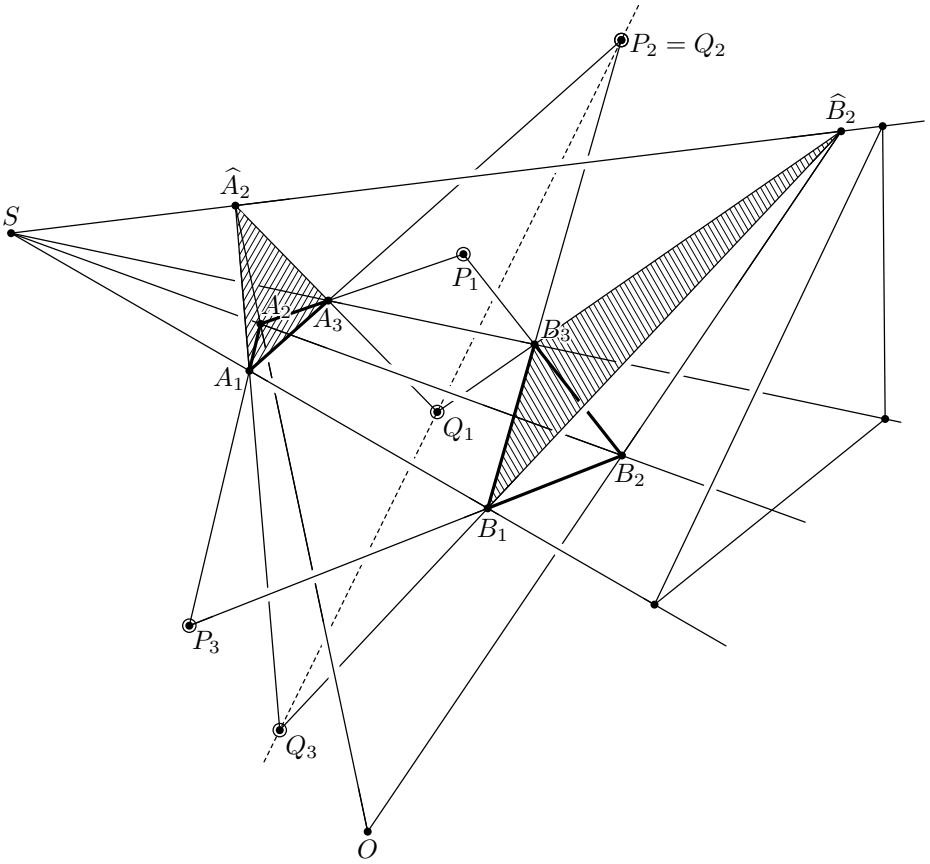


Figure 12.5. Proof of Desargues' theorem

Using the 3-D version of the theorem, we can now construct the three collinear points Q_1, Q_2, Q_3. Now rotate the line $S\widehat{B}_2$ about S downward

in the vertical plane until it coincides with SB_2. Since the mobile points Q_1, Q_2, Q_3 will always be collinear and, when they reach the horizontal plane, they will coincide with the points P_1, P_2, P_3, it follows that these three points are collinear. This proves the theorem. □

12.7.2. Pappus' Theorem. *Suppose the points A_1, A_2, A_3 are collinear, and the points B_1, B_2, B_3 are collinear. Let P_1, P_2, P_3 be the intersection points of the lines A_2B_1 and A_1B_2, A_1B_3 and A_3B_1, A_2B_3 and A_3B_2, respectively. Then the points P_1, P_2, P_3 are collinear.*

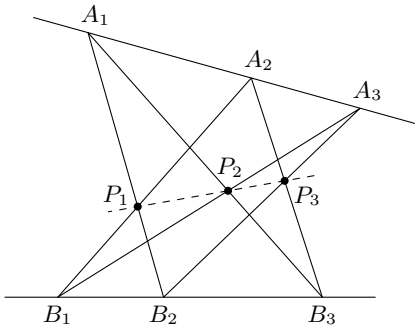


Figure 12.6. Pappus' theorem

Sketch of the proof. By Theorem 12.3.3, we can assume that $A_1A_2B_1B_2$ is a square. Using the coordinate system with basis $\overrightarrow{A_1A_2}$ $\overrightarrow{A_1B_1}$, it is an easy exercise to prove that the points P_1, P_2, P_3 are collinear.

12.7.3. Pascal's Theorem. *The points A, B, C, D, E, F lie on a conic. Let P_1, P_2, P_3 be the intersection points of the lines AB and ED, AF and CD, CB and EF, respectively. Then the points P_1, P_2, P_3 are collinear.*

The theorem is illustrated by Figure 12.7, in which the conic is a circle. In fact, Pascal actually proved the theorem in this particular case without any loss of generality – he knew all conics are projectively equivalent to the circle. Here we do not present the (not very difficult) proof of his theorem.

12.7.4. Remark. Note that the theorem is true in $\mathbb{R}P^2$ as well as in $\mathbb{R}^2$. To formulate it in full generality as a Euclidean theorem, one has to consider several singular cases (which arise when one of the points P_i "goes to infinity"); in these cases the proof differs somewhat from the proof in the generic case. Note also that the Euclidean versions have metric proofs (see

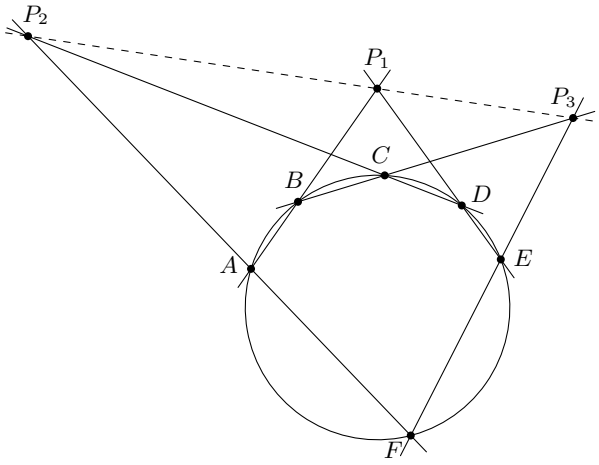


Figure 12.7. Pascal's theorem

Exercise 12.14), but the projective proof is, in a sense, more natural. Similar remarks hold for the Pappus and the Desargues theorems.

12.8. Projective space $\mathbb{R}P^3$

In this section we very briefly describe three-dimensional projective geometry.

12.8.1. *Definition of projective space.* The projective space $\mathbb{R}P^3$ can be defined in terms of homogeneous coordinates as explained in Subsection 12.2.3, but here we adopt a more geometric approach. Namely, we consider four-dimensional Euclidean space $\mathbb{R}^4$ and for the *points* of $\mathbb{R}P^3$ take the straight lines passing through the origin O of $\mathbb{R}^4$ and define the transformation group Proj(3) of $\mathbb{R}P^3$ as in the two-dimensional case (using GL(4) instead of GL(3)). We then define the *lines* of $\mathbb{R}P^3$ as the planes passing through the origin O and its *planes* as the three-dimensional hyperplanes of $\mathbb{R}^4$ passing through O.

The following basic statements immediately follow from the above definitions.

I. *One and only one "line" passes through any two distinct "points".*

II. *Any two distinct "planes" intersect in one and only one "line".*

Thus there are no parallel lines or parallel planes in this geometry. Moreover, there is no distance function in $\mathbb{R}P^3$, and so no measure of areas or angles, and no perpendiculars.

12.8.2. *Properties of projective transformations.* Without going into

details, let us just mention that there is a "five point theorem" similar to the "four point theorem" 12.3.3 and that the cross-ratio of four collinear points is invariant under projective transformations. There is neat theory of quadrics (surfaces given by second degree equations) in which, for example, the hyperboloid of two sheets is (projectively) equivalent to the hyperboloid of one sheet and to the ellipsoid.

12.8.3. *Projective duality in space.* Just as in $\mathbb{R}P^2$, in $\mathbb{R}P^3$ there is a *duality principle*, but a somewhat more sophisticated one: it involves not only points and lines, but also planes. After replacing the expressions "passing through", "intersecting in", etc. by appropriate versions of the notion of incidence and using the expressions "copunctal' and "coplanar" in the formulation of a theorem, we obtain the *dual theorem* simply by interchanging the words "point" and "plane" (and not changing the word "line", which is self-dual). The dual theorem will also be correct, since its proof can be obtained by "dualizing" the proof of the original theorem. For example, the properties I and II are dual to each other.

12.9. Problems

12.1 Five distinct collinear points A, B, C, D, E are given. Prove that

$$\langle A, B, C, D\rangle \cdot \langle A, B, D, E\rangle \cdot \langle A, B, E, C\rangle = 1.$$

12.2. How many different values does the cross-ratio of four points on a line take when the order of the points is changed?

12.3. Calculate the cross-ratio of four points $(x_i : y_i; 0)$, $i = 1, 2, 3, 4$ lying on the infinite line Λ_∞.

12.4. Prove Theorem 12.4.3.

12.5. Four planes pass through a common line l, while the line m intersects all four planes. Prove that the cross-ratio of the intersection points of m with the planes does not depend on the choice of m.

12.6. State and prove the theorem dual to the Pappus theorem. Draw the corresponding picture.

12.7. State and prove the theorem dual to Desargues' theorem. Draw the corresponding picture.

12.8*. Prove that under projective duality any point on a conic is taken to a line tangent to the dual conic.

12.9. Using Exercise 12.8, state and prove the theorem dual to Pascal's theorem (the dual theorem is known as *Brianchon's Theorem*). Draw the corresponding picture.

12.10. Three skew lines l, l_1, l_2 in $\mathbb{R}^3$ are given. To a point $A_1 \in l_1$ let us assign the point A_2 at which the line l_2 intersects the plane determined by A_1 and l. Prove that the assignment $A_1 \mapsto A_2$ is a projective map of l_1 onto l_2.

12.11. The lines $l_1, \dots l_{n-1}$ and l are given on the plane. The points $O_1, \dots, O_n$ are chosen on l. The lines containing the sides of a polygon $A_1, \dots, A_n$ pass through the points $O_1, \dots, O_n$ while its vertices $A_1, \dots, A_{n-1}$ move along the lines $l_1, \dots l_{n-1}$. Prove that the vertex A_n also moves along a straight line.

12.12. Compute the cross-ratios of the quadruple of points A, B, C, D in Figure 12.8.

12.13. Prove the triangle inequality for the hyperbolic metric by using appropriate projective transformations.

12.14. Prove the Euclidean version of Pascal's theorem for the case of the circle.

Chapter 13

"PROJECTIVE GEOMETRY IS ALL GEOMETRY"

The title of this chapter is a quotation from Arthur Cayley, the outstanding 19th century British mathematician, one of the founders of projective geometry. The aim of this chapter is to give a precise mathematical meaning to these words, namely to show that the three principal continuous geometries, parabolic (Euclid), hyperbolic (Lobachevsky), and elliptic (Riemann) are *subgeometries* of projective geometry. We will prove this in dimension two, i.e., show that the projective plane "contains" (in a certain precise sense) the hyperbolic plane, the elliptic plane, and the Euclidean plane. Since the discrete geometries that we also studied in this book are, in turn, subgeometries of the three principal continuous ones, this means that all the geometries studied so far in this course are parts of projective geometry.

But first we recall the notion of subgeometry, which appeared briefly in Chapter 1.

13.1. Subgeometries

13.1.1. Recall that two geometries $(X : G)$ and $(Y : H)$ are isomorphic if there is an equivariant bijection between them, i.e., a bijection between their sets of points and an isomorphism between their transformation groups which are compatible (for the detailed definition, see Chapter 1). Further, the geometry $(X : G)$ is a *subgeometry* of $(Y : H)$ if there is an injective map $i : X \to Y$ and a monomorphism $\gamma : G \to H$ compatible with the group actions, i.e., satisfying $(i(x))(\gamma(g)) = i(xg)$. (In this formula, we use the notation xg for the result of the action of the element $g \in G$ on the point $x \in X$; thus $(i(x))(\gamma(g))$ stands for the result of the action of the element $\gamma(g) \in H$ on the point $i(x) \in Y$.)

Of course any geometry isomorphic to the given one is its subgeometry, but we are interested in the case when it is a *proper* subgeometry, i.e., when i is not a bijection, or γ is not an isomorphism, or both.

13.1.2. Here are some toy examples of proper subgeometries:

• the motion group of the regular dodecagon (regular polygon of 12 sides) is a subgeometry of the dodecahedron with dihedral group $\mathbb{D}_{12}$ acting on it;

• the dihedral group $\mathbb{D}_6$ acting on the regular dodecahedron defines a subgeometry of the same dodecahedron with the dihedral group $\mathbb{D}_{12}$ acting on it;

• the dihedral group $\mathbb{D}_6$ acting on the regular dodecahedron defines a subgeometry of Euclidean plane geometry $(\mathbb{R}^2 : \mathrm{Isom}(\mathbb{R}^2))$.

13.2. The Euclidean plane as a subgeometry of $\mathbb{R}P^2$

13.2.1 The fact that the Euclidean plane $(\mathbb{R}^2 : \mathrm{Isom}(\mathbb{R}^2))$ is a subgeometry of the projective plane $(\mathbb{R}P^2 : \mathrm{Proj}(2))$ is rather obvious if we interpret $\mathbb{R}P^2$ (in the homogeneous coordinate model, see Section 11.2) as the plane

$$\Pi = \{(x_1, x_2, x_3) \in \mathbb{R}^3 \,|\, x_3 = 1\}$$

supplied with the "line at infinity" $\Lambda_\infty = \{(x_1 : x_2 : x_3) | x_3 = 0\}$, i.e., if we take $\mathbb{R}P^2 = \Pi \cup \Lambda_\infty$.

Indeed, let us define $i : \mathbb{R}^2 \to \mathbb{R}P^2 = \Pi \cup \Lambda_\infty$ in the obvious way, i.e., by setting $i((x_1, x_2)) := (x_1, x_2, 1)$ and define $\gamma : \mathrm{Isom}(\mathbb{R}^2) \to GL(3)$ as follows. Let $g \in \mathrm{Isom}(\mathbb{R}^2)$, let $(\overrightarrow{AB}, \overrightarrow{AC})$ be an orthonormal frame in $\mathbb{R}^2$ and $(\overrightarrow{A'B'}, \overrightarrow{A'C'})$ be its image under g. For $\gamma(g)$ we take the element of Proj(2) that takes the three lines OA, OB, OC to the three lines OA', OB', OC'. This construction is shown in the figure.

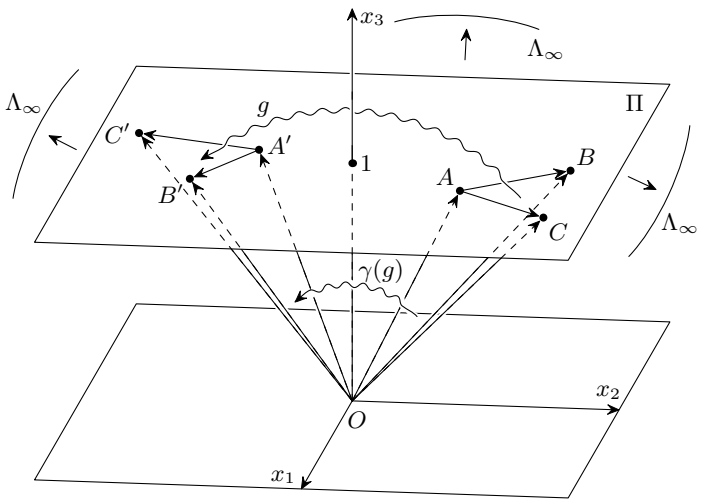


Figure 13.1. The Euclidean plane as a subgeometry of $\mathbb{R}P^2$

13.2.2. Theorem. *The construction described above shows that the Euclidean plane is a subgeometry of the projective plane.*

Proof. The theorem is obvious: clearly, i is injective, γ is a monomorphism, and the fact that compatibility holds is also immediate. □

13.3. The hyperbolic plane as a subgeometry of $\mathbb{R}P^2$

13.3.1. The fact that the hyperbolic plane $(\mathbb{H}^2 : M)$ is a subgeometry of the projective plane $(\mathbb{R}P^2 : \mathrm{Proj}(2))$ is best seen by using the Cayley–Klein model and interpreting $\mathbb{R}P^2$ (as in Section 12.2 above) as the plane

$$\Pi = \{(x_1, x_2, x_3) \in \mathbb{R}^3 \,|\, x_3 = 1\}$$

supplied with the "line at infinity" $\Lambda_\infty = \{(x_1 : x_2 : x_3) | x_3 = 0\}$, i.e., by taking $\mathbb{R}P^2 = \Pi \cup \Lambda_\infty$.

Recall that the Cayley–Klein model was defined as $(\mathbb{H}^2 : \mathrm{Isom}_\lambda(\mathbb{H}^2))$, where $\mathbb{H}^2$ is the unit open disk and λ is the metric given by the formula $\lambda(A, B) = (1/2)|\ln(\langle A, B, X, Y\rangle)|$ (for the details, see Section 9.2).

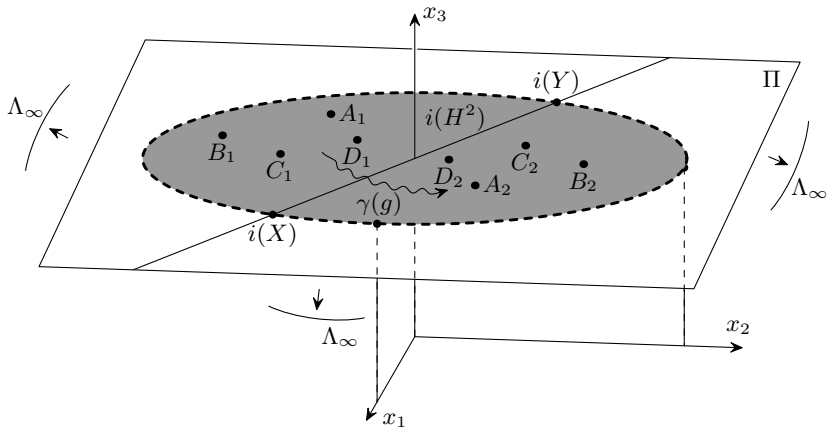


Figure 13.2. The Cayley–Klein model as a subgeometry of $\mathbb{R}P^2$

Now let us define $i : \mathbb{H}^2 \to \mathbb{R}P^2 = \Pi \cup \Lambda_\infty$ in the obvious way, i.e., by setting $i((x_1, x_2)) := (x_1, x_2, 1)$ and define $\gamma : \mathrm{Isom}_\lambda(\mathbb{H}^2) \to \mathrm{Proj}(2)$ as follows. Let $g \in \mathrm{Isom}_\lambda(H^2)$. Take four points $A, B, C, D \in \mathbb{H}^2$ in general position and consider their images $Ag, Bg, Cg, Dg \in \mathbb{H}^2$ under g. Denote

$$A_1 = i(A), \quad B_1 = i(B), \quad C_1 = i(C), \quad D_1 = i(D) \in \mathbb{H}^2,$$

$$A_2 = i(Ag), \quad B_2 = i(Bg), \quad C_2 = i(Cg), \quad D_2 = i(Dg) \in \mathbb{H}^2.$$

The two quadruples of points $A_i, B_i, C_i, D_i \in \Pi$, $i = 1, 2$, are in general position, and so by Theorem 12.3.3 there exists a unique projective transformation taking A_1, B_1, C_1, D_1 to A_2, B_2, C_2, D_2; we take this transformation to be $\gamma(g)$. The construction is shown on the figure.

Basically, this construction is simply the natural extension of the action of g from the open unit disk

$$\{(x_1, x_2, 1) \,|\, x_1^2 + x_2^2 < 1\} = i(\mathbb{H}^2)$$

to the entire projective plane. To any "straight line" of $\mathbb{H}^2$ (i.e., any chord XY of the unit circle) corresponds the straight line joining the points $i(X)$, $i(Y)$ in the projective plane; to parallel or nonintersecting lines in $\mathbb{H}^2$ (chords of the unit circle) correspond straight lines in $\mathbb{R}P^2$ that actually intersect (at a point outside the disk $i(\mathbb{H}^2)$, possibly on the "infinite line" Λ_∞).

13.3.2. Theorem. *The construction described above shows that the hyperbolic plane is a subgeometry of the projective plane.*

Proof. The map i is obviously injective, so that it remains to show that the restriction of $\gamma(g)$ to the open disk $\{(x_1, x_2, 1) \,|\, x_1^2 + x_2^2 < 1\} = i(\mathbb{H}^2)$ coincides with γ. This is a consequence of the fact that projective transformations preserve the cross ratio of any four collinear points, and therefore preserve the distance λ between points inside $i(\mathbb{H}^2)$ (λ being the absolute value of the logarithm of the appropriate cross ratio). But g is an isometry (with respect to λ), it coincides with the restriction of $\gamma(g)$ to $i(\mathbb{H}^2)$ on three noncollinear points, therefore it coincides with this restriction on all of $i(\mathbb{H}^2)$. This proves the theorem.□

13.3.3. Remark. It can be proved that the subgroup of Proj(2) that takes the circle $\{(x_1, x_2, 1) \,|\, x_1^2 + x_2^2 = 1\}$ to itself is actually isomorphic to $\mathrm{Isom}_\lambda(\mathbb{H}^2)$), and this isomorphism is often used to establish various formulations expressing the fact that the hyperbolic plane is "a part of" the projective plane. We do not need this fact in our approach to this topic, so we omit the proof here.

13.4. The Riemannian elliptic plane as a subgeometry of $\mathbb{R}P^2$

13.4.1. As in the two previous sections, we regard $\mathbb{R}P^2$ as the plane Π with the line at infinity Λ_∞ added to it. Our model of Riemannian two-dimensional elliptic geometry $\mathbb{E}\mathrm{ll}^2$ will be the standard one, i.e., the unit sphere with its antipodal points identified: $\mathbb{E}\mathrm{ll}^2 = (\mathbb{S}^2/_{Ant} : \mathrm{O}(3))$. We think of this sphere as lying on the plane Π, touching it at the point $(0, 0, 1)$.

We first construct the inclusion (which will actually be a bijection) of $\mathbb{S}^2/_{Ant}$ to $\mathbb{R}P^2$ by simply projecting it from the center of the sphere onto $\Pi \cup \Lambda_\infty$. Note that "straight lines" in $\mathbb{S}^2/_{Ant}$ (i.e., great circles of the sphere with diametrically opposed points identified) will be mapped to straight lines of the projective plane, in particular, the equator of the sphere will be mapped to the "infinite line" Λ_∞. Note also that spherical triangles (not intersecting the equator) will be projected to ordinary rectilinear triangles in Π, but their angles will not be preserved.

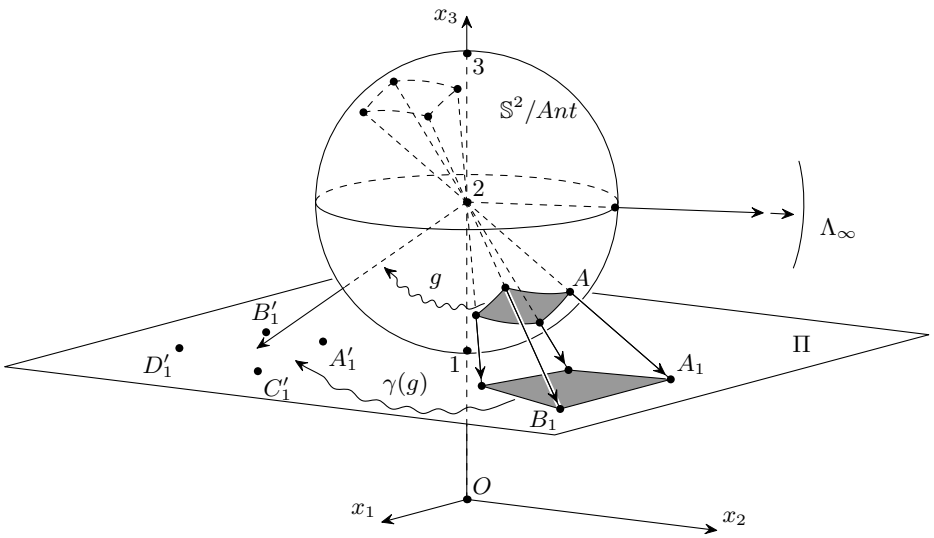


Figure 13.3. Bijection between the elliptic plane and $\mathbb{R}P^2$

To construct the monomorphism $\gamma : O(3) \to \mathrm{Proj}(2)$, we choose two perpendicular arcs AB and AC (that do not intersect the equator) and denote by BCD the triangle symmetric to triangle ABC with respect to the line BC. Denote by A_1, B_1, C_1, D_1 the central projections of the points A, B, C, D to the plane Π. Now suppose $g \in O(3)$ takes the points A, B, C, D to A', B', C', D', and denote by A_1', B_1', C_1', D_1' their projections to Π. We define $\gamma(g)$ as the projective transformation that takes A', B', C', D' to A_1', B_1', C_1', D_1' (such a projection exists and is unique by Theorem 12.3.3). The construction is shown on the figure.

13.4.2. Theorem. *The construction described above shows that the Riemannian elliptic plane is a subgeometry of the projective plane.*

Proof. The theorem is an easy consequence of the following lemma, whose proof is the object of Exercise 13.3.

13.4.3. Lemma. *The map γ described above is a monomorphism of* O(3) *to* Proj(2).

Indeed, the monomorphism γ is compatible with the map i by construction, so that the theorem follows. □

13.5. Problems

13.1. Prove that any projective transformation of the projective plane $\mathbb{R}P^2$ preserves the cross-ratio of collinear points

13.2. Prove that, conversely, any transformation of the projective plane that preserves the cross-ratio of all collinear points is projective.

13.3. Prove Lemma 13.4.3.

13.4. Give an example of a spherical triangle whose angle sum is close to 2π and describe its image under the central projection defined in §12.4.

13.5. Show that for any $\varepsilon > 0$ and any positive number S, there exists a spherical triangle of area less than ε whose image under the central projection defined in Section 13.4 is of area greater than S.

13.7. Prove that the subgroup of projective transformations that take the unit circle centered at the origin to itself is isomorphic to the isometry group of the hyperbolic plane.

13.8. Generalize and solve the previous problem by replacing the circle by an arbitrary oval (nondegenerate second degree curve).

.

Chapter 14
FINITE GEOMETRIES

A finite geometry is geometry whose set of points is finite. In that situation, the possibilities for the transformation group are extremely varied, and Klein's definition of geometry is too general to single out those finite geometries that actually deserve to be called geometries. Thus one must impose restrictions on the group actions involved, and this is done by using coordinates from linear spaces over finite fields. Another approach involves introducing the notion of "straight line" and imposing conditions (axioms) which make the geometries "projective" or "affine" in a certain sense.

Unfortunately, the two approaches are not equivalent, the axiomatic approach yielding a wider class of finite planes than the algebraic coordinate one. However, it turns out that the two approaches *are* equivalent if and only if Desargues' theorem holds in the finite geometry considered.

It should be noted that some basic natural questions about finite geometries are at present unanswered and that these geometries are the object of active ongoing research. Some of these questions and related conjectures are mentioned in Section 14.11.

14.1. Small finite geometries

In this section, we try to classify all the geometries with a "small" number of points. By classifying we mean listing (without repetitions) all the geometries with a given number of points $k := |X|$ up to isomorphism. Recall that two geometries are isomorphic if there is an equivariant bijection between them, i.e., a bijection between their sets of points and an isomorphism between their transformation groups which is compatible with the bijection (for the detailed definition, see Chapter 1).

There is of course only one geometry with one point. For $|X| = 2$ there are two geometries (with $|G| = 2$ and $|G| = 1$). For $|X| = 3$ there are four: the symmetries (= isometries) of the vertices of the equilateral triangle ($G = \mathbb{S}_3$), the motions of the vertices of the equilateral triangle ($G = \mathbb{Z}_3$), the symmetries of the vertices of the isosceles triangle ($G = \mathbb{Z}_2$). For $|X| = 4$ there are ten: the symmetries of the regular tetrahedron, its motions, the symmetries of the square, its motions, the rotations of the square by 0 and

π, the symmetries of the rhombus, and four more geometries obtained when the transformation group has a fixed point (the same one for each element).

For $|X| \geq 5$ the situation becomes too complicated to handle, while for $|X| \geq 10$, even a supercomputer is powerless.

To continue our study, we need to specify some reasonable classes of finite geometries. To do that, we need some algebra.

14.2. Finite fields

The modern logical foundation of ordinary Euclidean affine geometry is the notion of vector space over the real number field. To construct something similar in the finite case, we need *finite fields*.

14.2.1. Theorem. *For any $q = p^m$, where p is prime and m is a positive integer, there exists exactly one (up to isomorphism) field consisting of q elements, called the finite field of order q and denoted by $F(q)$. There are no other finite fields.*

We will not prove this theorem (the proof belongs to algebra courses) and only present the simplest nontrivial example $F(4) = \{\mathbf{0}, \mathbf{1}, \mathbf{2}, \mathbf{3}\}$ by displaying its addition and multiplication tables:

+	**0**	**1**	**2**	**3**
0	**0**	**1**	**2**	**3**
1	**1**	**2**	**3**	**0**
2	**2**	**3**	**0**	**1**
3	**3**	**0**	**1**	**2**

×	**0**	**1**	**2**	**3**
0	**0**	**0**	**0**	**0**
1	**0**	**1**	**2**	**3**
2	**0**	**2**	**3**	**1**
3	**0**	**3**	**1**	**2**

In order to get a feeling for the structure of the fields $F(q)$, we invite the reader to construct the addition and multiplication tables for, say, $F(3^2)$.

14.3. Example: the finite affine plane of over $F(5)$

In this section we will construct a finite affine plane geometry starting from the finite field $F(p)$, where p is a prime number (i.e., in the case $m = 1$). To make the construction more concrete, we will carry it out for $p = 5$, although it works for any prime p.

14.3.1. Let us define the *affine plane* $AF(5)$ of order 5 as the set $\{(x, y) | x \in F(5), y \in F(5)\}$ of pairs (coordinates of *points*). As in ordinary Euclidean geometry, two points $T = (a, b)$, $S = (c, d)$ determine a *vector*

$\overrightarrow{TS} = \{c-a, d-b\}$. We will define *straight lines* as in analytic geometry, i.e., by setting $A(t) = A_0 + t\overrightarrow{v}$, where A_0 is a point, $\overrightarrow{v}$ is a vector, and t runs over $F(5)$. For example, if we take $A_0 = (0,0)$ and $\overrightarrow{v} = (1,2)$, we obtain the "straight line"

$$\big\{(0,0),(1,2),(2,4),(3,1),(4,3)\big\}.$$

Thus we obtain a total of 30 straight lines, 25 points, 5 points on each line, and 6 lines passing through each point. In Figure 14.1, we have shown the six lines passing through the point $(0,0)$.

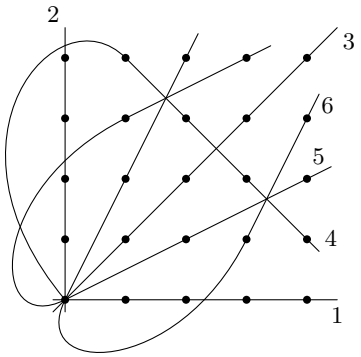


Figure 14.1. Six lines in $AF(5)$

Arguing in the same way in the general case, we obtain $p^2 + p$ straight lines, p^2 points, p points on each line, and $p+1$ lines passing through each point.

14.3.2. The same result can be obtained by using the orbit space of an appropriate geometry. Let $\mathbb{Z} \oplus \mathbb{Z} \subset \mathbb{R}^2$ be the integer lattice on the plane and let $(\mathbb{Z} \oplus \mathbb{Z} : G)$ be the geometry defined by the transformation group G, isomorphic to $\mathbb{Z} \oplus \mathbb{Z}$, acting by coordinate shifts by 5, i.e.,

$$G \ni (k,l) : (m,n) \mapsto (m+5k, n+5l).$$

The orbit space of this action consists of 25 "points". We identify them with the 25 points of the lattice with nonnegative coordinates less than 5. The "straight line" passing through two points of this 5 by 5 square are defined as follows: construct the Euclidean line joining these two points in $\mathbb{R}^2$, take all the integer points on this line and reduce both their coordinates mod 5, obtaining three more points in the square; together with the two given points, they constitute a " straight line".

Geometrically, you can visualize this as the covering of the torus by the plane: under this map the points of the square lattice are "wrapped around" the 25 points on the torus.

14.4. Example: the finite affine plane over $F(2^2)$

We now start our constructions with the field $F(2^2)$. Define the *affine plane* over $F(4)$ as the set $\{(x,y)|x \in F, y \in F\}$ of pairs (coordinates of *points*). Using the same approach as in Section 14.3 (including the "vector definition" of straight lines), we obtain, for example, the "straight line" consisting of the points $\big\{(0,0),(1,1),(2,2),(3,3)\big\}$. Now consider the line passing through $(0,0)$ along the vector $\{1,2\}$; it consists of the points

$$\big\{(0,0),(1,2),(2,0),(3,2)\big\}.$$

But there is another line passing through the two points $(0,0)$, $(2,0)$, namely the "horizontal" line

$$\big\{(0,0),(1,0),(2,0),(3,0)\big\}.$$

Thus the fundamental fact that "through two points there passes one and only one straight line" does not hold in the "affine geometry" with straight lines defined as above!

Nevertheless, a reasonable affine geometry with 4 points on each line can be constructed on the set of points P by defining straight lines in a different way. In particular, the "straight line" that passes through the points $(0,0)$, $(1,2)$ must contain two more points ($(2,3)$ and $(3,1)$) and is unique. In this geometry, there are $16 = q^2$ points, $20 = q^2 + q$ straight lines, $4 = q$ points on each line, and $5 = q + 1$ lines pass through each point. The five lines passing through the point $(0,0)$ are shown in Figure 14.2.

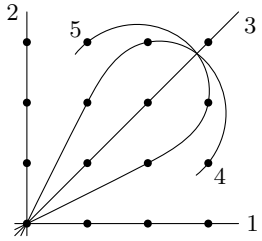


Figure 14.2. Five lines in $AF(4)$

The result is a geometry called the *finite affine plane* over the field $F(4)$ and is denoted by $AF(4)$. The set $AF(4)$ is indeed a geometry in the sense of Klein, $(AF(4) : \Gamma)$, if for the transformation group Γ we take the set of all bijections of $AF(4)$ that map lines into lines.

In the general case, i.e., when $F = F(q)$, $q = p^m$, $m > 1$, with prime p, one can also construct the finite affine plane $AF(q)$, but the direct construction is rather tedious, and we omit it. However, we will present a neat indirect construction via finite projective geometries in Section 14.8.

First, we give an example of a finite projective geometry.

14.5. Example of a finite projective plane

14.5.1. Let $AF(4)$ be the finite affine plane for $q = 2^2$. We say that two lines of $AF(4)$ are *parallel* if they coincide or have no common points. Parallelism is an equivalence relation, and so all lines are partitioned into equivalence classes of parallel lines. It is easy to see that there are 5 such classes. To $AF(4)$ let us add 5 points (called *points at infinity*) and agree that they all lie on one straight line (the *line at infinity*). The set thus obtained is called the *projectivization* of the affine plane $AF(4)$ and is denoted by $PF(4)$; it has 21 points, 21 straight lines, 5 points on each line, 5 lines passing through each point, and any two distinct lines have exactly one common point. The projective plane $PF(4)$ is shown in Figure 14.3.

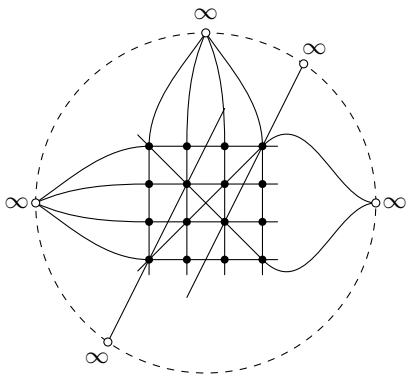


Figure 14.3. Projectivization of $AF(4)$

14.5.2. The construction described above for $q = 4$ actually works for any $q = p^m$ with prime p. One obtains the projective geometry $PF(q)$; it has $q^2 + q + 1$ points, $q^2 + q + 1$ straight lines, $q + 1$ points on each line, $q + 1$

lines passing through each point, and any two distinct lines of $PF(q)$ have exactly one common point.

14.6. Axioms for finite affine planes

14.6.1. A more traditional approach to finite geometries is the axiomatic approach. A finite affine plane is a nonempty finite set of elements P (called *points*) with a family L of subsets (called *lines*) that satisfy the axioms:

Aff.1. *There is exactly one line passing through any two distinct points.*

Aff.2. *There is exactly one line parallel to a given line and passing through a given point.* (Two lines are called *parallel* if they have no common points or if they coincide.)

Aff.3. *There is a generic triangle* (*three points not belonging to one and the same line*).

Here the second axiom ensures that the dimension of the set of points is less than or equal to 2. The third axiom ensures that its dimension is greater than or equal to 2. Thus the dimension of the set of points is two, this set can be regarded as a "plane". The construction of the two simplest affine planes (with 4 and 9 points) is the object of Exercise 14.1.

14.3.2. Theorem. (i) *For every $q = p^m$, where p is prime and m is a positive integer, there exists an affine geometry $\mathsf{P} = AF(q)$ with q points on a line.*

(ii) *The geometry $\mathsf{P} = AF(q)$ has q^2 points, a family of $q^2 + q$ subsets L that satisfies the axioms* Aff.1–Aff.3.

(iii) *If Γ_q is the group of bijections of P that map lines* (*i.e., elements of L*) *into lines, then (P, Γ_q) is a geometry in the sense of Klein called an affine Galois plane of order q.*

The existence of $AF(q)$ (item (i) of the theorem) will be proved in 14.8.3. The proof of items (ii)-(iii) is a series of exercises (14.2-14.6) in the problem section.

14.7. Axioms for finite projective planes

14.7.1. A finite projective plane is a nonempty finite set of elements P (called *points*) with a family L of subsets (called it lines) that satisfy the following axioms:

Proj.1. *There is exactly one line passing through a pair of distinct points.*

Proj.2. *There is exactly one point contained in a pair of distinct lines.*

Proj.3. *There exist four points that determine six distinct lines.*

Proj.4. *There exist four lines that determine six distinct points.*

Actually the fourth axiom is redundant (it follows from the first three), we include it for the sake of symmetry.

The simplest finite projective plane (called the *Fano plane*) is shown in Figure 14.4. It has 7 points, 7 lines, 3 points on each line, and 3 lines passing through each point. The four points in the middle of the picture satisfy the axiom Proj.3. The Fano plane can be constructed from the four point affine plane by adding the "line at infinity", as explained in 14.5.1.

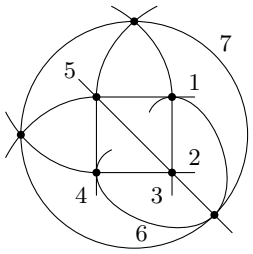


Figure 14.4. The Fano plane

14.7.3. *Projective duality.* Just as in the case of the real projective plane $\mathbb{R}P^2$, the finite projective plane satisfies the *Duality Principle: Interchanging the words "point" and "line" in the statement of any theorem and accordingly modifying the wording of the incidence relations, one obtains another theorem.* This principle follows from the fact that the four axioms split into two pairs dual to each other. However, the finite projective projective plane obtained from a given one by duality is not necessarily isomorphic to the given one. Questions of duality are rather delicate in the finite case, and we do not discuss them here.

14.7.3. Theorem. *If (P,L) is a finite projective plane, then there exists a natural number n, called the order of the plane, such that:*

(i) *each line contains $n+1$ points;*

(ii) *each point is contained in $n+1$ lines;*

(iii) *the number of points is equal to the number of lines and equal to n^2+n+1.*

14.7.4. Remark. The theorem does not assert the existence of finite projective planes: in it, it is *assumed* that a finite projective plane is given

and thus it only asserts that if a plane satisfying axioms Proj.1–Proj.3 exists, then its number of lines and points satisfies the constraints (i)-(iii).

14.7.5. *Proof of Theorem* 14.7.3. Suppose (P,L) is a finite projective plane, $l, m \in \mathsf{L}$, and let $a \in \mathsf{P}$ be a point not lying on l nor on m (such a point exists by axiom Proj.3). Consider the map f of the set of points of the line l to the set of points of m that assigns to each point $x \in l$ the intersection point of the lines xa and m. Axioms Proj.1–Proj.2 imply that f is well defined and bijective. Denoting the number of points on l by $n+1$, we see that item (i) is proved. Item (ii) follows by the Duality Principle. To prove (iii), fix some point $a \in \mathsf{P}$. Each line passing through a passes through n other points, and so $|\mathsf{P}| = (n+1)n + 1 = n^2 + n + 1$. By duality we have $|\mathsf{L}| = n^2 + n + 1.$, which concludes the proof. □

14.7.6. Remarks. (1) One can pass from the finite affine plane to the projective plane by adding $q+1$ "points at infinity" (corresponding to each class of parallel lines) and one new line (the line of all points at infinity). Conversely, one can pass from a projective plane to an affine plane by removing one line (with all its points). Unfortunately, the result is not well defined: it may depend on the choice of the line!

(2) There is no uniqueness theorem for projective planes of order p^m for $m > 1$ (for example, there is are several nonisomorphic projective planes of order 9, see Exercise 14.9).

(3) It is not known at present for what values of q there exist projective planes of order q. Specifically, this question is unanswered already for $q = 12$. This question, and other open questions, as well as related conjectures, are briefly discussed in Section 14.11.

14.8. Constructing projective planes over finite fields

In this section, we give a constructive definition of the finite projective planes based on linear spaces over finite fields, similar to the definition of the real projective plane $\mathbb{R}P^2$ (*cf.* 12.1).

14.8.1. *Main construction.* Consider the three-dimensional vector space V over the finite field $F = F(p^m)$, where p is prime. Denote by P the set of one-dimensional subspaces of V, which we now call *points*, and the set L of two-dimensional subspaces, which we now call *lines*; we say that a line $l \in \mathsf{L}$ *passes through* a point $p \in \mathsf{P}$ (or p is contained in l, or l *contains* p) if we have the inclusion of linear spaces $p \subset l$.

14.8.2. Theorem. (i) *The construction described above yields a finite projective plane* (P, L) *of order* $q = p^m$.

(ii) *If we define the transformation group of* P *as the set of bijections* Γ *of* P *that take lines to lines, then* (P, Γ) *is a geometry in the sense of Klein.*

Proof. All four axioms Proj.1–Proj.4 are immediate consequences of the main construction. Item (ii) is the object of Exercise 14.6 .□

The geometry thus constructed is called the *finite projective space* over the field $F(p^m)$and is denoted by $PF(p^m)$.

14.8.3. Corollary. *The finite affine plane of order* $q = p^m$, *where* p *is any prime and* m *is any natural number, exists.*

Proof. To construct the required plane, it suffices to remove one line (and all its points) from the finite projective plane of order q. □

This fact completes the proof of item (i) of Theorem 14.3.2.

14.9. The Desargues theorem

The Desargues theorem, which we proved for the real projective plane $\mathbb{R}P^2$, is not true for arbitrary finite projective planes. However, we have the following statement.

14.9.1. Theorem. *The Desargues theorem holds for the finite projective planes* $PF(p^m) = (\mathsf{P}, \mathsf{L})$, *i.e., three lines* $x_1y_1, x_2y_2, x_3y_3 \in \mathsf{L}$ *intersect at one point if and only if the intersection points* $z_1, z_2, z_3 \in \mathsf{P}$ *of the pairs of lines* x_2x_3 *and* y_2y_3, x_3x_1 *and* y_3y_1, x_1x_2 *and* y_1y_2, *respectively, are collinear.*

Proof. In the proof, we will use the model of $PF(p^m)$ given by the construction 14.8.1, i.e., we regard points as one-dimensional linear subspaces of the vector space over $F(p^m)$ and lines as two-dimensional subspaces.

First let us note that the Desargues theorem is self-dual, and therefore it suffices to prove the "only if" part, i.e., assuming that the lines A_1B_1, A_2B_2, A_3B_3 intersect at one point (which we denote by S), to show that the intersection points P_1, P_2, P_3 are collinear. If the point S lies in each of the three lines P_1P_2, P_2P_3, P_3P_1, then there is nothing to prove, so we can assume that $S \notin P_2P_3$.

In our model the points S, A_i, B_j, P_k are actually one-dimensional linear spaces, and we will use the same lower case letters s, a_i, b_j, p_k to denote nonzero vectors belonging to (and therefore determining) the corresponding linear spaces.

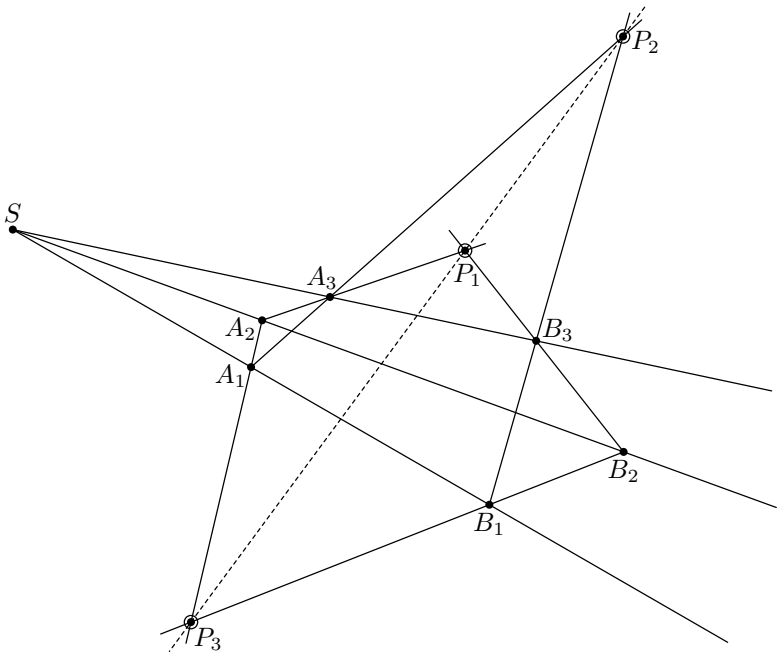


Figure 14.5. Desargues' theorem

Now since the vectors s, a_1, b_1 belong to the same two-dimensional space, they are linearly dependent, and (by an appropriate choice of these vectors in their linear spaces) we can write $b_1 = a_1 + s$. It is easy to see that the vectors a_1, p_2, p_3 are linearly independent, and therefore we can put

$$b_1 = \alpha_1 a_1 + \alpha_2 p_2 + \alpha_3 p_3,$$

where $\alpha_1, \alpha_2, \alpha_3 \in F$. Consider the linear operator φ on V given by

$$\varphi(a_1) = b_1, \varphi(p_2) = p_2, \varphi(p_3) = p_3.$$

Then we have

$$\varphi(s) = \varphi(b_1 - a_1) = (\alpha_1 - 1)b_1 + \alpha_2 p_2 + \alpha_3 p_3 = (\alpha_1 - 1)b_1 + b_1 - \alpha_1 a_1 = \alpha_1 s.$$

The linear operator φ is nondegenerate, it takes linear subspaces to linear subspaces of the same dimension. In particular, we have

$$\varphi(A_1) = B_1,\ \varphi(P_2) = P_2,\ \varphi(P_3) = P_3,\ \varphi(S) = S.$$

The vectors p_2, p_3 form a basis of the line P_2P_3 (regarded as a two-dimensional vector space), and so the operator φ is the identity on this line. Now if Λ is any line passing through S, then, since φ leaves S in place as well as the intersection point of the lines Λ and P_2P_3, it follows that $\varphi(\Lambda) = \Lambda$.

Now the point A_2 lies on the lines SA_2 and B_1P_3, and therefore $\varphi(A_2)$ is the intersection point of the lines SA_2 and B_1P_3, and so $\varphi(A_2) = B_2$. Similarly, $\varphi(A_3) = B_3$. Thus $\varphi(A_2A_3) = B_2B_3$. Now let P be the intersection point of the lines A_2A_3 and P_2P_3. Then the point $\varphi(P)$ lies on the line B_2B_3 and, at the same time, $\varphi(P) = P$. Therefore, $P = P_1$ and P_1 lies on the line P_2P_3, which was to be proved. □

14.9.2. *Remark.* Note that this proof (like the proof given in 12.7.1) is, in a certain sense, "three-dimensional": when we replaced points by vectors in the above proof, we were essentially adding a point (the origin of coordinates in the three-dimensional space over $F(p^m)$) lying outside the plane containing all the given points.

14.10. Algebraic structures in finite projective planes

Until now, we have been using algebra (finite fields) to construct geometric objects (finite affine and projective planes). Now we will try to move in the opposite direction, i.e., analyze what the geometric axioms for the finite projective plane imply concerning the algebraic structure of the projective line. Unfortunately, it will turn out that the natural and optimistic expectation that axioms Proj.1–Proj.4 imply that there are $p^m + 1$ points on each line (for some prime p and natural number m) and that these points can be added and multiplied in a natural way, thereby forming a field isomorphic to $F(p^m)$, does not come true. The situation is much more complicated, in the general case one can obtain an algebraic structure from the axioms, but is not that of a field: its multiplication is not commutative and there only one distributive law (see 14.10.3 below)

14.10.1. *Introducing coordinates.* Let (P, L) be a finite projective plane of order $n \geq 2$. (Recall that this means that (P, L) satisfies axioms Proj.1–Proj.4) and one of its lines (and therefore all lines) contains n points). Denote by F a set of n elements; we stress that F is a set of arbitrary symbols, it is not a field, in fact at first it has no algebraic operations defined on it. Our aim is to supply F with an algebraic structure (hopefully that of a field) and use it to introduce coordinates in our finite projective plane (P, L).

We begin by choosing two arbitrary elements of F that we denote by 0 and 1. By ∞ we denote a symbol that does not belong to F. Using axiom

Proj.3, let us choose an *initial quadrilateral* in our plane, i.e., four points no three of which lie on one line. Denote these points by $(0,0), (0), (\infty), (1,1)$ and denote the six lines passing through these points as follows

$$[0,0] : = (0,0)(0), \qquad [0] := (0,0)(\infty),\ [\infty] := (0)(\infty),$$
$$[1] : = (1,1)(\infty),\ [0,1] := (1,1)(0),\ [1,0] := (0,0)(1,1).$$

These six lines intersect in seven points (four of which belong to the initial quadrilateral), and we denote the other three as follows

$$(1,0) := [1][0,0],\ (0,1) := [0][0,1],\ (1) := [\infty][1,0],$$

where the juxtaposition of two lines determines their intersection, e.g. the formula $(1,0) = [1][0,0]$ means that $(0,1)$ is the intersection point of the lines $[1]$ and $[0,0]$.

If there are no other points in P, then $n = 2$ and it is easy to see that we have obtained the Fano plane. The reader will profit by looking at Fig.14.4 and supplying its points with coordinates as indicated in the construction described above.

If there are other points left, then $n > 2$ and we denote by a an arbitrary element of F other than 0 or 1. For any such a, we define new points and lines by setting

[a,0]:=(0,0)(a), (1,a):=[1][a,0], [0,a]:=(0)(1,a),
(a,a):=[0,a][1,0], [a]:=(a,a)(∞), (a,0):=[a][0,0], (0,a):=[0,a][0].

If there are any other elements b in F other than $0, 1, a$, we set

$$(a,b) := [a][0,b], \quad [a,b] := (a)(0,b).$$

Thus we have supplied all the points of our finite projective plane with coordinates, and we know what the intersection points of any two lines are.

14.10.2. *Addition and multiplication.* Now we can define the sum and product of two arbitrary elements $a, b \in F$ by setting

$$(a, a+b) := [a][1,b], \quad (a, a\cdot b) := [a][b,0].$$

The motivation behind this definition is that it is compatible with the addition and multiplication induced on points on the projective line in the case of the finite projective plane over the field $F(p^m)$. The reader is invited to return to the definition of finite projective planes over a field, check that they

can be supplied with coordinates as specified above and that the operations defined above coincide with the ones induced by the field $F(p^m)$.

As we noted above, it is not always true that these operations supply F with a field structure. They satisfy axioms of a structure weaker than that of a field, which we now define.

14.10.3. *Almost fields.* An *almost field* is a set F with two binary operations, called *addition* and *multiplication*, such that under addition F is an Abelian group with neutral element 0, the set $F \setminus 0$ is a group (not necessarily Abelian) under multiplication and the right distributive law is satisfied, i.e., $(a+b)c = ac + bc$.

When the left distributive law is not satisfied (such examples of almost fields exist), the almost field is not even a ring. We will not describe examples of this type or study almost fields in detail: they are complicated and rather ugly, and we will limit our exposition to the statements (without proofs) of two beautiful theorems and of some open problems.

14.10.4. Theorem. (i) *Given any finite almost field F, a projective plane over F can be determined by using the construction from Section* 14.8 *with F replacing the field $F(p^m)$.*

(ii) *Given any finite projective plane of order n, there is an almost field F (of order $n-1$) using which the projective plane can be constructed as indicated in* (i).

The proof of (i) is similar to that in Section 14.8, while (ii) can be proved by a tedious series of geometric constructions needed to verify the numerous axioms of almost fields.

14.10.5. Theorem. *A finite projective plane is a projective plane over the field $F(p^m)$ if and only Desargue's theorem holds in it.*

The "only if" part was proved above (see 14.8.1), while the "if" part is another complicated series of artificial geometric constructions ensuring the required algebraic axioms.

14.11. Open problems and conjectures

The main open problem here is the following: *For what values of q does there exist a finite projective plane of order q and for what values of q is the finite projective plane of order q unique?*

We know that there exists one and only one projective plane of the orders 2,3,4,5,7,8 (see Exercises 14.10–14.11). We also know certain number-theoretic constraints forbidding projective planes of certain orders.

14.11.3. Theorem. [Brack–Raiser] *Let $q \equiv 1$ or 2 (mod 4). If there exists a projective plane of order q, then q can be presented as the sum of squares of two natural numbers.*

For the proof, see [??]. This theorem forbids projective planes of orders 6, 14,21, 22, 30, etc.

14.11.3. Conjecture. *The order q of any finite projective plane is a prime number $q = p$ or a power of a prime $q = p^m$.*

The first natural number q which does not meet the assumptions of the conjecture is 6, and indeed on can prove (see Exercise 14.9) that there is no finite projective plane of order 6. The next such number is 10, and it is only in 1991 that it was established, with the aid of a supercomputer, that the conjecture holds there also. But already for $q = 12$ the existence of a projective plane of order q is an open question.

14.11.3. Conjecture. *All the projective planes of prime order p are Desargues.*

There are non-Desarguian projective planes of nonprime order. The "smallest" one is of order 9 (Exercise 14.15).

14.12. Problems

14.1. Construct an affine geometry having 4 points and a finite affine geometry having 9 points.

14.2. Suppose that one of the lines of the affine plane (P, L) from Theorem 14.1 consists q points. Prove that the plane P consists of q^2 points.

14.3. Suppose that one of the lines of the affine plane (P, L) from Theorem 14.1 consists q points. Prove that all other lines consist of q points.

14.4. Suppose that one of the lines of the affine plane (P, L) from Theorem 14.1 consists q points. Prove that L consists of $q^2 = q$ lines.

14.5. Suppose that one of the lines of the affine plane (P, L) consists q points. Prove that $q + 1$ lines pass through each point.

14.6. Prove that the finite affine plane $AF(p^m)$ is a geometry in the sense of Klein.

14.7. In the affine plane consisting of q^2 points for $q = 3$, construct the system of lines passing through one of the points.

14.8. Describe the projectivization of the affine plane from Exercise 14.5.

14.9*. Prove that there does not exist a finite projective plane of order $q = 6$.

14.10. Prove that the projective planes of order 2,3,4,5 are unique.

14.11*. Prove that the projective planes of order 7 and 8 are unique.

14.12*. Does there exist a finite affine plane of order $q = 6$?

14.13*. Find two nonisomorphic finite affine planes of order $q = 9$.

14.14. By adding "points at infinity" to the affine geometries of orders 3,4,5, construct the corresponding finite projective planes.

14.14.** Give an example of a finite projective plane from which one can obtain nonisomorphic affine planes by removing one line.

14.15*. Construct a non-Desarguian projective plane of order 9.