Lectures on Groups and Their Connections to Geometry

Anatole Katok

Vaughn Climenhaga

Department of Mathematics, Pennsylvania State University,
University Park, Pennsylvania 16802
E-mail address: katok_a@math.psu.edu

Department of Mathematics, Pennsylvania State University,
University Park, Pennsylvania 16802
E-mail address: climenha@math.psu.edu

Contents

CHAPTER 1

Elements of group theory

Lecture 0. Wednesday, Thursday, and Friday, August 26–28

a. Binary operations. We learn very early on that numbers are more than just static symbols which allow us to quantify one thing or another. Indeed, they can be added, subtracted, multiplied, and (usually) divided, and this toolbox of arithmetic operations is largely responsible for the phenomenal variety of uses we are able to make of the concept of "number".

The study of arithmetic is the study of the behaviour of these tools and of the numbers with which they interact. Given the utility of these arithmetic tools, one may reasonably ask if their usefulness is inextricably bound up in their domain of origin, or if we might profitably apply them to something besides numbers (although just *what* else we ought to use in place of numbers is not obvious).

Fortunately for our purposes (otherwise this course would not exist), the latter alternative turns out to be the correct one. As so often happens in mathematics, the arithmetic tools of addition, multiplication, etc. can be abstracted and generalised for use in a much broader setting; this generalisation takes us from *arithmetic* to *algebra*, and leads us eventually into an incredibly rich field of mathematical treasures.

We begin our story where so many mathematical stories begin, with a set X, which we then equip with some structure. In this case, the key structure we will place on X is a *binary operation*—that is, a function which takes two (possibly equal) elements of X as its input, and returns another (possibly equal) element as its output.

Formally, a binary operation is a map from the direct product $X \times X$ to the original set X. Rather than using the functional notation $f\colon X \times X \to X$ for this map, so that the output associated to the inputs a and b is written as $f(a,b)$, it is standard to represent the binary operation by some symbol, such as $\star$, which is written between the two elements on which it acts—thus in place of $f(a,b)$ we write $a \star b$.

This notation becomes intuitive if we consider the familiar examples from arithmetic. If X is some class of numbers (such as $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{R}$, or $\mathbb{C}$), then the usual arithmetic operators are binary operations; for example, addition defines a map $f_+\colon X \times X \to X$ by $f_+(a,b) = a + b$, subtraction defines a map $f_-(a,b) = a - b$, and so on.

Exercise 0.1. In fact, subtraction and division only define binary operations on some of the classes of numbers listed above. On which sets do they fail to be binary operations, and why?

When we need to indicate which binary operation a set is equipped with, we shall do so by writing the two as an ordered pair. Thus $(\mathbb{N}, +)$ denotes the natural numbers equipped with addition, while $(\mathbb{R}, \cdot)$ denotes the real numbers with multiplication.

Since we tend to think of arbitrary binary operations as generalisations of either addition or multiplication, it is common to refer to $a \star b$ as either the "sum" or the "product" of a and b, even when $(X, \star)$ is arbitrary.

Example 0.1. Although addition and multiplication are the two primary examples, we can define many other binary operations on the above sets of numbers, not all of which are particularly interesting:

(1) $a \star b = ab + 2$
(2) $a \star b = a + b - 5$
(3) $a \star b = a - 3$
(4) $a \star b = 4$

Example 0.2. As suggested in the introduction, we can define binary operations on sets which are not sets of numbers. For example, let X be an arbitrary set, and consider the *power set* $P(X)$, which is the set of all subsets of X. We may define a binary operation on $P(X)$ by taking intersections: $S_1 \star S_2 = S_1 \cap S_2$, where $S_1, S_2 \subset X$. Alternately, we may define $\star$ by taking unions, and consider $(P(X), \cup)$. Other set operations lead to similar constructions.

Example 0.3. Addition of vectors in $\mathbb{R}^n$ is a binary operation. In the particular case $n = 3$, another binary operation is given by the cross product $\mathbf{u} \times \mathbf{v}$. Note that the dot product $\mathbf{u} \cdot \mathbf{v}$ is *not* a binary operation, since it returns a scalar (an element of $\mathbb{R}$), rather than a vector in $\mathbb{R}^3$.

Example 0.4. Given a set X, let $F(X)$ be the set of all maps from X to itself; then composition defines a binary operation on $F(X)$, with $(f \circ g)(x) = f(g(x))$ for $f, g \in F(X)$. This example will eventually turn out to be of much greater importance than may initially be apparent.

Example 0.5. Let $M(n, \mathbb{R})$ be the set of $n \times n$ matrices with real entries; matrix multiplication defines a binary operation on $M(n, \mathbb{R})$.

b. Monoids, semigroups, and groups. As given so far, the concept of binary operation is really too general to be of much use (as one may perhaps divine from Example 0.1). Given $a, b, c \in X$, the definition of binary operation tells us that $a \star b$ and $b \star c$ are again elements of X, and so we may consider the two elements $(a \star b) \star c$ and $a \star (b \star c)$. If these two elements differ, then it is not clear which one is to be thought of as the "product" (to use the multiplicative terminology) of the three elements a, b, and c. This motivates the following definition.

Definition 0.6. A binary operation $\star$ is *associative* on a set X if $(a \star b) \star c = a \star (b \star c)$ for every $a, b, c \in X$.

Example 0.7.

(1) The usual notions of addition and multiplication are associative, while subtraction and division are not.
(2) The set operations $\cup$ and $\cap$ are associative on $P(X)$.
(3) Vector addition is associative, while the cross product is not: $(\mathbf{u} \times \mathbf{u}) \times \mathbf{v} = \mathbf{0}$, while $\mathbf{u} \times (\mathbf{u} \times \mathbf{v}) \neq \mathbf{0}$ provided $\mathbf{u}$ and $\mathbf{v}$ are nonzero.

Exercise 0.2.

(a) Show that composition of functions is associative.
(b) Show that matrix multiplication is associative.

Associativity says that we may simply write the product of three elements as $a \star b \star c$, as the resulting element is the same whether we bracket the product as $(a \star b) \star c$ or $a \star (b \star c)$.

Exercise 0.3. Let $\star$ be associative on X, and show that given any elements $a_1, \dots, a_n \in X$, the product $a_1 \star \cdots \star a_n$ is the same no matter where one puts the brackets.

Definition 0.8. A set X together with an associative binary operation $\star$ is called a *monoid*.

Remark. We must emphasise that apart from associativity, no other hypotheses whatsoever are placed on $\star$. In particular, we do not assume that $\star$ is commutative—that is, that $a \star b = b \star a$—one may easily check that although this property holds for the familiar cases of multiplication and addition of numbers, it fails for composition of functions and for matrix multiplication.

It turns out that monoids are too general to be as useful as we would like. We can remedy the situation somewhat by adding one more requirement.

Definition 0.9. Given a monoid $(X, \star)$, an *identity element* is an element $e \in X$ such that $e \star a = a = a \star e$ for every $a \in X$. A monoid which possesses an identity element is known as a *semigroup*.

Example 0.10.

(1) In an additive number semigroup, such as $(\mathbb{C}, +)$, the identity element is 0.
(2) In a multiplicative number semigroup, such as $(\mathbb{Z}, \cdot)$, the identity element is 1.
(3) $(P(X), \cup)$ is a semigroup with identity element $\emptyset$; $(P(X), \cap)$ is a semigroup with identity element X.
(4) $(F(X), \circ)$ is a semigroup whose identity element is the identity map.
(5) $(M(n, \mathbb{R}), \cdot)$ is a semigroup whose identity element is the $n \times n$ identity matrix.

Remark. The similarity in form between the last two statements in Example 0.10 is no coincidence. Although $n \times n$ matrices can be viewed simply as square arrays of numbers, and matrix multiplication as merely an arbitrary rule, a mathematical version of "Simon says", it is much more enlightening to consider elements of $M(n, \mathbb{R})$ as linear maps from $\mathbb{R}^n$ to itself. Then matrix multiplication corresponds to composition of linear maps, which gives an easy solution of Exercise 0.2—certainly much simpler than trying to verify associativity using the formula for multiplying three matrices together!

This is an early example of the connection between algebra and geometry—here a geometric interpretation (matrices are linear maps) led to a much simpler proof of an algebraic fact (matrix multiplication is associative). We will explore many such connections in this course.

We will be interested in semigroups which satisfy one additional requirement.

Definition 0.11. Let $(X, \star)$ be a semigroup. An element $b \in X$ is the *inverse* of $a \in X$ if $a \star b = e = b \star a$. A semigroup in which every element possesses an inverse is a *group*.

Thus a group is a set equipped with an associative binary operator and an identity element, in which every element has an inverse.

Exercise 0.4. One might ask if both equalities in Definition 0.11 are strictly necessary—that is, once we know that $a \star b = e$, does it follow that $b \star a = e$ as well? Show that this is not necessarily the case, and so we do need to specify that b is a *two-sided* inverse.

Example 0.12.

(1) $(\mathbb{Z}, +)$, $(\mathbb{R}, +)$, and $(\mathbb{C}, +)$ are groups: the inverse of a is $-a$. $(\mathbb{N}, +)$ is not a group, as no negative numbers are included, and hence no positive numbers have inverses.

(2) $(\mathbb{R}^+, \cdot)$ and $(\mathbb{C} \setminus \{0\}, \cdot)$ are groups: the inverse of a is $1/a$. $(\mathbb{R}, \cdot)$ and $(\mathbb{C}, \cdot)$ are not groups, as 0 has no inverse; similarly, $(\mathbb{Z}, \cdot)$ and $(\mathbb{N}, \cdot)$ are not groups.

(3) Neither $(P(X), \cup)$ nor $(P(X), \cap)$ are groups; indeed, they do not have *any* invertible elements, in contrast to the previous examples of $\mathbb{R}$ and $\mathbb{C}$.

(4) $(F(X), \circ)$ is not a group if X has more than one element (consider the map $f \colon X \to X$ which maps everything to one element). However, we may obtain a group by restricting our attention to the set $S(X)$ of *invertible* maps in $F(X)$—that is, the set of bijections (one-to-one and onto maps) from X to itself.

(5) Similarly, $M(n, \mathbb{R})$ is not a group for $n \geq 1$, as matrices with zero determinant are non-invertible. By removing such matrices, we obtain the *general linear group*

$$GL(n, \mathbb{R}) = \{A \in M(n, \mathbb{R}) \mid \det A \neq 0\},$$

which will be central to many of our later investigations.

The notation we use will usually reflect one of the first two examples above. In additive notation, the inverse of a is written $-a$; in multiplicative notation, we write it as a^{-1}.

c. Multiplication tables and basic properties. Take the set of all integers, and partition it into two subsets, one containing all the even integers, the other containing all the odds. Denote the former by E and the latter by O. Then the set $X = \{E, O\}$ inherits the binary operations of addition and multiplication from the integers; the sum of two even numbers is even, thus we write $E + E = E$, and so on. All the possible combinations are summed up in the following tables:

$\cdot$	E	O
E	E	E
O	E	O

$+$	E	O
E	E	O
O	O	E

Such a table will be referred to as a *multiplication table* (even though the binary operation in question may be addition), and can be constructed for any binary operation on a finite set. From the table we can read off the value of $a \star b$ by looking at the entry in the row corresponding to a and the column corresponding to b, and thus the table gives us complete information about the binary operation. For example, if we write $\mathbb{Z}/3\mathbb{Z} = \{0, 1, 2\}$ for the set of possible remainders upon division by 3, and take as our binary operation either addition or multiplication modulo 3, we obtain the following:

$\cdot$	0	1	2
0	0	0	0
1	0	1	2
2	0	2	1

$+$	0	1	2
0	0	1	2
1	1	2	0
2	2	0	1

In general, there is no easy way to see from a multiplication table whether a given binary operation is associative, short of checking every possible combination (which can become remarkably tedious). However, once we have verified associativity (through tedious means or otherwise), the other group axioms can be checked relatively easily.

A *left identity element* is $e_\ell \in X$ such that $e_\ell \star a = a$ for every $a \in X$; the various elements $e_\ell \star a$ are the entries in the row of the table labeled with e_ℓ, and so the statement that e_ℓ is a left identity amounts to the statement that the corresponding row merely repeats the entries in the top row of the matrix (the column labels). Thus we see from the above tables that:

(1) E is a left identity for $(\{E, O\}, +)$.
(2) O is a left identity for $(\{E, O\}, \cdot)$.
(3) 0 is a left identity for $(\mathbb{Z}/3\mathbb{Z}, +)$.
(4) 1 is a left identity for $(\mathbb{Z}/3\mathbb{Z}, \cdot)$.

Similarly, a *right identity element* is $e_r \in X$ such that $a \star e_r = a$ for every $a \in X$; this corresponds to a column in the matrix which repeats the row labels, and we see that for the four examples given, the left identities are also right identities, and hence true two-sided identity elements. Thus all of these examples are semigroups.

REMARK. In fact, to check that a monoid X is a semigroup, it suffices merely to check that X has both a left identity e_ℓ and a right identity e_r, for then we have $e_\ell = e_\ell e_r = e_r$ (using the definitions of left and right identities), and hence the two agree.

Checking for the existence of an inverse is also simple. Given an element $a \in X$, a *left inverse* for a is an element b such that $b \star a = e$; this corresponds to the occurrence of the identity e in the column headed by a. Thus every element of X has a left inverse if and only if the identity element appears in every column of the multiplication table.

Similarly, every element of X has a right inverse if and only if the identity element appears in every row of the multiplication table. Thus if X is a group, then the identity element appears in every row and column.

It is not *a priori* obvious that this necessary condition is also sufficient—after all, the definition of a group requires the left and right inverses of a to be the same. However, we may observe that if b and b' are left and right inverses for a, respectively, then $b \star a = a \star b' = e$, and hence $b = b \star (a \star b') = (b \star a) \star b' = b'$. It follows that the left and right inverse agree, and are equal to a^{-1}.

EXAMPLE 0.13.

(1) $(\{E, O\}, +)$ and $(\mathbb{Z}/3\mathbb{Z}, +)$ are groups.
(2) $(\{E, O\}, \cdot)$ is not a group, as E has no inverse.
(3) $(\mathbb{Z}/3\mathbb{Z}, \cdot)$ is not a group, as 0 has no inverse.

REMARK. Once we know that $(X, \star)$ is a group, we can deduce an even stronger result about its multiplication table. Not only does the identity element appear in each row and each column, but *every* element of X does. To see this, let $a, b \in X$ be arbitrary, and let $c = b \star a^{-1}$. Then $b = c \star a$, and thus b appears in the column headed by a. This shows that every element appears in every column, and a similar argument shows that every element appears in every row.

We can also make the dual observation that no element can appear twice in the same row or column (if X is finite, this is equivalent to the result in the previous paragraph, but for infinite X it is independent). To see this, consider the column headed by a, and let $b, c \in X$ be arbitrary. Then if $b \star a = c \star a$, we have $(b \star a) \star a^{-1} = (c \star a) \star a^{-1}$, and hence $b = c$. Since the entries in the column headed by a are all of the form $b \star a$, it follows that they are all distinct.

One further comment about inverses is in order. What is the inverse of $a \star b$? A naïve guess would be $a^{-1} \star b^{-1}$—the reader is invited to verify that this does not work out. Instead, one must reverse the order, and we find that $(a \star b)^{-1} = b^{-1} \star a^{-1}$. In fact, this is perfectly intuitive. Instead of group elements, think of a and b as standing for daily actions; let a stand for putting on socks, and b for putting on shoes. These two things are of course not commutative—putting on shoes before socks produces a most ineffective set of footwear. Their inverses are a^{-1} (taking off socks) and b^{-1} taking off shoes. Then common sense and daily experience tell us that in order to invert the element $a \star b$ (putting on socks, then shoes), we must first apply b^{-1} (remove shoes), and *then* apply a^{-1} (remove socks).

d. Groups and arithmetic. The two most familiar binary operations are addition and multiplication; we have already seen these applied to various classes of numbers (natural, integer, real, complex). One further example which bears mentioning is the integers taken modulo a fixed n. This arises when we fix $n \in \mathbb{N}$ and consider the equivalence relation on the integers which is given by congruence modulo n: $a \equiv b \bmod n$ if and only if n divides $b - a$. As with any equivalence relation, this induces a partition of the integers into *equivalence classes*, such that two integers are congruent modulo n if and only if they lie in the same equivalence classes.

The equivalence class of $a \in \mathbb{N}$ may be denoted

$$[a]_n = \{b \in \mathbb{Z} \mid a \equiv b \bmod n\};$$

there are exactly n equivalence classes, and the collection of these classes is denoted $\mathbb{Z}/n\mathbb{Z}$. We see that $\mathbb{Z} = [0]_n \cup [1]_n \cup \cdots \cup [n-1]_n$, and so

$$\mathbb{Z}/n\mathbb{Z} = \{[0]_n, [1]_n, \dots, [n-1]_n\}.$$

Observe that the example $\{E, O\}$ from before is just another way of writing $\mathbb{Z}/2\mathbb{Z}$.

Addition and multiplication both define binary operations on $\mathbb{Z}/n\mathbb{Z}$ in the obvious way; that is,

$$[a]_n + [b]_n = [a+b]_n, \qquad [a]_n \cdot [b]_n = [ab]_n.$$

Exercise 0.5. Check that these operations are well-defined; that is, that if $a, a', b, b' \in \mathbb{Z}$ are such that $[a]_n = [a']_n$ and $[b]_n = [b']_n$, then $[a+b]_n = [a'+b']_n$, and similarly for multiplication.

For simplicity of notation, we will from now on write a in place of $[a]_n$, with the understanding that we work modulo n; thus $\mathbb{Z}/n\mathbb{Z}$ may be thought of as comprising the integers from 0 to $n-1$.

It may easily be checked that $(\mathbb{Z}/n\mathbb{Z}, +)$ is a group, in which the identity element is 0 and the inverse of a is $n - a$. This group has the important property that it is generated by a single element; that is, if we consider the elements 1, $1+1$, $1+1+1$, and so on, we eventually get every element of the group. Such a group is called *cyclic*, and the element which is repeatedly added to itself (in this case 1) is called a *generator*.

Exercise 0.6. Find necessary and sufficient conditions for a to be a generator of $\mathbb{Z}/n\mathbb{Z}$.

The situation with $(\mathbb{Z}/n\mathbb{Z}, \cdot)$ is more complicated. This is a semigroup with identity element 1, but is not yet a group, because 0 has no inverse. In order to obtain a group, we should restrict our attention to invertible elements—are there any of these? a is invertible if and only if there exists b such that $ab \equiv 1 \bmod n$; equivalently, we require the existence of $b, d \in \mathbb{Z}$ such that $ab - nd = 1$. Using the Euclidean algorithm, we know that such integers exist if and only if a and n are relatively prime (their greatest common divisor is 1).

We now restrict our attention to the set

$$(\mathbb{Z}/n\mathbb{Z})^* = \{[a]_n \mid a \text{ and } n \text{ are relatively prime}\}.$$

Exercise 0.7. Show that $((\mathbb{Z}/n\mathbb{Z})^*, \cdot)$ is a group.

The fact that $(\mathbb{Z}/n\mathbb{Z})^*$ is a group under multiplication gives us an algebraic structure with which to work; we can use this structure to give simple proofs of some arithmetic facts.

Example 0.14. Let p be prime, so that $(\mathbb{Z}/p\mathbb{Z})^* = \{1, \dots, p-1\}$. We compute the residue class of $(p-1)!$ modulo p by observing that $(p-1)!$ is the product of all the elements of $(\mathbb{Z}/p\mathbb{Z})^*$, and so

$$[(p-1)!]_p = [1]_p[2]_p \cdots [p-1]_p. \tag{0.1}$$

Since $(\mathbb{Z}/p\mathbb{Z})^*$ is a group, every element has an inverse; for each $1 \leq a \leq p-1$ there exists b such that $[a]_p[b]_p = [1]_p$. Assuming $a \neq b$, we may remove $[a]_p$ and $[b]_p$ from (0.1) without changing the product. Repeating this process until we have removed every element which is not its own inverse, we see that

$$[(p-1)!]_p = \prod_{\{1 \leq a \leq p-1 \mid [a]_p^2 = [1]_p\}} a.$$

Now $[a]_p^2 = [1]_p$ if and only if $a^2 - 1 = (a-1)(a+1)$ is a multiple of p. Since p is prime, this implies that either $a-1$ or $a+1$ is a multiple of p, and hence $[a]_p$ is either $[1]_p$ or $[p-1]_p$. It follows that $[(p-1)!]_p = [p-1]_p$, or equivalently, $(p-1)! \equiv p-1 \bmod p$.

We may summarise this by saying that p divides $(p-1)! + 1$ for every prime number p, a fact which is not at all obvious without invoking the group structure of $(\mathbb{Z}/p\mathbb{Z})^*$.

Example 0.15. Fermat's Little Theorem states that if p is prime and $1 \leq a < p$, then $a^{p-1} \equiv 1 \bmod p$. To see this, we once again consider the product of all elements in $(\mathbb{Z}/p\mathbb{Z})^*$, as in (0.1), and then consider the following product:

$$[a^{p-1}(p-1)!]_p = [a]_p[2a]_p \cdots [(p-1)a]_p. \tag{0.2}$$

Because $(\mathbb{Z}/p\mathbb{Z})^*$ is a group, the elements in (0.2) are just a permutation of the elements in (0.1) (recall our discussion of multiplication tables). It follows that the products are the same, since we are in the abelian setting, and thus

$$[(p-1)!]_p = [a^{p-1}]_p[(p-1)!]_p.$$

Multiplying both sides on the right by $[(p-1)!]_p^{-1}$ gives the result.

Example 0.16. Fermat's Little Theorem can be generalised to Euler's Theorem, which states that $a^{\varphi(n)} \equiv 1 \bmod n$ whenever a and n are relatively prime, where $\varphi(n)$ is Euler's totient function, which counts the number of integers $1 \leq k < n$ which are relatively prime to n. This number is exactly the cardinality of the group $(\mathbb{Z}/n\mathbb{Z})^*$, and so the same argument as in Exercise 0.15 yields a proof of this result.

From now on we will use multiplicative notation for arbitrary groups, unless otherwise specified, and will often indicate the binary operation by juxtaposition, writing ab (or perhaps $a \cdot b$) in place of $a \star b$ and referring to the resulting element as the *product* of a and b. The identity element will usually be denoted by 1 rather than e. We will also denote exponents in the standard way: $a^0 = 1$ and $a^{n+1} = a^n \cdot a$. Negative exponents are defined by $a^{-n} = (a^{-1})^n$, where a^{-1} is the inverse of a.

Definition 0.17. Let X be an arbitrary group. The *order* of $a \in X$ is the smallest positive integer k such that $a^k = 1$; if no such integer exists, we say that the order of a is ∞.

Exercise 0.8. Show that if X is finite, then the order of every element divides the cardinality of the group.

e. Subgroups. Let G be a group, and suppose that $H \subset G$ is also a group with the same binary operation $\star$. Then H is called a *subgroup* of G. Since H is a group in its own right, it must satisfy the following properties:

(1) If $a, b \in H$, then $a \star b \in H$.
(2) If $a \in H$, then $a^{-1} \in H$.
(3) $e \in H$.

The first of these states that H is closed under multiplication, which guarantees that $\star$ is a genuine binary operation on H. Then associativity of $\star$ on H follows from associativity on G.

The second and third requirements guarantee that H is closed under taking inverses (hence inverses exist *in* H) and has an identity element. In fact, given an arbitrary non-empty subset $H \subset G$, it suffices to check (1) and (2) to see that H is a group; choosing any $a \in H$, (2) implies that $a^{-1} \in H$, and then (1) implies that $a \star a^{-1} = e \in H$.

Exercise 0.9. Show that $H \subset G$ is a subgroup if and only if $ab^{-1} \in H$ whenever $a, b \in H$.

The technique of passing to a subgroup is a standard source of new groups to study, and helps us to relate various groups with which we are already familiar.

Example 0.18.

(1) $(\mathbb{Z},+)$ and $(\mathbb{R},+)$ are subgroups of $(\mathbb{C},+)$.
(2) $(\mathbb{R}^+,\cdot)$ is a subgroup of $(\mathbb{C}\setminus\{0\},\cdot)$.
(3) The unit circle $S^1 = \{z \in \mathbb{C} \mid |z| = 1\}$ is closed under multiplication and inverses, and contains the identity, hence $(S^1,\cdot)$ is a subgroup of $(\mathbb{C}\setminus\{0\},\cdot)$.

A richer set of examples comes from considering the group $S(\mathbb{R}^n)$ of all invertible maps from $\mathbb{R}^n$ to itself. By considering classes of maps with certain other properties, we obtain various important subgroups.

Example 0.19. We have already observed that $n \times n$ real matrices may be thought of as linear maps on $\mathbb{R}^n$; thus $GL(n,\mathbb{R})$ may be viewed as a subgroup of $S(\mathbb{R}^n)$. An important subgroup of $GL(n,\mathbb{R})$ (and thus of $S(\mathbb{R}^n)$) is the *special linear group*

$$SL(n,\mathbb{R}) = \{A \in M(n,\mathbb{R}) \mid \det A = 1\}.$$

The proof that $SL(n,\mathbb{R})$ is a subgroup is an easy application of the fact that the identity matrix has determinant 1 and that the determinant is multiplicative ($\det AB = \det A \cdot \det B$).

Example 0.20. An *isometry* of $\mathbb{R}^n$ is a map $f\colon \mathbb{R}^n \to \mathbb{R}^n$ such that $d(f(\mathbf{x}), f(\mathbf{y})) = d(\mathbf{x},\mathbf{y})$ for every $\mathbf{x},\mathbf{y} \in \mathbb{R}^n$, where d is the Euclidean distance

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

Let $\mathrm{Isom}(\mathbb{R}^n)$ denote the set of all isometries of $\mathbb{R}^n$. It is easy to check that the composition of two isometries is an isometry, that the identity map is an isometry, and that the inverse of an isometry is an isometry, and it follows that $\mathrm{Isom}(\mathbb{R}^n)$ is a subgroup of $S(\mathbb{R}^n)$.

Example 0.21. Let $UT(n,\mathbb{R})$ denote the set of all upper triangular matrices in $GL(n,\mathbb{R})$; that is, all invertible $n \times n$ matrices A such that $A_{ij} = 0$ whenever $i > j$, and thus all entries below the main diagonal vanish. We claim that $UT(n,\mathbb{R})$ is a subgroup of $GL(n,\mathbb{R})$; clearly it contains the identity matrix, and it is not hard to show that it is closed under multiplication. However, it is not immediately clear why the inverse of an upper triangular matrix is upper triangular. One proof of this fact may be given using Cramer's rule, which gives an explicit formula for the inverse of a matrix.

Another proof is as follows. Observe that an $n \times n$ matrix A is upper triangular if and only if it can be decomposed as $A = DU$, where D is a diagonal matrix and U is upper triangular with all diagonal entries equal to 1. Now $A^{-1} = (DU)^{-1} = U^{-1}D^{-1}$; one may check that D^{-1} is a diagonal matrix, and thus it suffices to show that U^{-1} is upper triangular.

U has the form $I + X$, where all non-zero entries of X lie above the main diagonal. An easy computation shows that $X^n = 0$ (for X^2, the entries immediately above the main diagonal vanish, for X^3, the entries in the two diagonals above the main diagonal vanish, and so on)—such a matrix is called *nilpotent*. We may adapt the familiar formula $(1+x)^{-1} = 1 - x + x^2 - x^3 + \cdots$, since in this case the sum terminates, and we have

$$U^{-1} = (I+X)^{-1} = I - X + X^2 - \cdots + (-1)^{n-1}X^{n-1}.$$

Each X^k is upper triangular, thus U^{-1} is upper triangular, and we are done.

f. Homomorphisms and isomorphisms. As is the case with any mathematical structure, when we consider a map from one group to another, we are not interested in just any old map, but rather one which somehow preserves the algebraic structure given by the group. This is made precise by the following definition.

DEFINITION 0.22. Let G and H be groups. A map $\varphi\colon G \to H$ is a *homomorphism* if $\varphi(ab) = \varphi(a)\varphi(b)$ for every $a, b \in G$. If in addition φ is a bijection (one-to-one and onto), it is called an *isomorphism*, and we say that G and H are *isomorphic*.

This definition may be phrased as follows: a homomorphism is a map φ such that given any two elements $a, b \in G$, we may either multiply a and b, and then apply φ, or apply φ to a and b, and then multiply the resulting elements of H, with the same result.

Another way of saying the same thing is to state that φ is a homomorphism if the following diagram commutes:

$$\begin{tikzcd}
G \times G \arrow[r, "\varphi\times\varphi"] \arrow[d, "\star_G"] & H \times H \arrow[d, "\star_H"] \\
G \arrow[r, "\varphi"] & H
\end{tikzcd}$$

Here $\varphi \times \varphi$ is the map $(\varphi\times\varphi)(a,b) = (\varphi(a), \varphi(b))$.

PROPOSITION 0.23. *Let $\varphi\colon G \to H$ be a homomorphism. Then the following hold:*

(1) $\varphi(e_G) = e_H$;
(2) $\varphi(a^{-1}) = \varphi(a)^{-1}$ for every $a \in G$.

PROOF. We need only perform some simple computations.

(1) $\varphi(e_G) = \varphi(e_G e_G) = \varphi(e_G)\varphi(e_G)$. Multiplying both sides by $\varphi(e_G)^{-1}$ gives the result.
(2) $e_H = \varphi(e_G) = \varphi(aa^{-1}) = \varphi(a)\varphi(a^{-1})$, and multiplying both sides by $\varphi(a)^{-1}$ gives the result. □

EXAMPLE 0.24.

(1) Fix $a \in \mathbb{Z}$, and define $\varphi\colon \mathbb{Z} \to \mathbb{Z}$ by $\varphi(n) = an$. Then φ is a homomorphism. It is an isomorphism if and only if $n = \pm 1$.
(2) Fix $a \in \mathbb{R}$, and define $\varphi\colon \mathbb{R} \to \mathbb{R}$ by $\varphi(x) = ax$. Then φ is a homomorphism, and is also an isomorphism provided $a \neq 0$.

(3) Consider the exponential map $\varphi\colon \mathbb{R} \to \mathbb{R}^+$ given by $\varphi(x) = e^x$. We have $\varphi(x + y) = e^{x+y} = e^x e^y = \varphi(x)\varphi(y)$, and so φ is an isomorphism from $(\mathbb{R}, +)$ to $(\mathbb{R}^+, \cdot)$.
(4) Twist the exponential map by a right angle in the complex plane—that is, consider $\varphi(x) = e^{ix}$, which maps $\mathbb{R}$ to $\mathbb{C}$. Then φ is a homomorphism from $(\mathbb{R}, +)$ to $(\mathbb{C} \setminus \{0\}, \cdot)$, but is not an isomorphism, as it is neither one-to-one nor onto.

The last example provides a good illustration of an important general principle. From the viewpoint of group theory, two isomorphic groups are completely identical, and so it is of interest to know when two groups are isomorphic, as this often lets us translate problems from one setting into another, in which they may be more tractable, or which may give us new insights. A homomorphism does not give us such a clean equivalence: $(\mathbb{R}, +)$ and $(\mathbb{C} \setminus \{0\}, \cdot)$ have very different properties. However, if we can find a way to make the homomorphism a bijection, then we will have an isomorphism which carries some genuine information.

To do this, we must first make the map $\varphi\colon G \to H$ surjective by restricting our attention to the image of the map—that is, the set

$$\operatorname{Im} \varphi = \{\varphi(a) \mid a \in G\} \subset H.$$

Exercise 0.10. Show that $\operatorname{Im} \varphi$ is a subgroup of H.

In the present example, this corresponds to considering φ as a map from $(\mathbb{R}, +)$ to $(S^1, \cdot)$, in which case it becomes onto. To make it one-to-one, we will need to introduce the *kernel* of a homomorphism, *normal subgroups*, and *quotient groups*.

Lecture 1. Monday, August 31

a. Generalities. Let us briefly review where we stand. We began by defining a group—a set together with a binary operation which is associative, has an identity element, and with respect to which every element is invertible. We then moved on to define and discuss the concepts of subgroups, homomorphisms, and isomorphisms.

These last three concepts, which are absolutely foundational to the study of groups, are in fact not specific to group theory, but are really much more general in scope. Let us explain what we mean by this.

In most branches of modern mathematics, one begins by considering a set equipped with a certain sort of structure, which is defined by a list of axioms. In the present case, the structure is a binary operation, and the axioms are listed above. If we consider a set with *two* binary operations, and if in addition we require them to interact in a certain way (which mirrors the relationship between addition and multiplication of integers), then we are dealing with another sort of algebraic object called a *ring.* Further examples abound: a set with a linear structure is a *vector space*, a set with a notion of convergence is a *topological space*, a set with a notion of distance is a *metric space*, and so on and so forth.

Whatever structure we consider on a set X, we may then consider subsets of X which inherit a similar structure; thus we obtain subgroups, subrings, subspaces, etc. Given two sets X and Y with a particular structure, we may also consider maps $f\colon X \to Y$ which preserve that structure. This is the general notion of a *morphism*; depending on the structure we are studying, we may refer to such maps as homomorphisms, linear maps, continuous maps, isometries, etc.

Invertible morphisms—in the present case, isomorphisms—are particularly important, because they allow us to treat X and Y as equivalent, from the point of view of the structure being studied, by defining an *equivalence relation* on the class of all sets endowed with that structure. This allows us to pose the problem of *classifying* such objects, which in the case of groups may be phrased as follows: *Produce an explicit list of groups such that no two groups on the list are isomorphic to each other, and such that every group is isomorphic to a group on the list.*

If we consider sets with no additional structure, then the relevant equivalence relation is nothing but the existence of a bijection; two sets X and Y are equivalent *as sets* if there exists a bijection from one to the other. Thus the classification problem for sets reduces to producing a list of all possible cardinalities. For finite sets, this is straightforward, as we need merely produce one set with each finite cardinality; this amounts to constructing the natural numbers.[1]

[1] For infinite sets, the matter becomes somewhat more delicate, as one encounters the continuum hypothesis and other such set theoretic beasts.

Another familiar example is sets with a linear structure—vector spaces. A complete classification of finite-dimensional vector spaces is given by the list of Euclidean spaces $\mathbb{R}^n$; every finite-dimensional vector space is isomorphic to the Euclidean space with the same dimension.[2]

It turns out that classifying groups, even only finite, is a much harder problem than classifying either finite sets or finite-dimensional vector spaces. It is too much to ask for a classification of *all* (even finite) groups, and so one must instead proceed by classifying particular classes of groups. For certain classes of group, e.g. abelian ones, this problem is manageable, as will will soon explain. On the other hand, if one considers (finite) groups in a natural sense opposite to abelian (such groups are called *simple* and we will define this notion in due time), their classification has been one of the outstanding achievements of twentieth-century algebra.

b. Cyclic groups. We begin by describing a class of groups which are in some sense the "simplest possible" (but far from being simple in the sense alluded to above).

Definition 1.1. Given a group G and an element $g \in G$, the *subgroup generated by* g if $\langle g \rangle = \{g^n \mid n \in \mathbb{Z}\}$. We say that G is *cyclic* if there exists $g \in G$ such that $\langle g \rangle = G$. Such a g is called a *generator* of G.

Note that the elements g^n may not be distinct; we may have $g^m = g^n$ for some $m \neq n$.

Remark. Let $g \in G$ be arbitrary (not necessarily a generator). Since subgroups are closed under the binary operation, any subgroup which contains g must also contain $\langle g \rangle$. It follows that g is a generator if and only if it is not contained in any subgroups of G other than G itself.

Example 1.2.

(1) The infinite group $(\mathbb{Z}, +)$ (which from now on will simply be denoted $\mathbb{Z}$) is cyclic; its generators are ± 1.
(2) The group $n\mathbb{Z} = \{\dots, -2n, -n, 0, n, 2n, \dots\}$ is cyclic; its generators are $\pm n$. In fact, this group is isomorphic to $\mathbb{Z}$ via the map $\varphi\colon \mathbb{Z} \to n\mathbb{Z}$ given by $\varphi(a) = na$.
(3) The group of residue classes $(\mathbb{Z}/n\mathbb{Z}, +)$ (from now on simply denoted $\mathbb{Z}/n\mathbb{Z}$) is a cyclic group with n elements. As with $\mathbb{Z}$, the element 1 is a generator; however, there are typically other generators as well. For example, in the case $n = 5$, we see that the subgroup generated by 2 is $\{0, 2, 4, 1, 3\}$, and so 2 generates $\mathbb{Z}/5\mathbb{Z}$.

Theorem 1.3. *Every cyclic group is isomorphic to either* $\mathbb{Z}$ *or* $\mathbb{Z}/n\mathbb{Z}$ *for some* $n \in \mathbb{N}$.

[2] Again, for infinite-dimensional vector spaces life is more interesting, and in this context one typically considers vector spaces with some additional structure, such as Banach spaces or Hilbert spaces.

Proof. Let $G = \langle g \rangle$. Suppose that $g^n \neq e$ for every $n \in \mathbb{Z}$, $n \neq 0$. Then $g^m \neq g^n$ for all $m \neq n$ (otherwise $g^{m-n} = e$), and hence the map

$$\begin{aligned} \varphi\colon \mathbb{Z} &\to G \\ n &\mapsto g^n \end{aligned}$$

is one-to-one. φ is onto since g generates G, and it is easy to check that φ is a homomorphism, and hence an isomorphism.

Now suppose there exists n such that $g^n = e$, and let n be the smallest positive integer with this property. Define $\varphi\colon \mathbb{Z}/n\mathbb{Z} \to G$ by $\varphi(k) = g^k$. The fact that φ is a homomorphism follows from the fact that $g^n = e$; injectivity follows since n is minimal with this property; and surjectivity follows since g generates G. Thus φ is an isomorphism. □

When we study a group G, one of the most important insights into its structure comes from determining its subgroups. Since Theorem 1.3 tells us that the groups $\mathbb{Z}$ and $\mathbb{Z}/n\mathbb{Z}$ are universal models for all cyclic groups, we can understand the subgroup structure of cyclic groups by understanding the subgroups of these two examples.

Proposition 1.4. *Every subgroup of the infinite cyclic group $\mathbb{Z}$ is of the form $n\mathbb{Z}$ for some $n \in \mathbb{Z}$.*

Proof. Given a subgroup $G \subset \mathbb{Z}$, let n be the minimal positive element in G. It follows that $n\mathbb{Z} \subset G$. Now suppose there exists $k \in G$ such that $k \notin n\mathbb{Z}$. Then there exists $1 \leq r < k$ and $q \in \mathbb{Z}$ such that $k = qn + r$; since G is a subgroup and since $k, n \in G$, we have $r = k - qn \in G$ as well. This contradicts the assumption that n is minimal, and we conclude that $G = n\mathbb{Z}$. □

Proposition 1.5. *Every subgroup of a finite cyclic group $\mathbb{Z}/n\mathbb{Z}$ is of the form $\langle k \rangle$, where k is a factor of n.*

Proof. Once again, fix a subgroup $G \subset \mathbb{Z}/n\mathbb{Z} = \{0, 1, \dots, n-1\}$, and let k be the minimal positive element of G. The same argument as in Proposition 1.4 shows that $G = \langle k \rangle$. To see that k divides n, let $q \in \mathbb{N}$ and $0 \leq r < k$ be such that $n = qk - r$, and thus $qk \equiv r \bmod n$. It follows that $r \in G$, and by the minimality of k, we must have $r = 0$, hence k divides n. □

Corollary 1.6. *If p is prime, then $\mathbb{Z}/p\mathbb{Z}$ has no nontrivial subgroups.*

Corollary 1.7. *An element $a \in \mathbb{Z}/n\mathbb{Z}$ is a generator if and only if a and n are relatively prime.*

Proof. By Proposition 1.5, every subgroup of $\mathbb{Z}/n\mathbb{Z}$ is of the form $\langle k \rangle$, where k divides n. $a \in \mathbb{Z}/n\mathbb{Z}$ is a generator if and only if it is not contained in any such subgroup (except for the case $k = 1$). But $a \in \langle k \rangle$ if and only if k divides a, and so this is the statement that no factor of n divides a, with the exception of 1, which is the statement that a and n are relatively prime. □

c. Direct products. Having described the simplest possible groups—cyclic groups—we now examine ways to build more complicated groups from these basic building blocks.

Definition 1.8. Let G and H be groups. The *direct product* of G and H is the set

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

together with the binary operation

$$(g_1, h_1) \cdot (g_2, h_2) = (g_1 g_2, h_1 h_2).$$

Exercise 1.1. Show that $G \times H$ is a group with identity element (e_G, e_H).

Remark. Although the groups we deal with may not be abelian, the operation of taking a direct product is commutative in the sense that $G \times H$ and $H \times G$ are isomorphic groups.

Example 1.9. Consider the group $V = (\mathbb{Z}/2\mathbb{Z}) \times (\mathbb{Z}/2\mathbb{Z})$, which is the direct product of two cyclic groups of order two; this is often called the *Klein four-group*. We see that $V = \{(0, 0), (0, 1), (1, 0), (1, 1)\}$, and that every element of V (besides the identity $(0, 0)$) has order two. In particular, V has no generators, and so is not cyclic.

Example 1.10. Now consider $G = (\mathbb{Z}/2\mathbb{Z}) \times (\mathbb{Z}/3\mathbb{Z})$. We have $G = \{(0, 0), (0, 1), (0, 2), (1, 0), (1, 1), (1, 2)\}$, and a simple computation shows that the subgroup generated by $(1, 1)$ is

$$\langle (1, 1) \rangle = \{(0, 0), (1, 1), (0, 2), (1, 0), (0, 1), (1, 2)\} = G.$$

Thus $(1, 1)$ is a generator, and G is cyclic. It follows from Theorem 1.3 that G is isomorphic to $\mathbb{Z}/6\mathbb{Z}$.

What is the difference between the previous two examples? Why is the direct product cyclic in one case, and not in the other? The answer lies in the following result.

Proposition 1.11. *Let G and H be finite cyclic groups. The direct product $G \times H$ is cyclic if and only if $|G|$ and $|H|$ are relatively prime.*

Proof. By Theorem 1.3, it suffices to consider the case $G = \mathbb{Z}/m\mathbb{Z}$, $H = \mathbb{Z}/n\mathbb{Z}$. We first show that if m and n are relatively prime, then $G \times H$ is cyclic; indeed, $(1, 1)$ is a generator.

To see this, observe that $\langle (1, 1) \rangle = \{(a, a) \mid a \in \mathbb{Z}\}$, and that $(a, a) = (k, \ell)$ in $G \times H$ if and only if there exist $p, q \in \mathbb{Z}$ such that $a = pm + k = qn + \ell$. In this case we have $pm - qn = \ell - k$. Now it follows from our assumption that m and n are relatively prime that for every (k, ℓ), we may use the Euclidean algorithm to find such a p and q. Letting $a = pm + k$, we have $(k, \ell) = (a, a) \in \langle (1, 1) \rangle$, and it follows that $(1, 1)$ is a generator.

Now suppose m and n are *not* relatively prime, and let $d > 1$ be an integer which divides both of them. Let $k = mn/d$, and observe that for any $(a, b) \in G \times H$ we have $(ka, kb) = (0, 0)$ since m and n both divide k.

It follows that the order of (a, b) divides k; in particular, because $k < mn = |G \times H|$, (a, b) is not a generator. This is true for any choice of a and b, hence $G \times H$ is not cyclic. □

d. Classification. We can now offer a complete description of the structure of all finite cyclic groups. Given a cyclic group $G = \mathbb{Z}/n\mathbb{Z}$, write the prime factorisation of n as

$$n = p_1^{k_1} \cdots p_m^{k_m},$$

where $p_i \neq p_j$ for $i \neq j$. The different factors $p_i^{k_i}$ are relatively prime, and so Proposition 1.11 implies that

$$G = (\mathbb{Z}/p_1^{k_1}\mathbb{Z}) \times \cdots \times (\mathbb{Z}/p_m^{k_m}\mathbb{Z}). \tag{1.1}$$

Exercise 1.2. Let p be prime and $k \in \mathbb{N}$. Show that $\mathbb{Z}/p^k\mathbb{Z}$ has exactly one subgroup of order p^j for each $0 \leq j \leq k$, and no other subgroups.

Exercise 1.2 shows that the cyclic groups whose order is a power of a prime form a nice set of building blocks with which to work, since their subgroup structure is quite transparent.

In fact, (1.1) is just a particular case of the more general *structure theorem for finite abelian groups*, which we only formulate now and will proof later in this course.

Theorem 1.12. *Let G be a finite abelian group. Then G can be written in the form* (1.1)*, where the p_i are not necessarily distinct.*

Of course, there is more to life than finite abelian groups. But this gives us a good handle on one particular corner of the algebraic world, and shows us how a more general class of groups (finite abelian groups) can be built up from a much simpler one (cyclic groups of prime power order).

We have seen a number of results having to do with divisibility relationships between the orders of various elements in a group, and between the order of a subgroup and the order of a group. In fact, these relationships hold far beyond the cyclic setting.

Theorem 1.13 (Lagrange's Theorem). *Let G be a finite group and H a subgroup of G. Then $|H|$ divides $|G|$.*

Proof. We need the notion of a *coset*. Fix an element $g \in G$. The *left coset* of H corresponding to g is

$$gH = \{gh \mid h \in H\};$$

the *right coset* Hg is defined similarly. Whether we consider left or right cosets is a somewhat arbitrary choice; for now we consider left cosets, although right cosets would work just as well.

The key observation is that the left cosets of H partition G; to see this, fix two cosets g_1H and g_2H, and suppose that $g_1H \cap g_2H \neq \emptyset$. Then there exist $h_1, h_2 \in H$ such that $g_1h_1 = g_2h_2 \in g_1H \cap g_2H$, and it follows that

$g_2^{-1}g_1 = h_2^{-1}h_1 \in H$ (using the definition of a subgroup). From this we obtain $g_2^{-1}g_1H = H$, and multiplying on the left by g_2 yields $g_1H = g_2H$.

The preceding argument shows that any two cosets g_1H and g_2H are either disjoint or equal; it follows that the left cosets of H partition G, and since $|gH| = |H|$ for every $g \in G$, we obtain

$$|G| = |H| \cdot (\text{number of left cosets of } H \text{ in } G).$$

Thus $|H|$ divides $|G|$, and we are done. □

Definition 1.14. The number $|G|/|H|$ is called the *index* of the subgroup H, and may be defined for subgroups of arbitrary (not necessarily finite) groups as the number of left (or right) cosets.

Observe that the result of Exercise 0.8 may be obtained as a corollary of Lagrange's Theorem. The following consequence of Lagrange's Theorem will also be helpful.

Corollary 1.15. *If $|G|$ is prime, then G is cyclic.*

Proof. Fix $g \in G$, $g \neq e$. Then $|\langle g\rangle| > 1$, and since $|\langle g\rangle|$ divides $|G|$, which is prime, we must have $|\langle g\rangle| = |G|$, which is only possible if $\langle g\rangle = G$. It follows that g is a generator; hence G is cyclic. □

Armed with these results, we sally forth to classify the smallest groups we can find. (One must begin somewhere.)

Groups of order ≤ 3 are easy: any group of order 1 is trivial, while Corollary 1.15 tells us that any group of order 2 or 3 (and also 5) is cyclic, and hence isomorphic to $\mathbb{Z}/2\mathbb{Z}$ or $\mathbb{Z}/3\mathbb{Z}$ by Theorem 1.3.

We have already encountered two non-isomorphic groups of order 4: the cyclic group $\mathbb{Z}/4\mathbb{Z}$, and the Klein four-group $(\mathbb{Z}/2\mathbb{Z}) \times (\mathbb{Z}/2\mathbb{Z})$.

Proposition 1.16. *Any group of order 4 is isomorphic to one of these two groups.*

Proof. Let $|G| = 4$. We need to show that is G is not cyclic it is isomorphic to $(\mathbb{Z}/2\mathbb{Z}) \times (\mathbb{Z}/2\mathbb{Z})$. For, every element of such a groups G has order 2 and hence is equal to its inverse Let $g_1 \neq g_2$ be two non-identity elements of G; both g_1g_2 and g_2g_1 are not equal to identity and neither equals g_1 or g_2. Hence those elements coincide with the only remaining element element of the group. Thus the map $\phi : G \to (\mathbb{Z}/2\mathbb{Z}) \times (\mathbb{Z}/2\mathbb{Z})$ defined as $\phi(e) = (0,0)$, $\phi(g_1) = (0,1)$, $\phi(g_2) = (0,1)$, $\phi(g_1g_2) = (1,1)$ is an isomorphism. □

Once again, any group of order 5 is cyclic, and hence isomorphic to $\mathbb{Z}/5\mathbb{Z}$. At the next step, though—groups of order 6—we suddenly find ourselves with a little variety in our life. Of course one has the cyclic group $\mathbb{Z}/6\mathbb{Z}$, but now a new breed of group appears, which we discuss next.

e. Permutation groups. We have proven by exhaustion that every group with ≤ 5 elements is abelian. We now exhibit a non-abelian group with 6 elements. Consider an equilateral triangle T in the plane, and let S_3 be its set of symmetries. That is, S_3 is the set of all isometries of the plane which map T to itself; of course the identity map is in S_3, but we also have two rotations (by $\pi/3$ and by $2\pi/3$ around the centre of the triangle) and three reflections (through the three lines of symmetry), so that $|S_3| = 6$.

Equipping S_3 with the binary operation of composition, we obtain a group, called the *symmetric group of order* 3, whose structure we will study in more detail in the next lecture. For now, we observe that if we label the three vertices of T and simply keep track of where the vertices are mapped, then we know the entire action of any given element of S_3. Thus using the labels $\{1, 2, 3\}$, we see that S_3 may also be thought of as the group of all permutations of the set $X_3 = \{1, 2, 3\}$; to use our earlier notation, this is $S(X_3)$.

More generally, one may consider the symmetric group of order n, which consists of all permutations of n elements, and is denoted $S_n = S(X_n)$.

Proposition 1.17. $|S_n| = n!$.

Proof. Image of 1 can be defined in n different ways; that fixed image of 2 can be defined in $n - 1$ different ways and so on. □

Since $n!$ is a highly divisible number, Lagrange's Theorem permits many possible orders for the subgroups of S_n. This leaves the door open for a high degree of complexity in the algebraic structure of S_n, although it does not guarantee it. In fact, this complexity is present, and we state two ways in which it is manifested.

The first way, which is quite simple, is the observation that the groups S_n are nested; that is,

$$S_2 \subset S_3 \subset S_4 \subset \cdots \subset S_n \subset S_{n+1} \subset \cdots .$$

(In fact, S_n contains not just one, but n subgroups which are isomorphic to S_{n-1}.)

The second (rather deeper) way in which the complex structure of S_n manifests itself is the following:

Theorem 1.18. *Every finite group is isomorphic to a subgroup of S_n (for sufficiently large n).*

Proof. Notice that every group acts by right translations on the set of its elements. To be precise, consider for $g \in G$ the transformation R_g that maps every element $h \in G$ into hg. Then composition R_{g_1} and R_{g_2} (in that order) equals to $R_{g_1 g_2}$. Hence $g \mapsto R_g$ is an injective homomorphism into the group of all bijections of G. If G is finite its bijections can be identified with permutations of order $|G|$ simply by numbering elements of G in an arbitrary way. Thus the homomorphism described above is an isomorphism between G and a subgroup of $S_{|G|}$. □

A subgroup of S_n is called a *permutation group*. Theorem 1.18 says that every finite group is isomorphic to a permutation group. This indicates that we should not expect to gain complete understanding of the structure of subgroups of S_n, since any behaviour which occurs in any finite group occurs in them as well. This will not prevent us into gaining insights into important structural properties of that group.

Lecture 2. Wednesday, September 2

a. Representations. We begin with a theorem which was stated last time.

THEOREM 2.1. *Let G be a finite group. Then G is isomorphic to a subgroup of S_n, where $n = |G|$.*

PROOF. Given $g \in G$, define a map $R_g : G \to G$ by $R_g(h) = hg$. This may be thought of as *right translation* by g; one could also define *left translation* by $L_g(h) = gh$. It follows from the axioms of a group that R_g is a bijection:

(1) If $R_g h_1 = R_g h_2$, then $h_1 g = h_2 g$; hence $h_1 = (h_1 g)g^{-1} = (h_2 g)g^{-1} = h_2$, so R_g is one-to-one.
(2) Given an arbitrary $h \in G$, we have $R_g(hg^{-1}) = hg^{-1}g = h$, so R_g is onto.

Now the map $\varphi : g \mapsto R_g$ defines a map from G into $S(G)$, the group of all bijections of G onto itself. Because G is finite, $S(G)$ may also be written as $S_{|G|} = S_n$, and so it remains to show that φ is an isomorphism onto its image.

Given $g_1, g_2 \in G$, we write $R_{g_1} R_{g_2}$ for the element of S_n obtained by applying first R_{g_1}, then R_{g_2}; this could also be written as $R_{g_2} \circ R_{g_1}$, where it is important to note that the order is reversed.

To see that φ is a homomorphism, we observe that for every $h, g_1, g_2 \in G$, we have

$$(R_{g_1} R_{g_2})h = R_{g_2}(R_{g_1} h) = R_{g_2}(hg_1) = hg_1 g_2 = R_{g_1 g_2} h.$$

It follows that $\varphi(g_1 g_2) = \varphi(g_1)\varphi(g_2)$.

Finally, observe that $\varphi(g)$ is the identity map if and only if $hg = h$ for all $h \in G$. Since the identity element is the only element with this property, the kernel of φ is trivial, hence φ is one-to-one. It follows that φ is a bijective homomorphism—an isomorphism—from G onto $\varphi(G) \subset S_n$. □

EXERCISE 2.1. Carry out the analogous construction using L_g instead of R_g. Be careful with the ordering....

Theorem 2.1 is a concrete illustration of one of the purposes for which homomorphisms and isomorphisms are designed—to build bridges between the structure of different groups. When such bridges are built, it is often the case that some of the structure of the first group is lost in translation; this motivates the following definition.

DEFINITION 2.2. Given a homomorphism $\varphi : G \to H$, the *kernel* of φ is the set

$$\ker \varphi = \{g \in G \mid \varphi(G) = e_H\}. \tag{2.1}$$

Notice that $\ker \varphi$ is a subgroup of G; If $\varphi(g_1) = \varphi(g_2) = e_H$ then $\varphi(g_1 g_2) = \varphi(g_1)\varphi(g_2) = e_H e_H = e_H$

Furthermore if $\ker\varphi$ is trivial then φ is injective since $\varphi(g_1) = \varphi(g_2)$ implies that $\varphi(g_1^{-1}g_2 = e_H$

REMARK. It may be helpful to realise that if V and W are vector spaces, then they are also abelian groups (with the binary operation of vector addition). In this case, a linear map $\varphi\colon V \to W$ is also a group homomorphism, and the kernel of φ is nothing but the null space.

Any element in the kernel of φ is in some sense erased by the action of the homomorphism. Thus homomorphisms for which the kernel is trivial ($\ker\varphi = \{e\}$) are of particular importance; such a homomorphism is called an *embedding*, and we say that G is *embedded* into H. If we consider φ as a map from G to the image $\varphi(G) \subset H$, then it is a bijection, and so G and $\varphi(G)$ are isomorphic. The subgroup $\varphi(G)$ is the *isomorphic image* of G in H.

Using this language, the moral of the story told in Theorem 2.1 can be stated as follows: There is a class of universal objects (the symmetric groups) into which every finite group can be embedded. Each such embedding gives us a concrete realisation of a particular group, representing it as a permutation group.

There are other classes of universal objects which we might use to represent abstract groups. The most important such class is the class of matrix groups, and chief among these is the *general linear group*

$$GL(n,\mathbb{R}) = \{A \in M(n,\mathbb{R}) \mid \det A \neq 0\}, \tag{2.2}$$

where $M(n,\mathbb{R})$ is the set of all $n \times n$ matrices with real entries.

DEFINITION 2.3. A homomorphism $\varphi\colon G \to GL(n,\mathbb{R})$ is called a *linear representation* of G. If $\ker\varphi$ is trivial, we say that the representation is *faithful.*

Thus a faithful linear representation of a group G is an embedding of G into a matrix group; to put it another way, a representation of G is a concrete realisation of G as a collection of invertible matrices, and the representation is faithful if distinct elements of G correspond to distinct matrices.

PROPOSITION 2.4. *The symmetric group S_n can be embedded into $GL(n,\mathbb{R})$.*

PROOF. S_n is the group of all permutations of n symbols. To embed S_n into $GL(n,\mathbb{R})$, we must find a one-to-one homomorphism $\varphi\colon S_n \to GL(n,\mathbb{R})$. This involves assigning to each permutation $g \in S_n$ an invertible linear map $\varphi(g)\colon \mathbb{R}^n \to \mathbb{R}^n$.

An easy way to do this is to fix n linearly independent elements in $\mathbb{R}^n$—say the standard basis vectors $\mathbf{e}_1, \dots, \mathbf{e}_n$—and to let $\varphi(g)$ be the unique linear map which permutes these elements according to the permutation given by g. That is, each $g \in S_n$ corresponds to a bijection $g\colon \{1, \dots, n\} \to \{1, \dots, n\}$. Define the action of $\varphi(g)$ on the vectors $\mathbf{e}_1, \dots, \mathbf{e}_n$ by

$$\varphi(g)\mathbf{e}_i = \mathbf{e}_{g(i)};$$

since the vectors $\mathbf{e}_i$ are linearly independent, $\varphi(g)$ extends uniquely to a linear map on their span, which is $\mathbb{R}^n$.

The map φ is a homomorphism since a linear map is uniquely defined by its action on any basis; by the same reason $\ker \varphi$ is trivial. □

Choosing the standard basis vectors for the n linearly independent elements in the proof of Proposition 2.4, the linear transformations $\varphi(g) \in GL(n, \mathbb{R})$ will be represented by *permutation matrices*—that is, matrices for which every row and column contains $n - 1$ entries equal to 0 and a single entry equal to 1.

Remark. We will see later that linear representation of S_n described above is not the "most economical"; for example S_n has a faithful representation in $GL(n - 1, \mathbb{R})$.

Combining Theorem 2.1 and Proposition 2.4, we see that any finite group can be embedded into a matrix group; to put it another way, every finite group has a faithful linear representation.

b. Automorphisms: Inner and outer. Not all elements of a group G are created equal. For example, the identity element is distinct from any other element in that it leaves all elements unchanged under multiplication. This property is completely intrinsic to the structure of the group, and is independent of any particular concrete realisation or representation of the group which we may choose.

Let us make this notion of "intrinsic property" more precise. An isomorphism $\varphi \colon G \to G$ is called an *automorphism*, and may be thought of as a "symmetry" of the group G. For example, the map $\varphi \colon \mathbb{Z} \to \mathbb{Z}$ defined by $\varphi(n) = -n$ is an automorphism, while the map $n \mapsto 2n$ is not. Given two elements $g, h \in G$, if there exists an automorphism φ from G to itself that maps g to h, we ought to consider g and h as having the same intrinsic properties.

One may consider a similar notion in the geometric setting. If X denotes an equilateral triangle, then the three vertices of X have the same intrinsic geometric properties (for example, the property of "being a vertex at which the angle is $\pi/3$"), while points which lie on a side, but not a vertex, have different intrinsic geometric properties. This is reflected in the fact that if $x, y \in X$ are both vertices, we can find an isometry $f \colon X \to X$ (such an isometry is called a *symmetry*) such that $f(x) = y$, while no such symmetry can be found if x is a vertex and y is not. Notice that the situation changes if one considers triangles that is not equilateral; here not all vertices are intrinsically the same anymore.

Another geometric example is the circle, for which *any* two points have the same intrinsic geometric properties; given any $x, y \in S^1$, there exists an isometry $f \colon S^1 \to S^1$ such that $f(x) = y$. Thus the fact that "no points on the circle are any more special than any others" is once again reflected in the presence of a great many symmetries.

Passing from geometry back to algebra, we replace the notion of symmetry with the notion of automorphism. The phenomenon illustrated above happens here as well; some groups have more automorphisms than others. Groups with many automorphisms are somehow "more symmetric", while groups with fewer automorphisms are "less symmetric". In the former case, many elements play very similar roles within the group structure, while in the latter, there are more "special" elements.

An example of an intrinsic algebraic property is the order of an element. If φ is any automorphism, then g and $\varphi(g)$ have the same order. Thus if g is the only element of G with order 2 (for example), then it cannot be mapped to any other element of G by an automorphism, and thus is quite special, just like the identity element.

Since identity is distinguished from all other elements the most symmetric groups are those where there is an automorphism that maps any given non-identity element into any other non-identity element. Among abelian groups there are some such groups.

For example, for the additive group $\mathbb{R}$ of real numbers multiplication by ant non-zero number is an automorphism (this is another formulation of the distributive property). Among finite groups, all prime cyclic groups $\mathbb{Z}/p\mathbb{Z}$ have this property by essentially the same reason: multiplication by any non-zero element is an automorphism. But there are other examples.

Exercise 2.2. Show that in the direct product of any number of $\mathbb{Z}/2\mathbb{Z}$ there exists an automorphism that maps any non-zero element to another given non-zero element.

It turns out that non-abelian groups have certain automorphisms due to their non-commutativity.

Example 2.5. Fix $h \in G$, and define a map $I_h \colon G \to G$ by

$$I_h(g) = h^{-1}gh. \tag{2.3}$$

Then $I_h(g_1g_2) = h^{-1}(g_1g_2)h = h^{-1}g_1hh^{-1}g_2h = I_h(g_1)I_h(g_2)$; hence I_h is a homomorphism. The map $g \mapsto gh$ and the map $g \mapsto h^{-1}g$ are both bijections, and so I_h is as well. Thus I_h is an automorphism of G.

Definition 2.6. An automorphism of the form (2.3) is called an *inner automorphism*. An automorphism φ which cannot be written in the form (2.3) is called an *outer automorphism*.

Of course, if G is abelian then every inner automorphism is trivial, and so the only possible automorphisms are the outer automorphisms. Thus in some sense, what we are doing here is using non-commutativity to our advantage, by building symmetries of G which are not present in the abelian case. This suggests that the presence of a small number of inner automorphisms corresponds to a large amount of commutativity, while the presence of many inner automorphisms corresponds to a large amount of non-commutative behaviour.

We have noticed above that some abelian groups are very symmetric. Now we consider some other cases.

Exercise 2.3. Show that the only automorphisms of $\mathbb{Z}$ are $n \mapsto n$ and $n \mapsto -n$.

Now let us find the automorphisms of $\mathbb{Z} \times \mathbb{Z}$. Since the group is abelian, there are no inner automorphisms, and so we must turn to other techniques.

Let A be an automorphism of $\mathbb{Z}\times\mathbb{Z}$, and let $A(1,0) = (a,b)$ and $A(0,1) = (c,d)$. Then $A(m,n) = (am + cn, bm + dn)$, and we see that A acts on $\mathbb{Z} \times \mathbb{Z}$ as multiplication by the matrix $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$, and so each automorphism of the group corresponds to a 2×2 matrix with integer entries.

What properties does this matrix have? Since A is invertible, this matrix must be invertible; furthermore, the inverse of A is again an automorphism, and hence the inverse of the matrix must again have integer entries. If $\det X = \pm 1$ this is the case not only for 2×2 but for $n \times n$ matrices as well due to the Kramer's rule: elements of the X^{-1} are polynomials in the elements of X with integer coefficients divided by the determinant. But the inverse is true as well. For 2×2 matrices this is particularly easy to see. If X^{-1} is an integer matrix then all elements of X divide its determinant which is possible only if the latter is ± 1.

Thus the automorphisms of $\mathbb{Z} \times \mathbb{Z}$ correspond to 2×2 integer matrices with determinant equal to ± 1.

Exercise 2.4. Determine necessary and sufficient conditions on (a,b) and (c,d) in $\mathbb{Z} \times \mathbb{Z}$ for there to exist an automorphism which maps (a,b) to (c,d).

c. Cosets and factor groups. Moving away from automorphisms for the moment, let us return to another facet of the internal structure of a group G—its subgroups. Because of the algebraic structure which is imposed, a subgroup $H \subset G$ is much more than just a subset. Indeed, in the proof of Lagrange's Theorem (Theorem 1.13), we introduced the notion of *coset*, which is a translation of H, and saw that the left cosets of H form a partition of G (and the same is true of the right cosets).

Remark. If G is abelian, then left and right cosets coincide for every group, and so we may simply speak of "cosets" without any ambiguity.

Example 2.7.

(1) The real line $\mathbb{R}$ is a subgroup of the complex plane $\mathbb{C}$; its cosets are the horizontal lines
$$iy + \mathbb{R} = \{x + iy \mid x \in \mathbb{R}\},$$
where each value of $y \in \mathbb{R}$ determines a different coset.

(2) The set of multiples of n, denoted $n\mathbb{Z}$, is a subgroup of the integers $\mathbb{Z}$. Its cosets are the residue classes modulo n, which already appeared in the definition of the cyclic groups $\mathbb{Z}/n\mathbb{Z}$.

The visual image provided by the first of these examples is that the cosets are somehow "parallel" images of the subgroup which do not pass through the origin. The relationship between a subgroup and its cosets is exactly the same as the relationship between a linear subspace and an affine subspace, to draw once more upon the language of linear algebra.

Given a subgroup $H \subset G$, we want to examine the group structure of the part of G which is missed by H—the part which lies "transverse" to H. This is done by turning the cosets themselves into a group.

First we recall that the binary operation on elements of G defines a binary operation on subsets of G:

$$AB = \{ab \mid a \in A, b \in B\}. \tag{2.4}$$

This is defined for *all* subsets $A, B \subset G$, whether or not they are subgroups, cosets, or anything else significant.

We would like to use this binary operation to define a group structure on the set of (left) cosets of a subgroup H. However, there is a problem. Why should the set $(g_1H)(g_2H)$ be a left coset of H? It turns out that we need to ask a little more of H.

Definition 2.8. A subgroup $H \subset G$ is a *normal subgroup* if $gH = Hg$ for all $g \in G$; that is, every left coset is also a right coset. Equivalently, we may demand that $gHg^{-1} = H$ for all $g \in G$.

Proposition 2.9. *If H is a normal subgroup of G, then* (2.4) *defines a binary operation on the set of left cosets of H.*

Proof. Given any two cosets g_1H and g_2H, we have $g_2H = Hg_2$, and hence

$$\begin{aligned}(g_1H)(g_2H) &= \{g_1h_1g_2h_2 \mid h_1, h_2 \in H\}\\ &= g_1(Hg_2)H\\ &= g_1(g_2H)H\\ &= (g_1g_2)H.\end{aligned}$$

□

Definition 2.10. Given a normal subgroup $H \subset G$, the *factor group* G/H is the set of left cosets of H equipped with the binary operation defined in Proposition 2.9.

There is a close relation between normal subgroups and homomorphisms. Namely, the kernel of any homomorphism $\varphi : G \to H$ is a normal subgroups of G. This follows for the simple fact that identity element commutes with any other element: if $k \in \ker\varphi$ and g is arbitrary then $\varphi(gkg^{-1}) = \varphi(g)\varphi(g^{-1}) = e_H$. Furthermore, any homomorphic image of a group is isomorphic to the factor group by the kernel.

Example 2.11. If G is abelian, then as already noted, left and right cosets coincide for every subgroup; hence every subgroup is normal. In particular, consider the subgroup $n\mathbb{Z} \subset \mathbb{Z}$. Here, the binary operation on the set of cosets is simply addition modulo n, and we see that the factor group $\mathbb{Z}/n\mathbb{Z}$ is just the cyclic group with n elements, justifying our earlier notation.

Example 2.12. Recall that the index of a subgroup is the number of cosets. If H is a subgroup of index 2, then the only cosets are H itself and the complement $G \setminus H$. Given $g \in G$, we have $gH = Hg = H$ if $g \in H$, and otherwise we have $gH = Hg = G \setminus H$. Thus H is normal, and the factor group G/H is isomorphic to $\mathbb{Z}/2\mathbb{Z}$.

Example 2.13. Consider the group S_3, which comprises symmetries of an equilateral triangle. It has four non-trivial subgroups, which are of two types.

First, there is the subgroup of rotations by a multiple of $\pi/3$. This subgroup has 3 elements and has index 2; hence it is normal.

Second, we may consider any reflection $r \in S_3$ (recall that there are three such reflections). Each of these has order 2, and so $\{\mathrm{Id}, r\}$ is a subgroup of order 2 and index 3. These subgroups are *not* normal; if we rotate by $\pi/3$, apply r, and then rotate back, we obtain not r, but the reflection through a different line of symmetry.

d. Permutation groups. We now use the tools from the previous two sections to study the symmetric groups S_n. These groups are highly non-abelian, and have many subgroups and inner automorphisms.

An element $\sigma \in S_n$ is a permutation of the set $\{1, \dots, n\}$. Consider the trajectory of 1 under repeated iterates of this permutation: $1, \sigma(1), \sigma^2(1), \dots$. Eventually we will return to 1; let k_1 be the smallest positive integer such that $\sigma^{k_1+1}(1) = 1$, so the elements $1, \sigma(1), \dots, \sigma^{k_1}(1)$ are all distinct. These elements compose a *cycle* of the permutation σ; let us call it X_1.

It may well happen that $X_1 \subsetneqq \{1, \dots, n\}$. In this case, we may choose $a \in \{1, \dots, n\} \setminus X_1$ and produce another cycle X_2 which contains all the iterates of a under the permutation σ. Continuing in this manner, we can decompose $\{1, \dots, n\}$ into disjoint sets $X_1, \dots, X_t$ such that σ acts cyclically on each X_i.

Example 2.14. Let $\sigma \in S_6$ be defined by the following table:

a	1	2	3	4	5	6
$\sigma(a)$	2	3	1	5	4	6

Then $X_1 = \{1, 2, 3\}$, $X_2 = \{4, 5\}$, and $X_3 = 6$. We may write the permutation σ in the following *cyclic notation*:

$$\sigma = (1\ 2\ 3)(4\ 5)(6) \tag{2.5}$$

This notation means that 1 is mapped to 2, which is in turn mapped to 3, which is in turn mapped to 1. Similarly, 4 and 5 are interchanged, and 6 is fixed.

We will usually ignore cycles X_i which have only a single element; this allows us to write (2.5) more compactly as

$$\sigma = (1\ 2\ 3)(4\ 5). \tag{2.6}$$

Furthermore, the point at which each cycle is started is arbitrary; thus (2.6) is equivalent to

$$\sigma = (3\ 1\ 2)(5\ 4),$$

and so on.

Definition 2.15. A permutation σ is *cyclic* if no more than one of the sets X_i has more than one element; that is, σ has only one cycle.

The notation (2.5) expresses the permutation in the example as a product of cyclic permutations, where these permutations are commuting, since their corresponding cycles are disjoint. The preceding discussion shows that *any* permutation can be expressed as a product of commuting cyclic permutations, and so we will use the notation (2.6) for elements of S_n.

What are the inner automorphisms of S_n? Fix $h \in S_n$; then the inner automorphism $I_h \colon \sigma \mapsto h^{-1}\sigma h$ corresponds to a relabeling of the elements of $\{1, \dots, n\}$. Indeed, if $h = (1\ 2)(3\ 4)$ is the permutation which exchanges 1 with 2 and 3 with 4, and if σ is the permutation in (2.6), then $I_h\sigma$ is the permutation

$$I_h\sigma = (2\ 1\ 4)(3\ 5).$$

Observe that σ and $I_h\sigma$ have the same cycle structure, but we have moved around the labels within the cycles according to the permutation h.

This suggests that if we are interested in intrinsic properties of a permutation σ which persist under automorphisms, then we should not look at the individual elements in the cycles which compose σ, but rather at the cycle structure as a whole. Thus we let $k_1(\sigma), \dots, k_{t(\sigma)}(\sigma)$ denote the lengths of the cycles $X_1, \dots, X_t$ for σ. For notational convenience, we always assume that $k_1 \geq k_2 \geq \cdots \geq k_t$.

Proposition 2.16. *Fix $\sigma, \sigma' \in S_n$. There exists an automorphism φ such that $\varphi(\sigma) = \varphi(\sigma')$ if and only if σ and σ' have the same cycle structure; that is, $t(\sigma) = t(\sigma')$ and $k_i(\sigma) = k_i(\sigma')$ for all i.*

Proof. If σ and σ' have the same cycle structure, then we find a suitable relabeling h such that $I_h\sigma = \sigma'$. The proof that every automorphism preserves cycle structure is left as an exercise. □

Definition 2.17. A *transposition* is a permutation which interchanges two elements and does nothing else. Using the notation of (2.6), it has the form $(a\ b)$ for some $a, b \in \{1, \dots, n\}$.

Proposition 2.18. *Every permutation can be written as a product of transpositions.*

Proof. Since every permutation can be written as a product of cyclic permutations, it suffices to show that every cyclic permutation is a product of transpositions. One way to accomplish this for the cyclic permutation $(1\ 2\ \cdots\ k)$ is as follows:

$$(1\ 2\ \cdots\ k) = (1\ 2)(1\ 3)\cdots(1\ k).$$

A similar technique works for any other cyclic permutation upon relabeling. □

Transpositions are in some sense the simplest possible permutations. If we think of the numbers $1, \dots, n$ as representing books on a shelf, then a permutation represents a rearrangement of the bookshelf. A transposition represents the very simple act of interchanging two books, and Proposition 2.18 states that we can accomplish any rearrangement, no matter how complicated, by repeatedly interchanging pairs of books.

Definition 2.19. A permutation is *even* if it can be written as the product of an even number of transpositions, and it is *odd* if it can be written as the product of an odd number of transpositions.

Proposition 2.18 guarantees that every permutation is either even or odd. However, we have *a priori* to acknowledge the possibility that some permutation σ may be both even and odd; perhaps σ can be written as the product of an even number of transpositions in one way, and as the product of an odd number of transpositions in another way. To show that this does not actually happen, we will prove the following lemma:

Lemma 2.20. *If σ is a transposition, then σ cannot be written as the product of an even number of transpositions.*

Lecture 3. Friday, September 4

a. Parity and the alternating group. Parity—the distinction between even and odd—is an important idea in many areas of mathematics. We will see later how this appears in geometry. It plays a prominent role in our study of the symmetric groups. Last time we defined the notion of parity for permutations on n elements, and promised to prove that this notion is in fact well defined—that a permutation σ cannot be simultaneously even and odd.

To this end, given any $\sigma \in S_n$, consider the number

$$N(\sigma) = |\{\{i, j\} \subset \{1, \dots, n\} \mid i < j, \sigma(i) > \sigma(j)\}|. \tag{3.1}$$

If we list the numbers from 1 to n and then permute them by the action of σ, the number $N(\sigma)$ is the number of pairs $\{i, j\}$ in the resulting list which appear in the "wrong" order.

Example 3.1. In S_4, let $\sigma = (2\ 3)$ and $\tau = (1\ 2\ 4)$. Then the resulting lists are as follows:

a	1	2	3	4
$\sigma(a)$	1	3	2	4
$\tau(a)$	2	4	3	1

Thus $N(\sigma) = 1$, since the pair $\{2, 3\}$ is backwards and all the other pairs are in the usual order, while $N(\tau) = 4$, since the pairs $\{2, 1\}$, $\{4, 3\}$, $\{4, 1\}$, and $\{3, 1\}$ are backwards.

Lemma 3.2. *Let $\sigma \in S_n$ be any permutation, and suppose σ can be written as the product of k transpositions. Then k and $N(\sigma)$ have the same parity ($k \equiv N(\sigma) \pmod 2$).*

Proof. Consider the following *basic transpositions* in S_n:

$$T_i = \begin{cases} (i\ i+1) & i < n, \\ (n1) & i = n. \end{cases}$$

Observe that an arbitrary transposition $(i\ j)$, where $i < j$, can be written as

$$\begin{aligned} (i\ j) &= (i\ i+1)(i+1\ i+2)\cdots(j-2\ j-1)(j-1\ j)(j-2\ j-1)\cdots(i+1\ i+2)(i\ i+1) \\ &= T_i T_{i+1} \cdots T_{j-2} T_{j-1} T_{j-2} \cdots T_{i+1} T_i. \end{aligned}$$

Thus every transposition can be written as a the product of an odd number of *basic* transpositions. We draw two conclusions from this:

(1) Every permutation $\sigma \in S_n$ can be written as a product of basic transpositions.
(2) If σ can be written as the product of an even number of transpositions, then it can be written as the product of an even number of basic transpositions. A similar fact holds with "even" replaced by "odd".

Without loss of generality, then, it suffices to prove that k and $N(\sigma)$ have the same parity whenever σ can be written as the product of k *basic* transpositions.

To see this, observe that if we take the product of some $\tau \in S_n$ with any basic transposition T_i, then the parity of the number of backwards pairs changes. That is, we permute the numbers $\{1, \dots, n\}$ by the action of τ, and then the number of backwards pairs in the resulting list is $N(\tau)$. The action of T_i swaps two neighbouring elements of this list; if this pair was correctly ordered before, it is now backwards, and vice versa, while every other pair is unchanged. Thus $N(\tau T_i) = N(\tau) \pm 1$.

It follows that if σ is the product of an even number of basic transpositions, then $N(\sigma)$ is even, and if σ is the product of an odd number of basic transpositions, then $N(\sigma)$ is odd. □

As a consequence of Lemma 3.2, a permutation σ is even if and only if $N(\sigma)$ is even, and odd if and only if $N(\sigma)$ is odd. It follows that every permutation is either even or odd, but not both. Furthermore, we see that the product of two even or two odd permutations is even and the product of one even and one odd permutation is odd. This fact can be reformulated by considering the map $P\colon S_n \to \mathbb{Z}/2\mathbb{Z}$ which takes σ to the residue class of $N(\sigma)$. This map is a surjective homomorphism. The kernel of P is the set of all even permutations, which is a subgroup of S_n with index 2. This is called the *alternating group* on n elements, and is denoted A_n. Since it has index 2, it follows from Example 2.12 that it is normal.

This can also be seen directly; suppose $\tau \in A_n$ can be written as a product of $2k$ transpositions and $\sigma \in S_n$ can be written as a product of ℓ transpositions. Then σ^{-1} is the produce of those same ℓ transpositions, but in the reverse order. It follows that $\sigma\tau\sigma^{-1}$ can be written as a product of $2k + 2\ell$ transpositions; hence it is in A_n.

In some sense, the alternating group A_n contains all the non-commutativity present in S_n. To make this statement more precise, we need to introduce some new notions in an attempt to quantify the degree of non-commutativity of a group.

b. Solvable groups. Setting aside the symmetric groups for the moment, let G be an arbitrary group. G may be abelian or non-abelian; in the latter case, we want to examine the degree to which G fails to be abelian.

To this end, observe that two elements a and b commute if and only if $ab = ba$, or equivalently, if $aba^{-1}b^{-1} = e$. The expression on the left-hand side of the equation is called the *commutator* of a and b, and is denoted

$$[a, b] = aba^{-1}b^{-1}. \tag{3.2}$$

We see that G is abelian if and only if $[a, b] = e$ for all $a, b \in G$.

Observe that the property of being a commutator is intrinsic, in the following sense: If φ is any automorphism of G, then

$$\varphi([a, b]) = \varphi(aba^{-1}b^{-1}) = \varphi(a)\varphi(b)\varphi(a)^{-1}\varphi(b)^{-1} = [\varphi(a), \varphi(b)].$$

That is, the property of being a commutator is invariant under any automorphism of G.

We would like to study the structure of G by looking at the set C of commutators. However, C itself may not be a subgroup—it may happen that for some $a, b, c, d \in G$, the product $[a, b][c, d]$ cannot be written in the form $[g, h]$ for any $g, h \in G$. Thus we consider the subgroup generated by all commutators:

$$[G, G] = \bigcap \{H \subset G \mid H \supset C \text{ is a subgroup}\}. \tag{3.3}$$

That is, $[G, G]$ is the smallest subgroup of G which contains the set of commutators in G; we refer to $[G, G]$ as the *commutator subgroup* of G, or as the *derived group*.

At one extreme, we have abelian groups, for which $[G, G] = \{e\}$; at the other extreme, suppose $[G, G] = G$. Then G is in some sense as non-abelian as it could possibly be; such a group is called *perfect*. One may reasonably ask if any perfect groups exist; we will see examples in a little while.

Remark. A closely related concept is that of a *simple* group—that is, a group with no non-trivial normal subgroups. Because the commutator subgroup is normal, we see that every abelian simple group is cyclic of prime order, and every non-abelian simple group is perfect.[3] Simple groups are in some sense irreducible, and a major theme in group theory is to build more complicated groups from simple groups.

The derived group $[G, G]$ may be thought of as the part of G which is left over when we strip away a small commutative part. (This is made precise by the statement that the factor group $G/[G, G]$ is abelian.) What happens if we apply the same procedure to $[G, G]$ and take its derived group?

To this end, let $G_0 = G$, and define G_n inductively by $G_n = [G_{n-1}, G_{n-1}]$. In this manner we obtain a sequence of subgroups

$$G = G_0 \supset G_1 \supset G_2 \supset \cdots . \tag{3.4}$$

If G is finite, this sequence must terminate somewhere; that is, there must exist n such that $G_{n+1} = G_n$. This follows since if $G_{n+1} \neq G_n$, then $|G_{n+1}| < |G_n|$, and there does not exist an infinitely long decreasing sequence of positive integers.

Now there are two possibilities. If G_n is the group in which the sequence (3.4) terminates, then either G_n is trivial (the single-element group) or it is not. If G_n is trivial, we say that the group G is *solvable*. If G_n is non-trivial, then since $G_n = G_{n+1} = [G_n, G_n]$, it is an example of a non-trivial perfect group.

Example 3.3. Let $G = S_3$ be the symmetric group on 3 elements. Then any commutator is an even permutation, hence $[G, G] \subset A_3$. The only even permutations on 3 elements are the identity permutation and the cyclic

[3] However, not every perfect group is simple.

permutations (1 2 3) and (1 3 2); it follows that A_3 is a cyclic group of order 3, and hence that $[G, G]$ is either trivial or equal to A_3. Because S_3 is non-abelian, $[G, G]$ is non-trivial; thus $G_1 = [S_3, S_3] = A_3$.

Now $A_3 = \mathbb{Z}/3\mathbb{Z}$ is abelian, so $G_2 = [A_3, A_3] = \{\mathrm{Id}\}$, and it follows that S_3 is solvable.

The argument in this example is emblematic of the so-called "soft" approach, which emphasises the use of general principles, rather than explicit computations. One could also give an explicit computation of the derived group of S_3 and then of the derived group of A_3, and avoid invoking general results such as the fact that a cyclic subgroup of prime order has no non-trivial subgroups. This latter approach—argument by explicit computation—stands in contradistinction to the "soft" approach, and is referred to as "hard". The distinction between "soft" and "hard" arguments applies in nearly every area of mathematics, not just algebra. Of course, most arguments lie somewhere in between the two extremes.

With a little more "hard" work, we can show that the alternating group is the derived group of the symmetric group for *any* value of n.

Proposition 3.4. *$[S_n, S_n] = A_n$ for every n.*

Proof. $n = 1$ and $n = 2$ are easy, since S_n is abelian and A_n is trivial. For $n \geq 3$, we first observe that $[S_n, S_n] \subset A_n$ since any commutator is even. To get the other inclusion, we begin with a statement on the generators of A_n.

Lemma 3.5. *Every element of A_n can be written as a product of (not necessarily disjoint) cyclic permutations of order* 3.

Proof. Given $\sigma \in A_n$, we can write σ as a product $\tau_1\tau_2 \cdots \tau_{2k}$, where each τ_i is a transposition, and $\tau_i \neq \tau_{i+1}$. Thus it suffices to write every product of two distinct transpositions as a product of cyclic permutations of order 3.

If the two transpositions share an element (for example, (1 2) and (1 3)), then their product is a cyclic permutation of order 3, and we are done.

Finally, observe that (1 2 3)(2 3 4) = (1 3)(2 4), and a similar computation obtains any product of two disjoint transpositions as a product of two cyclic permutations of order 3. □

Thanks to Lemma 3.5, it suffices to show that any cyclic permutation of order 3 can be obtained as a commutator. The following computation shows the general principle:

$$[(1\ 3), (1\ 2)] = (1\ 3)(1\ 2)(1\ 3)(1\ 2) = (1\ 2\ 3).$$

Every other cyclic permutation of order 3 can be obtained analogously; it follows that $[S_n, S_n]$ contains all cyclic permutations of order 3, and hence contains all even permutations. □

Example 3.6. We compute the groups G_k for $G = S_4$. Proposition 3.4 shows that $G_1 = [S_4, S_4] = A_4$, which is a non-abelian group with 12 elements. Aside from the identity element, these fall into two classes:

(1) Products of two disjoint transpositions; for example, (1 2)(3 4).
(2) Cyclic permutations of order 3; for example, (1 2 3).

It turns out (though we do not show it here) that a non-trivial element of A_4 can be obtained as a commutator $[\sigma, \tau]$, where $\sigma, \tau \in A_4$, if and only if it is the product of two disjoint transpositions. There are three such elements; together with the identity, they form a group isomorphic to the Klein four-group V, and so $G_2 = [A_4, A_4] = V$.

Finally, V is abelian, and so $G_3 = [V, V] = \{\mathrm{Id}\}$. Thus S_4 is solvable.

The present discussion seems quite abstract; aside from some vague notion of being able to "peel away all the non-commutativity", it is not clear just what we gain from knowing that a group is solvable. There are various possible replies to this concern. For the time being, we ignore the more abstract and general replies, and content ourselves with the following observation: Historically, group theory did not arise from a Bourbaki-esque drive towards abstraction, but rather from a specific problem—the solution of polynomial equations by radicals. Évariste Galois, who was the first to call groups "groups", made the following remarkable observation.

Theorem 3.7. *To any polynomial $f(x)$ of degree n there can be associated a subgroup $G_f \subset S_n$. The roots of the equation $f(x) = 0$ can be expressed using radicals if and only if G_f is a solvable group.*

Theorem 3.7 can be put more directly as the statement that the equation is solvable if and only if the corresponding group is solvable. In light of this, the fact that S_3 and S_4 (and hence all of their subgroups) are solvable corresponds to the fact that formulae can be found for the solutions of cubic and quartic equations. It turns out that for $n \geq 5$, the alternating group is perfect: $[A_n, A_n] = A_n$. Thus the symmetric group S_n is not solvable, and coupled with the fact that one can produce polynomials for which the corresponding group is S_n, this implies that polynomial equations with degree greater than 4 cannot be solved by radicals.

c. Nilpotent groups. If H and K are subgroups of a group G, then we can construct the commutator subgroup $[H, K]$; this is the smallest subgroup of G which contains all elements of the form $[h, k]$, where $h \in H$ and $k \in K$. This allows us to form another important series of subgroups of G: Set $\tilde{G}_0 = G$, and define $\tilde{G}_n$ recursively by $\tilde{G}_n = [\tilde{G}_{n-1}, G]$. Once again, if G is finite, the sequence must stabilise eventually; if it stablises by reaching the trivial group ($\tilde{G}_n = \{e\}$), we say that G is *nilpotent*.

It follows immediately from the definitions that $\tilde{G}_n \supset G_n$ for all n, and thus every nilpotent group is solvable. Since every abelian group has

$\tilde{G}_1 = \{e\}$ and hence is abelian, we have

$$\{\text{abelian groups}\} \subset \{\text{nilpotent groups}\} \subset \{\text{solvable groups}\}.$$

At the other end of the spectrum are the simple groups; in between the two extremes, we can do little more for the time being than proclaim, "Here there be dragons".

Example 3.8. Despite the fact that A_3 is abelian, its commutator with the entire symmetric group S_3 is non-trivial. In fact,

$$[(1\ 2), (1\ 2\ 3)] = (1\ 2)(1\ 2\ 3)(1\ 2)(3\ 2\ 1) = (1\ 2\ 3),$$

and so $[S_3, A_3] = A_3$. Thus S_3 is *not* nilpotent, and neither is S_4.

Nilpotent groups will play an important role in this course. Their key feature is that they are somehow close enough to being abelian that they are well understood; in particular, they can be classified in a reasonable way.

The distinction between nilpotent and solvable groups illustrates a trade-off which often occurs in mathematics. In choosing what structures we study, we must balance two goals; we want to consider general enough objects that our results will apply to a broad range of examples, but we must consider a specific enough class of objects that we can obtain meaningful results. As a class of objects to study, general abstract groups are far too broad to admit truly useful general results, and thus we restrict our attention to more specific classes. The class of nilpotent groups is small enough to be well understood, while the task of classifying the much larger class of solvable groups is more difficult.

CHAPTER 2

Symmetry in the Euclidean world: Groups of isometries of planar and spatial objects

Lecture 4. Wednesday, September 9

a. Groups of symmetries. So far we have been doing abstract algebra. Now it is time to throw a little geometry into the mix. As previously mentioned, one natural way to obtain a group is to consider the set of all bijections from a set X to itself that preserve a particular structure on X. Thus we now turn our attention to groups which arise as *symmetries* preserving a geometric structure.

In order to give a precise definition of a symmetry of X, we must first decide just what sort of object X is, and then decide just what geometric properties of X such a symmetry ought to preserve. For now, we return to the geometry of the ancient Greeks, and consider geometric bodies in either the Euclidean plane $\mathbb{R}^2$ or three-dimensional Euclidean space $\mathbb{R}^3$.

Recall that both $\mathbb{R}^2$ and $\mathbb{R}^3$ are equipped with a notion of distance given by the Pythagorean formula; the distance between two points $\mathbf{x} = (x_1, x_2)$ and $\mathbf{y} = (y_1, y_2)$ in $\mathbb{R}^2$ is

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

and similarly in $\mathbb{R}^3$. Indeed, the formula generalises to any Euclidean space $\mathbb{R}^d$, but for the time being we will only consider the cases $d = 2$ and $d = 3$.

Definition 4.1. Given $X \subset \mathbb{R}^d$, a map $f\colon X \to X$ is called *isometric* if

$$d(f(\mathbf{x}), f(\mathbf{y})) = d(\mathbf{x}, \mathbf{y}) \tag{4.1}$$

for every $\mathbf{x}, \mathbf{y} \in X$. If f is both isometric and a bijection, then f is called an *isometry* of X. The set of isometries of X forms a group under composition; this group is denoted $\operatorname{Isom}(X)$.

Of course, distance is only one particular geometric structure. Recall that if γ_1 and γ_2 are any two curves in $\mathbb{R}^d$ which intersect in a point $\mathbf{p}$, then the angle between the curves is defined to be the angle between their tangent vectors at $\mathbf{p}$. We could just as easily consider maps $f\colon X \to X$ which preserve angles, in the sense that the angle between $f(\gamma_1)$ and $f(\gamma_2)$ at $f(\mathbf{p})$ is equal to the angle between γ_1 and γ_2 at $\mathbf{p}$ for every two smooth curves. Such maps are called *conformal*.

Exercise 4.1. Show that every isometry is conformal.

The converse of the statement in Exercise 4.1 is not true; there are conformal maps f which are not isometries. An obvious example is the map $f\colon \mathbf{x} \mapsto \lambda\mathbf{x}$, where $\lambda > 0$ is fixed; such a map is called a *homothety* around the point $\mathbf{0}$ (one may also consider homotheties around other points, but the formulae are not as simple).

A composition of an isometry and a homothety is called a *similarity transformation*. As H. S. M. Coxeter points out in his *Introduction to Geometry*, similarity transformations are in fact the natural group of transformations to consider on the Euclidean plane, since there is no universal unit of length, and so any statement in Euclidean geometry which is true for a particular figure is also true for the image of that figure under an arbitrary similarity transformation.

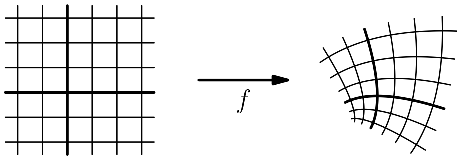


Figure 2.1. A conformal map which is not a similarity transformation.

Any similarity transformation is conformal. Less obviously, there are also conformal maps which are not similarity transformations; a geometric portrait of one such map is shown in Figure 2.1. The map f does not take lines to lines, but nevertheless it preserves angles. In fact, any holomorphic map $f\colon \mathbb{C} \to \mathbb{C}$ is conformal, and so there are many more conformal maps than there are similarity transformations or isometries.

We now have three different classes of symmetries of $\mathbb{R}^2$ (or $\mathbb{R}^d$, for that matter), which may be categorised according to which structures they preserve. The broadest class is the class of conformal maps, which preserve angles. Within the class of conformal maps we find the class of similarity transformations, which not only preserve angles, but map straight lines to straight lines.

Exercise 4.2. Show that if $f\colon \mathbb{R}^2 \to \mathbb{R}^2$ preserves angles and maps straight lines to straight lines, then f is a similarity transformation.

Finally, within the class of similarity transformations we find the class of isometries, which not only preserve angles and map straight lines to straight lines, but preserve distances as well. (In fact, the first two properties follow from this last fact.)

Angles, lines, distances... all of these are important geometric notions, and in fact, we may profitably study any of the three classes of transformations for various geometric objects X. For the time being, we will focus on the class of isometries as our symmetries of interest. Since this is the smallest of the three classes, we may reasonably expect that the study and classification of such symmetries will be simpler than for the other two classes.

b. Symmetries of bodies in $\mathbb{R}^2$. Given $X \subset \mathbb{R}^2$, we want to study the *symmetry group* of X—that is, the group of isometries $\mathrm{Isom}(X)$. At this point an issue of semantics arises: do we consider isometries f which are defined only on X (but not necessarily on the rest of $\mathbb{R}^2$), or do we restrict our attention to isometries of the entire Euclidean plane which happen to preserve X—for which $f(X) = X$? Whether there is a difference between the two approaches hinges on whether or not every isometry $f\colon X \to X$ can be uniquely extended to an isometry $F\colon \mathbb{R}^2 \to \mathbb{R}^2$.

In fact, every isometry can be so extended; however, this extension may not be unique. For example, suppose $X = \{\mathbf{p}\}$ is a single point, and consider the group

$$G_{\mathbf{p}} = \{f \in \mathrm{Isom}(\mathbb{R}^2) \mid f(\mathbf{p}) = \mathbf{p}\}.$$

Despite the fact that $\mathrm{Isom}(X)$ is a single element (since there is only one possible way to map a single point to itself), this single element extends to many different isometries of $\mathbb{R}^2$. Two sorts of extensions come immediately to mind:

(1) Rotations around $\mathbf{p}$. Given $\alpha \in \mathbb{R}$, let $R_\alpha\colon \mathbb{R}^2 \to \mathbb{R}^2$ denote the map which rotates the plane counterclockwise around $\mathbf{p}$ through an angle α.

(2) Reflections through lines containing $\mathbf{p}$. Given $\beta \in \mathbb{R}$, let $\ell_\beta \subset \mathbb{R}^2$ be the line through $\mathbf{p}$ that makes an angle β with the positive x-axis (equivalently, with the horizontal line from $\mathbf{p}$ to the right), and write L_β for the reflection in the line ℓ_β. Observe that $L_\beta = L_{\beta+\pi}$.

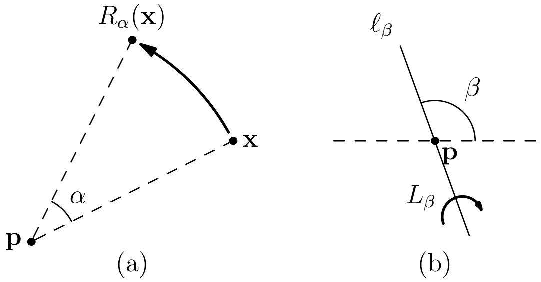


FIGURE 2.2. Isometries of $\mathbb{R}^2$ which fix $\mathbf{p}$.

Are these all the options? Clearly every R_α and L_β is in $G_{\mathbf{p}}$, and since $G_{\mathbf{p}}$ is a group under composition, it must also contain all the products of rotations and reflections.

PROPOSITION 4.2. *Given two reflections L_β and $L_{\beta'}$, let α be the angle between the corresponding lines—that is, the angle through which ℓ'_β must be rotated counterclockwise to reach ℓ_β—so $\alpha = \beta - \beta'$. Then $L_\beta \circ L_{\beta'} = R_{2\alpha}$.*

PROOF. The idea of the proof is shown in Figure 2.3, which illustrates the case $\beta' = 0$. Writing $\mathbf{y} = L_{\beta'}(\mathbf{x})$ and $\mathbf{z} = L_\beta(\mathbf{y})$, all one has to do is

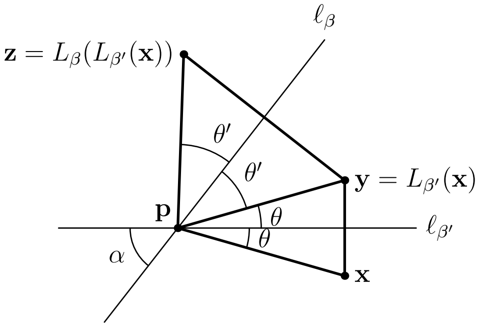


FIGURE 2.3. The product of two reflections is a rotation.

to observe that $d(\mathbf{x}, \mathbf{p}) = d(\mathbf{z}, \mathbf{p})$ and that the angle formed by the points $\mathbf{x}, \mathbf{p}, \mathbf{z}$ is equal to 2α. □

The product of two rotations is easy to handle: one sees immediately that $R_\alpha \circ R_{\alpha'} = R_{\alpha+\alpha'}$. It remains to determine what the product of a rotation and a reflection is.

EXERCISE 4.3. It turns out that the product of R_α and a reflection through a line ℓ is a reflection through the line ℓ' through $\mathbf{p}$ that makes an angle of $\alpha/2$ with ℓ.

(a) Prove this using geometric methods as in the proof of Proposition 4.2.
(b) Prove this using algebraic methods by observing that $R_\alpha \circ L_\beta = L_{\beta'}$ if and only if $R_\alpha = L_{\beta'} \circ L_\beta$ and applying the result of Proposition 4.2.

Thus the set of rotations around $\mathbf{p}$ and reflections in lines through $\mathbf{p}$ is closed under composition, and forms a subgroup of $G_{\mathbf{p}}$. In fact, it is the entire group.

In this case, the symmetries of X have extensions to $\mathbb{R}^2$ which are highly non-unique. This fact is peculiar to the case where X is a single point; we will show in the next lecture that the images of three non-collinear points determine an isometry, and so if X contains three such points, then every isometry of X has a unique extension to $\mathbb{R}^2$.

EXAMPLE 4.3. Let X be an equilateral triangle with centre $\mathbf{p}$. Then every symmetry of X fixes $\mathbf{p}$, and thus $\text{Isom}(X)$ is a subgroup of $G_{\mathbf{p}}$. We see that $R_\alpha(X) = X$ if and only if α is a multiple of $2\pi/3$, and that $L_\beta(X) = X$ if and only if ℓ_β is one of the three lines which connects a vertex of X to the midpoint of the opposite side. Thus

$$\text{Isom}(X) = \{\text{Id}, R_{2\pi/3}, R_{4\pi/3}, L_{\beta_1}, L_{\beta_2}, L_{\beta_3}\},$$

where β_i indicates the direction from $\mathbf{p}$ to the ith vertex, and the differences $\beta_i - \beta_j$ are multiples of $\pi/3$. Using the result of Proposition 4.2 and Exercise 4.3, one may easily verify that $\text{Isom}(X)$ is isomorphic to the symmetric group S_3; the same can be seen by labeling the vertices of the triangle and observing how they are permuted by each element of $\text{Isom}(X)$.

Example 4.4. Now we add a side and let X be a square. Once again, every symmetry of X fixes its centre $\mathbf{p}$, so $\text{Isom}(X)$ is again a subgroup of $G_{\mathbf{p}}$. This time the permissible rotation angles are multiples of $\pi/2$, and there are four possible reflections. Two of these reflections are through the lines connecting opposite vertices, and two are through lines connecting opposite midpoints.

Label the vertices clockwise with the numbers 1 through 4, we once again obtain a one-to-one homomorphism from $\text{Isom}(X)$ into the symmetric group S_4. This time, however, the homomorphism is not onto. For example, the permutation $(1\ 2)$ cannot be realised by any element of $\text{Isom}(X)$, as any symmetry of the square which interchanges vertices 1 and 2 must also interchange vertices 3 and 4. Instead, the isomorphic image of $\text{Isom}(X)$ is a subgroup of S_4 called the *dihedral group on* 4 *elements*, and denoted D_4.

Observe that since X has an even number of sides, the set of reflections can be partitioned into those which fix some vertices of the square and those which do not. The former (which are reflections through lines connecting opposite vertices) correspond to the elements $(1\ 3)$ and $(2\ 4)$ in S_4, while the latter correspond to $(1\ 2)(3\ 4)$ and $(1\ 4)(2\ 3)$.

In general, if X is a regular n-gon, then the isometry group of X contains n rotations and n reflections, and is isomorphic to the dihedral group $D_n \subset S_n$. We have $D_3 = S_3$, but since $n! > 2n$ for $n \geq 4$, the dihedral group is a proper subgroup for all larger values of n.

As n goes to infinity, the regular n-gons converge to the circle S^1, and we have $\text{Isom}(S^1) = G_{\mathbf{p}}$, since every isometry of $\mathbb{R}^2$ which fixes $\mathbf{p}$ also maps a circle centred at $\mathbf{p}$ to itself.

c. Symmetries of bodies in $\mathbb{R}^3$. Can we obtain the dihedral groups as symmetries of three-dimensional bodies? Let $X \subset \mathbb{R}^2 \subset \mathbb{R}^3$ be a regular polygon lying in the xy-plane (the horizontal plane). Then we see that in addition to the usual symmetries of X, we may also consider the composition of each such symmetry with reflection in the xy-plane. This reflection is an isometry and fixes X, thus isometries of X do not extend uniquely to isometries of $\mathbb{R}^3$; rather, each isometry of X has two possible extensions.

This ambiguity can be removed by considering a cone over X—that is, the set

$$\tilde{X} = \{(tx, ty, 1-t) \mid (x,y) \in X, 0 \leq t \leq 1\} \subset \mathbb{R}^3.$$

Then $\text{Isom}(\tilde{X})$ is isomorphic to a dihedral group, and every isometry of $\tilde{X}$ extends to a unique isometry of $\mathbb{R}^3$.

What about other three-dimensional bodies? The three-dimensional analogues of the regular polygons are the regular polyhedra—that is, polyhedra whose faces are all congruent regular polygons, and in which the same number of faces meet at each vertex. Such polyhedra are called *Platonic solids*, and there are exactly five of them: the tetrahedron, the cube, the octahedron, the icosahedron, and the dodecahedron. We examine the symmetry group of each of these in turn.

The tetrahedron. Let X be a tetrahedron, so the faces of X are triangles, and X has 4 faces, 6 edges, and 4 vertices. Labeling the vertices with the numbers 1 through 4, it is not difficult to check that any permutation of these numbers determines a unique isometry of X, and thus $\text{Isom}(X)$ is isomorphic to S_4.

The cube. Let X be a cube, so the faces of X are squares, and X has 6 faces, 12 edges, and 8 vertices. By labeling the vertices, we may find an isomorphic image of $\text{Isom}(X)$ in S_8; however, it is nowhere close to being the whole group, and so to understand the structure of $\text{Isom}(X)$ we turn to a more geometric approach. Namely, we observe that as in the two-dimensional case, every isometry of a cube centred at $\mathbf{p}$ must fix $\mathbf{p}$, and that the isometries of $\mathbb{R}^3$ which fix $\mathbf{p}$ are precisely the isometries of the sphere S^2 centred at $\mathbf{p}$. Thus $\text{Isom}(X)$ is a subgroup of $\text{Isom}(S^2)$.

The isometries of $\mathbb{R}^3$ can be divided into two classes. Recall that a basis $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ of $\mathbb{R}^3$ satisfies the *right-hand rule*, or is *right-handed*, if pointing your right index finger in the direction of $\mathbf{v}_1$ and your right middle finger in the direction of $\mathbf{v}_2$ results in your right thumb pointing in the direction of $\mathbf{v}_3$. We say that $f \in \text{Isom}(\mathbb{R}^3)$ is *orientation-preserving*, or *even*, if the image of a right-handed basis is right-handed; otherwise f is *orientation-reversing*, or *odd*.

We will see later on that every orientation-preserving (even) isometry of $\mathbb{R}^3$ which fixes a point $\mathbf{p}$ is a rotation around some axis ℓ which passes through $\mathbf{p}$; furthermore, every odd isometry can be obtained as the product of an even isometry with the map $\mathbf{x} \mapsto -\mathbf{x}$. Thus to understand $\text{Isom}(X)$, it suffices to understand the subgroup of even isometries of the cube, which we denote $\text{Isom}^+(X)$.

So we ask: What lines ℓ through $\mathbf{p}$ can be axes for rotations which map the cube to itself? We list the possibilities and count the number of (non-trivial) rotations they yield.

(1) If ℓ passes through the centres of two opposite faces, then rotation by any multiple of $\pi/2$ around ℓ is an isometry of the cube. Thus each axis yields 3 non-trivial rotations, and since there are 3 such axes, $\text{Isom}^+(X)$ contains 9 rotations which are generated by elements of order 4. 6 of these rotations have order 4 and the other 3 have order 2.

(2) If ℓ passes through the midpoints of two opposite edges, then rotation by π around ℓ is an isometry of the cube. Thus each axis yields only 1 non-trivial rotation, and since there are 6 such axes, we have found 6 more elements of order 2.

(3) If ℓ passes through two opposite vertices, then rotation by $2\pi/3$ around ℓ is an isometry of the cube (the rotation acts cyclically upon the 3 edges which meet this vertex). Thus each axis yields 2 non-trivial rotations, and since there are 4 such axes, $\text{Isom}^+(X)$ contains 8 rotations of order 3.

Adding everything up and remembering to include the identity, we see that the number of elements of $\text{Isom}^+(X)$ is

$$9 + 6 + 8 + 1 = 24.$$

So far we have met three groups of order 24: the symmetric group S_4, the dihedral group D_{12}, and the cyclic group $\mathbb{Z}/24\mathbb{Z}$. Is $\text{Isom}^+(X)$ one of these? Or is it something else entirely?

Since every element of $\text{Isom}^+(X)$ has order ≤ 4, the symmetry group of the cube cannot be $\mathbb{Z}/24\mathbb{Z}$ or D_{12}, as both of these contain elements of higher order. So perhaps it is S_4. To show this, we ought to find some geometric features of the cube which are permuted in an arbitrary way by the isometries of X.

If we label the main diagonals of the cube (the ones which pass through $\mathbf{p}$ and which connect opposite vertices) with the numbers 1 to 4, we find that this is exactly the case. This labeling yields the desired isomorphism.

The octahedron. It seems that life is becoming more difficult as we progress to the more involved Platonic solids. So we might expect that the symmetry group of the octahedron, which has 8 equilateral triangles as faces, and which has 12 edges and 6 vertices, might be even more troublesome to compute.

Fortunately for us, this turns out not to be the case. In fact, we have already answered this question! Observe that the octahedron is the dual polyhedron to the cube—that is, if we construct a polyhedron which has one vertex at the centre of each face of the cube, we obtain an octahedron, and vice versa. Thus the isometries of the octahedron are exactly the isometries of the cube, and we see that once again the group of even isometries is S_4. Notice that for tetrahedron the construction gives nothing new since the dual to it is another tetrahedron.

The icosahedron and dodecahedron. Armed with the realisation that dual polyhedra have the same isometry group, we can treat the icosahedron (20 triangular faces, 30 edges, and 12 vertices) and the dodecahedron (12 pentagonal faces, 30 edges, and 20 vertices) in one fell swoop. The technique is the same as for the cube; every even isometry must be a rotation, and we can classify the possible axes of rotation according to whether they connect vertices, midpoints of edges, or centres of faces.

Using the same approach as in the case of the cube we can count even isometries of the dodecahedron. There are four rotations by the multiples of $2\pi/5$ around each of the 6 axis connecting the centers of opposite faces, the rotation by π around each of 15 axis connecting midpoints of the pair of opposite (symmetric with respect to the center) edges, and two rotations by multiples of $2\pi/3$ around each of 10 principal diagonals adding the identity we obtain $24 + 20 + 15 + 1 = 60$ even isometries.

Icosahedron is dual to the dodecahedron so the symmetry groups are the same.

Since, the number of odd isometries is equal to the number of even isometries we obtain the total of 48 isometries for the cube/octahedron and 120 for the dodecahedron/icosahedron.

Exercise 4.4. List all odd isometries of tetrahedron, cube and dodecahedron.

It is interesting to understand algebraic nature of the groups thus obtained. Notice that both the cube and the dodecahedron are centrally symmetric and the central symmetry $\mathbf{x} \mapsto -\mathbf{x}$ commutes with all other isometries. This immediately implies that the full isometry group of the cube is $S_4 \times \mathbb{Z}/2\mathbb{Z}$. For the dodecahedron the group of even isometries has 60 elements, as many as the alternating group A_5. We will see later that, similarly to the cube, we can find five elements rigidly related to the dodecahedron such that every even permutation of these elements determines an even isometry. Thus the group of even isometries of the dodecahedron is isomorphic to A_5 and the full isometry group to $A_5 \times \mathbb{Z}/2\mathbb{Z}$ but not to S_5! Finally notice that S_4, the full isometry groups of the tetrahedron is not isomorphic to $A_4 \times \mathbb{Z}/2\mathbb{Z}$, an algebraic counterpart of the fact that tetrahedron is not centrally symmetric.

Notice also that we have $D_n \times \mathbb{Z}/2\mathbb{Z}$ as the symmetry group of a regular n-gon in $\mathbb{R}^3$ or of the rectangular prism based on such a polygon.

d. Isometries of the plane. There is more to life than regular polygons and polyhedra. However, we will not discover any new symmetries by considering other bounded figures made up of lines and polygons, as the examples we have already considered are the most symmetric piecewise linear objects in $\mathbb{R}^2$ and $\mathbb{R}^3$. In the case of the plane, the statement that every finite group of isometries is either cyclic or dihedral (attributed to Leonardo da Vinci, according to Hermann Weyl) will be proved in the next lecture after the basic properties of planar isometries are established; for three-dimensional space it will take us a bit longer.

For this and other purposes we consider more systematically the isometry group of the Euclidean plane, $\text{Isom}(\mathbb{R}^2)$.

We have already encountered rotations (which are even) and reflections (which are odd). These are all the isometries of $\mathbb{R}^2$ which fix a point; if we consider isometries with no fixed point, we immediately find the *translations*. To each $\mathbf{v} \in \mathbb{R}^2$ is associated the translation $T_{\mathbf{v}} \colon \mathbf{x} \mapsto \mathbf{x} + \mathbf{v}$. Less immediately evident are the *glide reflections*; a glide reflection is the product of a translation $T_{\mathbf{v}}$ and reflection in the line parallel to $\mathbf{v}$.

Is this it? Does every isometry of the plane fall into one of these four classes? We will begin the next lecture by showing that it does, and indeed these four classes can be neatly organised according to whether they are even or odd, and whether or not they have a fixed point:

	even	odd
fixed point	rotations	reflections
no fixed point	translations	glide reflections

Each of the four classes is closed under conjugation; that is, if f and g are isometries of $\mathbb{R}^2$, then $f \circ g \circ f^{-1}$ is in the same class as g is. Thus a necessary condition for two isometries to be conjugate is that they be in the same class.

For reflection, this necessary condition is also sufficient. We will show soon that any two reflections are conjugate.

In contrast, the even isometries—rotations and translations—have intrinsic invariants. For rotations this is the absolute value of the angle of rotation (but not the centre of rotation!)—two rotations are conjugate if and only if their angle of rotation is the same or opposite. For translations this is the distance a point is moved (but not the direction!)—two translations $T_{\mathbf{v}}$ and $T_{\mathbf{w}}$ are conjugate if and only if $\mathbf{v}$ and $\mathbf{w}$ have the same length. For he glide reflections the situation is similar to translations; they are conjugate if and only if the translational parts have the same length.

Remark. Our discussion here has taken the *synthetic* approach to geometry; that is, we have used the sorts of axiomatic arguments which date back to Euclid, eschewing coordinate systems or any other more modern tools. We will later see that we can gain further insight into the situation by using the tools of linear algebra, which will allow us to make use of the fact that isometries and linear maps are related, and from insights about the group structure of the latter.

On the whole, the algebraic approach is more algorithmic, it allows to prove things by more or less routine calculations, while synthetic one is more elegant and also "invariant" since it does not use auxiliary tools extrinsic to the Euclidean structure, such as fixing an origin or a coordinate system.

Lecture 5. Friday, September 11

a. Even and odd isometries. Staying with the synthetic approach for a little while longer, we now establish the basic properties of isometries of $\mathbb{R}^2$ which are needed to complete the treatment in the previous lecture. In particular, we prove the classification results stated there. Although our methods are geometric, we focus on the algebraic structure of $\mathrm{Isom}(\mathbb{R}^2)$, and use it to our advantage when possible.

The first property which we use to classify isometries is orientation. Given a point $\mathbf{p} \in \mathbb{R}^2$, there are two directions in which we can rotate something around $\mathbf{p}$: clockwise and counterclockwise. The choice of which of these is the positive direction of rotation and which is negative determines an *orientation* of the plane. By convention, we usually choose counterclockwise as the positive direction, and so clocks run in the negative direction (this turns out not to be a terribly effective method of time travel, though).

An isometry f is *orientation-preserving*, or *even*, if it preserves the positive and negative directions of rotation, and *orientation-reversing*, or *odd*, otherwise. We may think of this as follows: if C is a clock lying in $\mathbb{R}^2$ centred at $\mathbf{p}$, then $f(C)$ is a clock centred at $f(\mathbf{p})$. If C and $f(C)$ run in the same direction, then f is even; if they run in opposite directions, then f is odd.

Even and odd isometries may also be defined in terms of basis vectors, as suggested in the previous lecture for isometries of $\mathbb{R}^3$.

It is easy to see that the rules for composing even and odd isometries are the same as the rules for adding even and odd numbers; the product of two even isometries is even, the product of one even and one odd isometry is odd, and the product of two odd isometries is even. In particular, the set of even isometries forms a subgroup of $\mathrm{Isom}(\mathbb{R}^2)$, which we denote $\mathrm{Isom}^+(\mathbb{R}^2)$.

A more formal way of stating the above remark is the following: we may define a homomorphism $P\colon \mathrm{Isom}(\mathbb{R}^2) \to \mathbb{Z}/2\mathbb{Z}$ by

$$P(f) = \begin{cases} 2\mathbb{Z} & f \text{ is orientation-preserving,} \\ 1 + 2\mathbb{Z} & f \text{ is orientation-reversing,} \end{cases}$$

where we recall that $2\mathbb{Z}$ is the subgroup of even numbers, and $1 + 2\mathbb{Z}$ is the coset of odd numbers. Then $\ker(P) = \mathrm{Isom}^+(\mathbb{R}^2)$, and we see that the subgroup of even isometries has just two cosets—itself and its complement. Thus despite the fact that $\mathrm{Isom}^+(\mathbb{R}^2)$ and $\mathrm{Isom}(\mathbb{R}^2)$ are both infinite groups, the former is a subgroup of finite index—in fact, index 2. It follows from the remarks in Example 2.11 that $\mathrm{Isom}^+(\mathbb{R}^2)$ is a normal subgroup of $\mathrm{Isom}(\mathbb{R}^2)$.

b. Isometries are determined by the images of three points. With one fundamental tool in hand, we now turn to another question which arises in the classification of isometries—indeed, of any sort of mathematical object. If f is an isometry of $\mathbb{R}^2$, how much do we have to know about f to determine it completely?

Of course if we know the image of every point in $\mathbb{R}^2$ then we have determined f completely, and for an arbitrary map this would be necessary. However, f has extra structure—it is an isometry—and so we may hope to get away with less, and avoid having to specify uncountably many pieces of information. Ideally, we would like to determine f uniquely via only finitely many pieces of information; we begin by observing that this is quite easy to do if we go down a dimension and consider isometries of the real line.

Note that even and odd isometries of $\mathbb{R}$ can be distinguished according to whether or not they preserve the ordering of the real numbers. That is, $f \in \text{Isom}(\mathbb{R})$ is even if and only if $f(x) < f(y)$ whenever $x < y$.

Proposition 5.1. *Given $x, x' \in \mathbb{R}$, there are exactly two isometries which map x to x'. One of these isometries is even and the other is odd.*

Proof. Suppose $f \in \text{Isom}(\mathbb{R})$ is such that $f(x) = x'$. Given an arbitrary $y \in \mathbb{R}$, observe that

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

and hence $f(y) = f(x) \pm (y - x)$. Thus there are two possibilities for $f(y)$; one is greater than $f(x)$, the other is less than $f(x)$. One corresponds to an even isometry, the other to an odd isometry. Writing

$$\begin{aligned} f_E(y) &= y + (f(x) - x), \\ f_O(y) &= -y + (f(x) + x), \end{aligned}$$

we see that f_E and f_O are even and odd, respectively, and that these are the only two isometries of $\mathbb{R}$ which map x to x'. □

Remark. The proof of Proposition 5.1 also shows that every even isometry of $\mathbb{R}$ is a translation, and every odd isometry is a reflection. Observe that a translation $y \mapsto y + a$ can be obtained by first reflecting around the origin 0 and then reflecting around $a/2$ (or indeed, around any two points such that the second lies a distance $a/2$ to the right of the first). Thus every translation is the product of two reflections, and so the set of reflections generates the group $\text{Isom}(\mathbb{R})$, just as the set of transpositions generates the symmetric group S_n.

It follows from Proposition 5.1 that every isometry of $\mathbb{R}$ is determined by just two pieces of information: the image of a single point and the parity of the isometry. If we know the images of two points, then we can determine the parity, and thus the images of two points suffice to uniquely determine an isometry.

In the plane, we have an extra dimension to work with, and so we expect to require more information. It turns out that one more piece of information is all we need; the images of any two points are enough to uniquely determine an isometry of $\mathbb{R}^2$ up to parity, and so the images of three non-collinear points determine an isometry uniquely.

PROPOSITION 5.2. *Let* $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^2$ *be non-collinear, and suppose* $\mathbf{x}', \mathbf{y}', \mathbf{z}' \in \mathbb{R}^2$ *are such that*

$$\begin{aligned} d(\mathbf{x}', \mathbf{y}') &= d(\mathbf{x}, \mathbf{y}) = d_1, \\ d(\mathbf{x}', \mathbf{z}') &= d(\mathbf{x}, \mathbf{z}) = d_2, \\ d(\mathbf{y}', \mathbf{z}') &= d(\mathbf{y}, \mathbf{z}) = d_3. \end{aligned} \tag{5.1}$$

Then there exists a unique isometry $I \colon \mathbb{R}^2 \to \mathbb{R}^2$ *such that* $I\mathbf{x} = \mathbf{x}'$, $I\mathbf{y} = \mathbf{y}'$, *and* $I\mathbf{z} = \mathbf{z}'$.

PROOF. We "build up" an isometry $I \in \mathrm{Isom}(\mathbb{R}^2)$ which has the specified action on $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$. Let T be the translation which takes $\mathbf{x}$ to $\mathbf{x}'$. Then $T\mathbf{y}$ and $\mathbf{y}'$ both lie on the circle centred at $\mathbf{x}'$ with radius d_1; let R be the rotation around $\mathbf{x}'$ which takes $T\mathbf{y}$ to $\mathbf{y}'$.

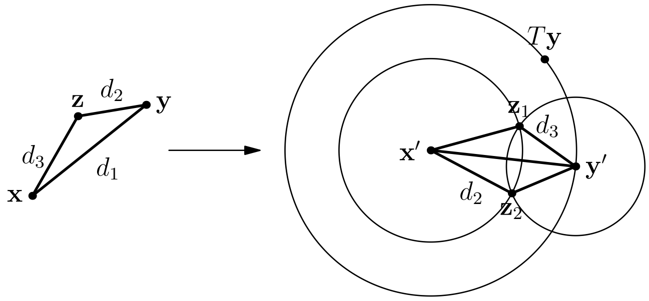


FIGURE 2.4. Images of three points determine an isometry.

Now $R \circ T$ is an isometry; hence $R \circ T(\mathbf{z})$ must lie on the circle centred at $\mathbf{x}'$ with radius d_2 and also on the circle centred at $\mathbf{y}'$ with radius d_3. These circles intersect in just two points, $\mathbf{z}_1$ and $\mathbf{z}_2$ (see Figure 2.4). Since $R \circ T$ is orientation-preserving, we have $R \circ T(\mathbf{z}) = \mathbf{z}_1$. Let L be reflection in the line through $\mathbf{x}'$ and $\mathbf{y}'$; then we have $L \circ R \circ T(\mathbf{z}) = \mathbf{z}_2$.

It follows from (5.1) that $\mathbf{z}'$ is either $\mathbf{z}_1$ or $\mathbf{z}_2$. We have exhibited one isometry which fulfills the former case ($R \circ T$) and one which fulfills the latter ($L \circ R \circ T$). This proves existence, and it remains to show uniqueness.

Let I be any isometry which takes $\mathbf{x}$ to $\mathbf{x}'$, $\mathbf{y}$ to $\mathbf{y}'$, and $\mathbf{z}$ to $\mathbf{z}'$. Then given any point $\mathbf{a} \in \mathbb{R}^2$, the image $I\mathbf{a}$ must lie on the circle centred at $\mathbf{x}'$ with radius $d(\mathbf{a}, \mathbf{x})$, and similarly for $\mathbf{y}'$ and $\mathbf{z}'$. These three circles intersect in exactly one point since their centres are not collinear, and so there is only one possibility for $I\mathbf{a}$. □

c. Isometries are products of reflections. The proof of Proposition 5.2 shows that every isometry of $\mathbb{R}^2$ can be written as the product of a rotation and a translation (if it is orientation-preserving) or of a rotation, a translation, and a reflection (if it is orientation-reversing).

We saw in the previous lecture that every rotation can be written as a product of two reflections. The same is true of translations; this was mentioned for isometries of $\mathbb{R}$ in the remark after Proposition 5.1, and is true in $\mathbb{R}^2$ as well (and indeed, in every $\mathbb{R}^d$). This may easily be seen by considering the composition of two reflections in parallel lines.

It follows, then, that every isometry of $\mathbb{R}^2$ can be written as a product of no more than five reflections. In fact, we can do even better than this.

Proposition 5.3. *Every isometry of $\mathbb{R}^2$ can be written as a product of no more than three reflections.*

Proof. Given an arbitrary isometry I, it suffices to consider the action of I on three non-collinear points $\mathbf{x}, \mathbf{y}, \mathbf{z}$, as in Proposition 5.2. Let ℓ_1 be the perpendicular bisector of the line segment from $\mathbf{x}$ to $I\mathbf{x}$, and let L_1 be reflection in ℓ_1. Observe that $L_1\mathbf{x} = I\mathbf{x}$.

Now let ℓ_2 be the perpendicular bisector of the line segment from $L_1\mathbf{y}$ to $I\mathbf{y}$. Since both these points are an equal distance from $I\mathbf{x}$, we see that $I\mathbf{x} \in \ell_2$. Let L_2 be reflection in ℓ_2, and observe that $L_2L_1\mathbf{y} = I\mathbf{y}$ and $L_2L_1\mathbf{x} = \mathbf{x}$.

Now as in the proof of Proposition 5.2, one of the following two things happens.

(1) I is even, in which case $I\mathbf{z} = L_2L_1\mathbf{z}$, and we have $I = L_2 \circ L_1$.
(2) I is odd, in which case $I\mathbf{z} = L_3L_2L_1\mathbf{z}$, where L_3 is reflection in the line ℓ_3 through $I\mathbf{x}$ and $I\mathbf{y}$, and we have $I = L_3 \circ L_2 \circ L_1$. □

Remark. Similar results to Propositions 5.2 and 5.3 are available in higher dimensions. For example, the same method of proof shows that every isometry of $\mathbb{R}^3$ can be written as a product of at most four reflections, and is uniquely determined by its action on four non-coplanar points.

We can use the result of Proposition 5.3 to provide a proof of the classification given at the end of the previous lecture. That is, we show that every even isometry of $\mathbb{R}^2$ is either a rotation (if it has a fixed point) or a translation (if it does not), and that every odd isometry is either a reflection (if it has a fixed point) or a glide reflection (if it does not).

The even isometries are easier to deal with, since they can be written as the product of only two reflections. Let $I = L_2 \circ L_1$, where L_i is reflection in the line ℓ_i. If ℓ_1 and ℓ_2 intersect, then I is a rotation around their point of intersection; if ℓ_1 and ℓ_2 are parallel, then I is a translation.

The odd isometries are trickier, since there are many possible configurations of three lines in $\mathbb{R}^2$. Nevertheless, we can reduce everything to the case of a translation composed with a reflection. To do this, let $I = L_3 \circ L_2 \circ L_1$. If ℓ_1 and ℓ_2 are parallel, then $L_2 \circ L_1$ is a translation, and so I has the form $L \circ T$, where L is a reflection and T is a translation.

If ℓ_1 and ℓ_2 are not parallel, then $R = L_2 \circ L_1$ is a rotation by 2θ, where θ is the angle between ℓ_1 and ℓ_2. Observe that R can be decomposed as a product of two reflections in many different ways, as shown in Figure 2.5—ℓ_2' can be chosen arbitrarily, provided ℓ_1' is chosen to make the angle between the two lines equal to θ. Thus for any such ℓ_1' and ℓ_2' we obtain $I = L_3 \circ$

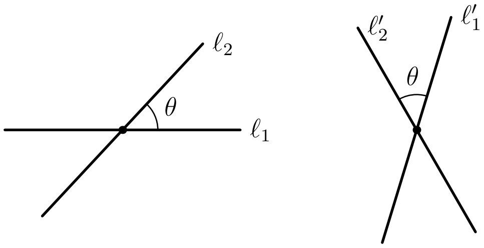


FIGURE 2.5. Decomposing a rotation as the product of two reflections in different ways.

$L_2' \circ L_1'$. In particular, we can choose ℓ_2' parallel to ℓ_3, so that $T = L_3 \circ L_2'$ is a translation, and I has the form $T \circ L_1'$.

Thus every odd isometry of $\mathbb{R}^2$ can be written as the product of a reflection and a translation. Let $T_{\mathbf{v}}$ be translation by a vector $\mathbf{v}$, and let L be reflection in a line ℓ. Decompose $\mathbf{v}$ as $\mathbf{v}_1 + \mathbf{v}_2$, where $\mathbf{v}_1$ is parallel to ℓ and $\mathbf{v}_2$ is orthogonal to ℓ. Then we have

$$T_{\mathbf{v}} = T_{\mathbf{v}_1} \circ T_{\mathbf{v}_2} = T_{\mathbf{v}_2} \circ T_{\mathbf{v}_1}.$$

Observe that since $\mathbf{v}_2$ and ℓ are perpendicular, $T_{\mathbf{v}_2} \circ L$ is reflection in a line ℓ' which is parallel to ℓ (in fact, $\ell' = T_{\mathbf{v}_2/2}\ell$). Similarly, $L \circ T_{\mathbf{v}_2}$ is reflection in a line ℓ'' parallel to ℓ.

It follows that $L \circ T = L \circ T_{\mathbf{v}_2} \circ T_{\mathbf{v}_1} = L' \circ T_{\mathbf{v}_1}$, and similarly for $T \circ L$. Thus every odd isometry I can be written as the product of a reflection around a line ℓ' and a translation by a vector $\mathbf{v}_1$ parallel to ℓ'. If $\mathbf{v}_1 = \mathbf{0}$, then I is a reflection, otherwise I is a glide reflection. This completes the classification of elements of $\mathrm{Isom}(\mathbb{R}^2)$.

Notice that glide reflections appear as products of commuting pairs of reflections and translations. Among three basic types of isometries, translations, rotations and reflections, this is the only case when two representatives of difference classes may commute.

d. Isometries in $\mathbb{R}^3$. A similar approach can be used to classify the isometries of three-dimensional Euclidean space. Once again, $\mathrm{Isom}(\mathbb{R}^3)$ is generated by the set of reflections; as before, a reflection is determined by its set of fixed points, but now this set is a plane instead of a line. Two planes in $\mathbb{R}^3$ are either parallel or intersect in a line ℓ. In the first case, the product of the corresponding reflections is a translation; in the second case, it is a rotation around ℓ.

It follows that translations and rotations generate $\mathrm{Isom}^+(\mathbb{R}^3)$. This was also true of $\mathrm{Isom}^+(\mathbb{R}^2)$, but in a somewhat trivial way, as *every* orientation-preserving isometry of the plane is either a translation or rotation. New types of isometries appear as products of commuting pairs of basic isometries. Unlike the planar case, in $\mathbb{R}^3$ there are such commuting pairs: if R is a rotation around an axis ℓ and T is a translation by a vector parallel to

ℓ, then $R \circ T = T \circ R$ is an isometry which is neither a translation nor a rotation, but something new.

There is also a new kind of orientation-reversing isometries. Following the same principle as before, one should look for commuting pairs of reflections with basic orientation preserving isometries. We still have a glide reflection as the product of a reflection and a translation along a vector parallel of the plane of reflection. In addition, if R is a rotation around an axis ℓ and L is reflection in a plane orthogonal to ℓ, then $L \circ R = R \circ L$ is an orientation-reversing isometry which is neither a reflection nor a glide reflection. In the particular case where R is rotation by π, we obtain the *central symmetry* $\mathbf{x} \mapsto -\mathbf{x}$.

We will show in due time, using a synthetic method based on representation of isometries as products of reflections, that every isometry in $\mathbb{R}^3$ belongs to one of the six types described above. This method however becomes too cumbersome when dimension goes up. In order to give a comprehensive classification of isometries in $\mathbb{R}^n$ we will resort to linear algebra instead.

e. The group structure of $\mathrm{Isom}(\mathbb{R}^2)$**.** So far we have analysed the structure of the *set* of isometries of the plane by providing a complete classification. We now turn our attention to the structure of the *group* of isometries, which we touched upon in our observation that this group is generated by the set of reflections.

We begin by considering $\mathrm{Isom}^+(\mathbb{R}^2)$, which is a normal subgroup of index two, as already noted. Let $\mathcal{T}$ denote the set of all translations; it is easy to see that $\mathcal{T}$ is a subgroup of $\mathrm{Isom}^+(\mathbb{R}^2)$.

PROPOSITION 5.4. *$\mathcal{T}$ is normal in* $\mathrm{Isom}^+(\mathbb{R}^2)$.

PROOF. It suffices to check that if T is a translation, then so is $R \circ T \circ R^{-1}$, where R is a rotation. To see this, we first define some notation.

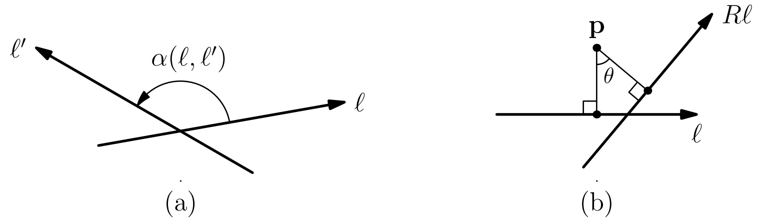


FIGURE 2.6. Characterising orientation-preserving isometries.

Consider two lines ℓ and ℓ', and suppose that ℓ and ℓ' have been "marked" with a positive direction (indicated in Figure 2.6(a) with an arrow). If ℓ and ℓ' are parallel or coincide, write $\alpha(\ell, \ell') = 0$; otherwise, write $\alpha(\ell, \ell')$ for the angle from the positive direction of ℓ to the positive direction of ℓ'.

If T is a translation, then $\alpha(\ell, T\ell) = 0$ for all lines ℓ. If R is a rotation by θ, then $\alpha(\ell, R\ell) = \theta$ for all marked lines ℓ (this follows from basic geometric arguments—see Figure 2.6(b)). Thus if I is any orientation-preserving isometry, we may define

$$\alpha(I) = \alpha(\ell, I\ell)$$

by choosing an arbitrary marked line ℓ. $\alpha(I)$ is defined up to multiples of 2π, and in fact, one may easily show that $\alpha\colon \mathrm{Isom}^+(\mathbb{R}^2) \to \mathbb{R}/2\pi\mathbb{Z}$ is a homomorphism. Now we see that $\mathcal{T} = \ker \alpha$, and since the kernel of any homomorphism is a normal subgroup, the proposition follows. □

Given a normal subgroup, we can take a factor group; what is the factor group $\mathrm{Isom}^+(\mathbb{R}^2)/\mathcal{T}$? It is not hard to see that this factor group is isomorphic to the group of rotations around a given point $\mathbf{p}$. Such rotations are in a bijective correspondence with the real numbers modulo 2π, and this correspondence is homomorphic; thus $\mathrm{Isom}^+(\mathbb{R}^2)/\mathcal{T}$ is isomorphic to $S^1 = \mathbb{R}/2\pi\mathbb{Z}$.

In fact, the isomorphism is provided by the map α which was constructed in the proof of Proposition 5.4; this is an example of the general principle that $G/\ker\varphi$ is isomorphic to $\mathrm{Im}\,\varphi$ for *any* homomorphism φ.

We can now make a rather strong statement about the group of even isomorphisms of the plane.

Proposition 5.5. $\mathrm{Isom}^+(\mathbb{R}^2)$ *is solvable.*

Proof. We compute the sequence of derived subgroups and show that it terminates in the trivial group. Given any two even isometries I_1 and I_2, we have

$$\alpha([I_1, I_2]) = \alpha(I_1 I_2 I_1^{-1} I_2^{-1}) = \alpha(I_1) + \alpha(I_2) - \alpha(I_1) - \alpha(I_2) = 0.$$

It follows that $[\mathrm{Isom}^+(\mathbb{R}^2), \mathrm{Isom}^+(\mathbb{R}^2)] \subset \mathcal{T}$.

Exercise 5.1. Show that $[\mathrm{Isom}^+(\mathbb{R}^2), \mathrm{Isom}^+(\mathbb{R}^2)] = \mathcal{T}$.

Now $\mathcal{T}$ is isomorphic to $\mathbb{R}^2$ via the map $\mathbf{v} \mapsto T_{\mathbf{v}}$, and so $\mathcal{T}$ is abelian. Thus $[\mathcal{T}, \mathcal{T}]$ is trivial, and we are done. □

Remark. Proposition 5.5 fails in higher dimensions; $\mathrm{Isom}^+(\mathbb{R}^3)$ is *not* solvable. This reveals a deep difference between isometries of the plane and isometries of three-dimensional space.

Remark. It follows from Proposition 5.5 that $\mathrm{Isom}(\mathbb{R}^2)$ is solvable as well. This is because the commutator of any two isometries is a product of four isometries, and hence is even, so the derived subgroup of $\mathrm{Isom}(\mathbb{R}^2)$ is contained in $\mathrm{Isom}^+(\mathbb{R}^2)$.

The discussion in this section should be compared with the proof of Proposition 5.2. We have just shown that the (non-abelian) group $\mathrm{Isom}^+(\mathbb{R}^2)$ can be constructed from the (abelian) groups $\mathbb{R}^2$ (the group of translations) and S^1 (the group of rotations around a given point). This corresponds exactly to the construction in the proof of Proposition 5.2 of an arbitrary even isometry as the product of a translation and a rotation.

Before turning our attention to other things, we find the conjugacy classes of elements of $\mathrm{Isom}^+(\mathbb{R}^2)$. Let $T_{\mathbf{v}}$ be translation by $\mathbf{v}$. Then since $\mathcal{T}$ is abelian, $T_{\mathbf{v}}$ can only be conjugated to something different from itself by using rotations. Furthermore, one may easily check that if R is a rotation, then $I = R \circ T \circ R^{-1}$ is a translation by $R\mathbf{v}$. The fact that I is a translation follows since $\mathcal{T}$ is normal; thus I is determined by the image of a single point $\mathbf{p}$. Taking $\mathbf{p}$ to be the centre of rotation of R, we see that $I\mathbf{p} = RTR^{-1}\mathbf{p} = \mathbf{p} + R\mathbf{v}$, hence I is translation by $R\mathbf{v}$.

Now let $R = R_\theta^{\mathbf{p}}$ be rotation by θ around a point $\mathbf{p}$, and let $I \in \mathrm{Isom}^+(\mathbb{R}^2)$ be arbitrary. Then $\alpha(IRI^{-1}) = \alpha(R)$, and hence IRI^{-1} is also rotation by θ; the centre of rotation may be different, however. Indeed, one can check that if T is any translation, then TRT^{-1} is rotation by θ around the point $T\mathbf{p}$.

Remark. Something slightly different happens if we take conjugacy classes in the whole group $\mathrm{Isom}(\mathbb{R}^2)$. By conjugating $R_\theta^{\mathbf{p}}$ with reflection in a line containing $\mathbf{p}$, we can obtain the rotation $R_{-\theta}^{\mathbf{p}}$. This illustrates the fact that passing to a larger group may make conjugacy classes larger.

Lecture 6. Monday, September 14

a. Finite symmetry groups. In Lecture 4 we examined the groups of symmetries for regular polygons and polyhedra in $\mathbb{R}^2$ and $\mathbb{R}^3$; these groups were finite and contained only rotations and reflections. Then in Lecture 5 we examined the group of all isometries of $\mathbb{R}^2$ (and of $\mathbb{R}^3$), which we saw to be much richer; it has infinitely many elements and contains not only rotations and reflections, but also translations and glide reflections (and in $\mathbb{R}^3$, even more possibilities).

The classification in the previous lecture gives us a complete understanding of individual isometries of $\mathbb{R}^2$. We now know that every isometry falls into one of four categories and we will try to see how can isometries from these different categories be put together to form subgroups of $\mathrm{Isom}(\mathbb{R}^2)$.

This question may be put another way. Given any geometric pattern in the plane—that is, a subset $X \subset \mathbb{R}^2$—the symmetry group of X is a subgroup of $\mathrm{Isom}(\mathbb{R}^2)$. Thus if we want to understand what patterns of symmetries a geometric object may have, we should understand the subgroups of $\mathrm{Isom}(\mathbb{R}^2)$. In particular, an object with finitely many symmetries corresponds to a finite subgroup of $\mathrm{Isom}(\mathbb{R}^2)$, an object with only discrete symmetries corresponds to a discrete subgroup, and so on.[1]

Let us now focus on objects with finitely many symmetries, such as polygons, and address the following specific question: What are the finite subgroups of $\mathrm{Isom}(\mathbb{R}^2)$?

There are two obvious possibilities. Given a rotation R by an angle $2\pi/n$ around some point $\mathbf{p}$, the group $C_n = \langle R\rangle = \{\mathrm{Id}, R, R^2, \dots, R^{n-1}\} \subset \mathrm{Isom}^+(\mathbb{R}^2)$ is a cyclic group of order n. If we let ℓ be a line through $\mathbf{p}$ and L a reflection in ℓ, then $D_n = \langle R, L\rangle$ is the dihedral group of order $2n$, which we already encountered as the group of symmetries of a regular n-gon.

In fact, this is it. There are no other finite subgroups.

Theorem 6.1.

(1) Every finite subgroup of $\mathrm{Isom}^+(\mathbb{R}^2)$ *is cyclic of the form* C_n *for some* $n \in \mathbb{N}$, $\mathbf{p} \in \mathbb{R}^2$.

(2) Every finite subgroup of $\mathrm{Isom}(\mathbb{R}^2)$ *is cyclic (if it contains only even elements) or dihedral (if it contains both even and odd elements).*

Proof. Let G be a finite subgroup of $\mathrm{Isom}(\mathbb{R}^2)$. We may immediately observe that G cannot contain a non-trivial translation or glide reflection, since every element of a finite group must have finite order. Thus either $G \subset \mathrm{Isom}^+(\mathbb{R}^2)$, in which case G contains only rotations, or $G \not\subset \mathrm{Isom}^+(\mathbb{R}^2)$, in which case $G \cap \mathrm{Isom}^+(\mathbb{R}^2)$ is a subgroup of index 2, and so G contains n rotations and n reflections.

[1] A rich variety of geometric patterns corresponding to various symmetry groups in both the Euclidean and non-Euclidean planes appears in the work of the Dutch twentieth century artist M. C. Escher.

Now we must show that all the rotations in a finite group have the same centre. There is an easy way to see it using the homomorphism $\alpha\colon \mathrm{Isom}^+(\mathbb{R}^2) \to \mathbb{R}/2\pi\mathbb{Z}$ constructed in the proof of Proposition 5.4. For, the commutator of any two rotations is a translation which is trivial if and only if rotations commute. However rotations around different points never commute as can be easily seen by looking at the images of centers of those rotations. Thus the group must contain a non-trivial translation and hence is infinite. This argument is strictly two-dimensional and we will now present another proof which illustrates a technique that is applicable in a broader setting. In particular, it works for isometries of Euclidean spaces of arbitrary dimension.

Lemma 6.2. *If $G \subset \mathrm{Isom}(\mathbb{R}^2)$ is a finite group of isometries, then there exists a point $\mathbf{p} \in \mathbb{R}^2$ such that $I\mathbf{p} = \mathbf{p}$ for every $I \in G$.*

Proof. The key observation to make is that isometries respect the centre of mass of a finite set of points; we prove this by induction. Given a finite set $X \subset \mathbb{R}^2$, let $\mathcal{C}(X)$ denote the centre of mass of X. We will show that

$$\mathcal{C}(I(X)) = I(\mathcal{C}(X)). \tag{6.1}$$

If X has only two points, $\mathbf{x}_1$ and $\mathbf{x}_2$, then $\mathcal{C}(X)$ is the midpoint $\mathbf{y}$ of the line segment $[\mathbf{x}_1, \mathbf{x}_2]$. The point $\mathbf{y}$ is the unique point in $\mathbb{R}^2$ such that

$$d(\mathbf{x}_1, \mathbf{y}) = d(\mathbf{x}_2, \mathbf{y}) = \frac{1}{2} d(\mathbf{x}_1, \mathbf{x}_2).$$

If I is an isometry, then we have

$$d(I\mathbf{x}_1, I\mathbf{y}) = d(I\mathbf{x}_2, I\mathbf{y}) = \frac{1}{2} d(I\mathbf{x}_1, I\mathbf{x}_2),$$

and it follows that $I\mathbf{y}$ is the midpoint of $[I\mathbf{x}_1, I\mathbf{x}_2]$, whence $I\mathbf{y} = \mathcal{C}(I(X))$.

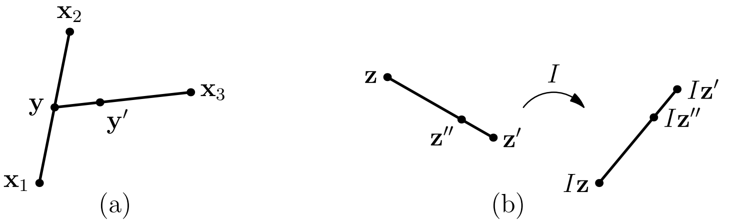


Figure 2.7. Isometries preserve ratios and centres of mass.

Now suppose X has n points $\mathbf{x}_1, \dots, \mathbf{x}_n$, and let $\mathbf{y} = \mathcal{C}(\{\mathbf{x}_1, \dots, \mathbf{x}_{n-1}\})$—the case $n = 3$ is shown in Figure 2.7(a). Then $\mathcal{C}(X)$ is the unique point $\mathbf{y}'$ which lies on the line segment $[\mathbf{y}, \mathbf{x}_n]$ for which

$$d(\mathbf{y}, \mathbf{y}') = \frac{1}{n} d(\mathbf{y}, \mathbf{x}_n), \qquad d(\mathbf{y}', \mathbf{x}_n) = \frac{n-1}{n} d(\mathbf{y}, \mathbf{x}_n). \tag{6.2}$$

Because I is an isometry, the relationships in (6.2) continue to hold for the points $I\mathbf{y}, I\mathbf{y}', I\mathbf{x}_n$. Thus if $I\mathbf{y} = \mathcal{C}(\{I\mathbf{x}_1, \dots, I\mathbf{x}_{n-1}\})$, then $I\mathbf{y}' = \mathcal{C}(I(X))$, and now induction shows that (6.1) holds for any finite X.

Armed with (6.1), we can complete the proof of the lemma.

Definition 6.3. Given a group $G \subset \mathrm{Isom}(\mathbb{R}^2)$ and a point $\mathbf{x} \in \mathbb{R}^2$, the *orbit* of $\mathbf{x}$ is

$$\mathrm{Orb}(\mathbf{x}) = \{I\mathbf{x} \mid I \in G\}.$$

Observe that G is finite if and only if $\mathrm{Orb}(\mathbf{x})$ is finite for every $\mathbf{x} \in \mathbb{R}^2$. Now suppose G is finite, and let $\mathbf{x} \in \mathbb{R}^2$ be arbitrary. For every $I \in G$, we have $I(\mathrm{Orb}(\mathbf{x})) = \mathrm{Orb}(\mathbf{x})$, since I simply permutes the points of the orbit. Using (6.1), this gives

$$I(\mathcal{C}(\mathrm{Orb}(\mathbf{x}))) = \mathcal{C}(I(\mathrm{Orb}(\mathbf{x}))) = \mathcal{C}(\mathrm{Orb}(\mathbf{x})).$$

That is, the centre of mass of any orbit is fixed by every element of the group. Taking $\mathbf{p} = \mathcal{C}(\mathrm{Orb}(\mathbf{x}))$ completes the proof of the lemma. □

Remark. The only property of isometries which we used in the proof of Lemma 6.2 was the fact that they preserve intervals and ratios. That is, if $\mathbf{z}, \mathbf{z}', \mathbf{z}''$ are three collinear points as in Figure 2.7(b), then $I\mathbf{z}, I\mathbf{z}', I\mathbf{z}''$ are still collinear, and furthermore,

$$\frac{d(I\mathbf{z}, I\mathbf{z}'')}{d(I\mathbf{z}'', I\mathbf{z}')} = \frac{d(\mathbf{z}, \mathbf{z}'')}{d(\mathbf{z}'', \mathbf{z}')}. \tag{6.3}$$

These properties hold for a more general class of maps, called *affine maps*. Thus we have in fact proved that any finite group of affine transformations in $\mathbb{R}^n$ has a fixed point.

Returning to the proof of Theorem 6.1, we see from Lemma 6.2 that if $G \subset \mathrm{Isom}^+(\mathbb{R}^2)$, then there exists $\mathbf{p} \in \mathbb{R}^2$ such that every element of G is a rotation around $\mathbf{p}$. Consider the (finite) set

$$\Theta = \{\theta \in [0, 2\pi) \mid R_\theta^{\mathbf{p}} \in G\},$$

and let α be the smallest positive number in θ. If $\alpha \neq 2\pi/n$ for some n, then there exists k such that $k\alpha \in (2\pi, 2\pi + \alpha)$, and consequently

$$(R_\alpha^{\mathbf{p}})^k = R_{k\alpha}^{\mathbf{p}} = R_{k\alpha - 2\pi}^{\mathbf{p}} \in G.$$

But now $0 < k\alpha - 2\pi < \alpha$, which contradicts the definition of α. Thus $\alpha = 2\pi/n$ for some n. Furthermore, every element of G is of the form $R_{k\alpha}^{\mathbf{p}}$ for some $0 \leq k < n$. To see this, fix $\beta \in \Theta$, and observe that if $\beta = k\alpha + \beta'$ for some $0 \leq \beta' < \alpha$, then

$$R_{\beta'}^{\mathbf{p}} = R_\beta^{\mathbf{p}} \circ (R_\alpha^{\mathbf{p}})^{-k} \in G,$$

and hence $\beta' \in \Theta$ as well. It follows that $\beta' = 0$, and so $\beta = k\alpha$.

This completes the proof in the case when $G \subset \text{Isom}^+(\mathbb{R}^2)$. For the general result, observe that if $I, I' \in G$ are orientation-reversing, then $I_e = I^{-1}I' \in G$ is orientation-preserving, and it follows that

$$I' = II_e \in IG^+,$$

where $G^+ = G \cap \text{Isom}^+(\mathbb{R}^2)$ is the even subgroup of G. Thus $G = G^+ \cup IG^+$, and so G^+ is a subgroup of index 2 in G. We know from above that G^+ is generated by a rotation $R^{\mathbf{p}}_{2\pi/n}$, it follows that G is generated by $R^{\mathbf{p}}_{2\pi/n}$ and L, where L is reflection in a line ℓ—the line ℓ contains $\mathbf{p}$ by Lemma 6.2. We have already seen that the group generated by such a rotation and reflection is the dihedral group D_n, and this completes the proof. □

Remark. We have already seen that the dihedral groups arise as the symmetry groups of regular polygons. Can we obtain the cyclic groups as the symmetry groups of geometric figures? To do so, we must construct a figure with no reflective symmetry; this may be done by taking a regular polygon and marking each side asymmetrically, as shown in Figure 2.8(a), to eliminate reflective symmetries. Another example is given by the triskelion (or trinacria); this shape, which appears on the flag of the Isle of Man, shown in Figure 2.8(b), has symmetry group C_3.

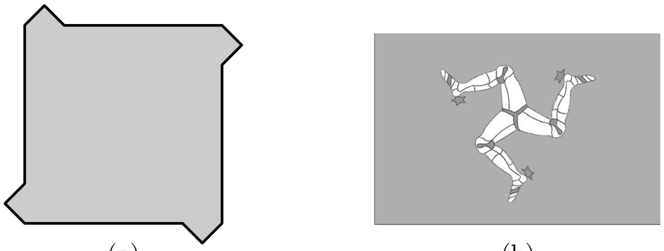

(a) (b)

Figure 2.8. Figures with a cyclic symmetry group.

b. Discrete symmetry groups. Having classified the finite subgroups of $\text{Isom}(\mathbb{R}^2)$, we now expand our horizons a bit and consider a broader class of subgroups—the discrete groups.

Definition 6.4. Fix $X \subset \mathbb{R}^2$. A point $\mathbf{p} \in \mathbb{R}^2$ is an *accumulation point* of X if there exists a sequence of points $\mathbf{x}_1, \mathbf{x}_2, \dots \in X$ such that $\lim_{n\to\infty} \mathbf{x}_n = \mathbf{p}$, but $\mathbf{x}_n \neq \mathbf{p}$ for all n.

We say that X is *discrete* if it does not have any accumulation points. A group $G \subset \text{Isom}(\mathbb{R}^2)$ is discrete if $\text{Orb}(\mathbf{x})$ is discrete for every $\mathbf{x} \in \mathbb{R}^2$.

Example 6.5. Any finite set is discrete; consequently, any finite subgroup of $\text{Isom}(\mathbb{R}^2)$ is discrete. This includes the cyclic groups C_n and the dihedral groups D_n.

Example 6.6. The set $\mathbb{Z}^2 = \{(a,b) \mid a,b \in \mathbb{Z}\} \subset \mathbb{R}^2$ is discrete. Consider the group G of translations by integer coordinates:

$$G = \{T_{\mathbf{v}} \in \mathrm{Isom}(\mathbb{R}^2) \mid \mathbf{v} = (a,b), a,b \in \mathbb{Z}\}.$$

Then $\mathrm{Orb}(\mathbf{x}) = \mathbf{x} + \mathbb{Z}^2$ is discrete for every $\mathbf{x} \in \mathbb{R}^2$; hence G is discrete but not finite.

Example 6.7. Suppose G contains a rotation $R_\theta^{\mathbf{p}}$ such that $\theta/2\pi$ is irrational. Then given any $\mathbf{x} \neq \mathbf{p}$, $\mathrm{Orb}(\mathbf{x})$ is an infinite subset of the circle centred at $\mathbf{p}$ with radius $d(\mathbf{p},\mathbf{x})$. Hence it must have accumulation points so that G is not discrete.

In fact, with a little extra effort we can show that the orbit is dense in the circle and hence every point on this circle is an accumulation point of $\mathrm{Orb}(\mathbf{x})$. To see that notice that rotation that maps any point on the orbit to any other point belongs to the group G. Hence G contains rotations by angle arbitrary close to zero. Applying iterates of such small rotations to $\mathbf{x}$ we obtain denser and denser subsets of the circle.

Remark. The above definition of a discrete group is extrinsic—that is, it relies on the action of G on the plane $\mathbb{R}^2$ and is given in terms of the orbits of points. An intrinsic definition can also be given by defining a notion of convergence in $\mathrm{Isom}(\mathbb{R}^2)$ itself: fixing three non-collinear points $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3 \in \mathbb{R}^2$, say that a sequence $\{I_n\} \subset \mathrm{Isom}(\mathbb{R}^2)$ converges to $I \in \mathrm{Isom}(\mathbb{R}^2)$ if and only if $I_n\mathbf{x}_i \to I\mathbf{x}_i$ for each $i = 1,2,3$. Three points is enough to determine an isometry, and so this implies that $I_n\mathbf{x} \to I\mathbf{x}$ for every $\mathbf{x} \in \mathbb{R}^2$. Then we may say that G is a discrete group if and only if it is discrete as a set—that is, it has no accumulation points.

Observe that this notion of convergence is not determined by the algebraic structure of the group. Let T be a translation and R a rotation by an irrational multiple of 2π. Then both $\langle T\rangle$ and $\langle R\rangle$ are isomorphic to $\mathbb{Z}$, but T is discrete, while R is not.

We see from the above examples that discrete groups form a broader class of groups than finite groups; however, they are still simple enough that we may hope to give a complete classification along the lines of Theorem 6.1.

So far we have seen two examples of infinite discrete groups: $\langle T_{\mathbf{v}}\rangle$ and $\langle T_{\mathbf{v}}, T_{\mathbf{w}}\rangle$, where $\mathbf{v}$ and $\mathbf{w}$ are linearly independent. In fact, given *any* discrete group $G \subset \mathrm{Isom}(\mathbb{R}^2)$, one may show that all of its non-finiteness comes from translations. To make this precise, we need the following definition.

Definition 6.8. Given a subgroup $G \subset \mathrm{Isom}(\mathbb{R}^2)$, the *translation subgroup* of G is

$$G_T = \mathcal{T} \cap G = \{I \in G \mid I \text{ is a translation}\}.$$

The precise statement of the claim that "all of the non-finiteness of G comes from translations" is that the translation subgroup has finite index. Since G_T is a normal subgroup of G, we may consider the factor group G/G_T, and we will see that this factor group is finite.

This illustrates a general principle in the theory of infinite groups: many infinite groups can be decomposed by finding a subgroup of finite index which takes a known form (such as the translation subgroup), and reducing questions about the whole group to questions about this subgroup (which is well understood) and about the quotient group (which is finite).[2]

We now address two questions which arise from the above ideas. First, what are the possible translation subgroups G_T? Second, what are the possible factor groups G/G_T?

PROPOSITION 6.9. *Given any discrete group $G \subset \mathrm{Isom}(\mathbb{R}^2)$, the translation subgroup G_T is one of the following:*

(1) The trivial group $\{\mathrm{Id}\}$.
(2) An infinite cyclic group $\langle T_{\mathbf{v}} \rangle$.
(3) A rank-2 abelian group $\langle T_{\mathbf{v}}, T_{\mathbf{w}} \rangle$, where $\mathbf{v}, \mathbf{w}$ are linearly independent.[3]

PROOF. Consider the orbit $\mathrm{Orb}(\mathbf{0})$ of the origin under the action of G_T. If $\mathrm{Orb}(\mathbf{0}) = \{\mathbf{0}\}$, then G_T is trivial; otherwise let $\mathbf{v}$ be the element of $\mathrm{Orb}(\mathbf{0})$ closest to the origin. Now $n\mathbf{v} \in \mathrm{Orb}(\mathbf{0})$ for every $n \in \mathbb{Z}$, and $\mathrm{Orb}(\mathbf{0})$ contains no other elements of the line ℓ through $\mathbf{0}$ and $\mathbf{v}$ (otherwise one of them would be closer to $\mathbf{0}$ than $\mathbf{v}$ is).

If these are all the points in $\mathrm{Orb}(\mathbf{0})$, then $G_T = \langle T_{\mathbf{v}} \rangle$; otherwise we may let $\mathbf{w}$ be the closest point in $\mathrm{Orb}(\mathbf{0})$ to $\mathbf{0}$ which does not lie on ℓ. We see that

$$\mathbf{v}\mathbb{Z} + \mathbf{w}\mathbb{Z} = \{a\mathbf{v} + b\mathbf{w} \mid a, b \in \mathbb{Z}\} \subset \mathrm{Orb}(\mathbf{0}). \tag{6.4}$$

Let P be the parallelogram whose vertices are $\mathbf{0}$, $\mathbf{v}$, $\mathbf{w}$, and $\mathbf{v} + \mathbf{w}$. Then if we do not have equality in (6.4), we can find a point $\mathbf{p} \in \mathrm{Orb}(\mathbf{0})$ which lies inside P. In fact, all four of the points $\mathbf{p}$, $\mathbf{p} - \mathbf{v}$, $\mathbf{p} - \mathbf{w}$, and $\mathbf{p} - \mathbf{v} - \mathbf{w}$ are contained in $\mathrm{Orb}(\mathbf{0})$ (an immediate consequence of the fact that $\mathbf{p} + a\mathbf{v} + b\mathbf{w} \in \mathrm{Orb}(\mathbf{0})$ for all integers a and b). One of these four points is closer to $\mathbf{0}$ than $\mathbf{w}$ is. To see this, notice that $\mathbf{p}$ lies on one side of a diagonal of P, say, the one connecting $\mathbf{v}$ and $\mathbf{w}$. then the sum of length of segments connecting $\mathbf{p}$ with $\mathbf{v}$ and $\mathbf{w}$ is less that the sum of lengths of the sides of P, i.e lengths of vectors $\mathbf{v}$ and $\mathbf{w}$. Hence at least one of the those segments is shorter than the longer side, i.e $\mathbf{w}$. This contradiction shows that equality holds in (6.4). This in turn implies that $G_T = \langle T_{\mathbf{v}}, T_{\mathbf{w}} \rangle$, and we are done. □

EXERCISE 6.1. Show that G/G_T is isomorphic to a finite group whose elements are all either rotations around a single point $\mathbf{p} \in \mathbb{R}^2$ or reflections in lines through $\mathbf{p}$.

[2] One sometimes says that if a finite index subgroup of G has a property P, then G is *virtually* P. For example, since the translation subgroup is abelian and is of finite index, any discrete subgroup of $\mathrm{Isom}(\mathbb{R}^2)$ is virtually abelian.

[3] The *rank* of a group G is the minimal number of generators of G—that is, the cardinality of the smallest set $X \subset G$ such that $\langle X \rangle = G$.

It follows from discreteness of G that any rotation in G must be by an angle which is a rational multiple of 2π. If the translation subgroup G_T is trivial, then there are no further restrictions; the cyclic and dihedral groups provide examples of finite (hence discrete) groups with trivial translation subgroups in which rotations of any rational angle (with respect to 2π) appear.

The situation is different if the translation subgroup is non-trivial. In this case, we have the following theorem, which places significant restrictions on the angles of rotations in G.

Theorem 6.10 (Crystallographic restriction theorem). *Let G be a discrete subgroup of* $\mathrm{Isom}(\mathbb{R}^2)$*, and suppose that G contains a non-trivial translation. Suppose furthermore that G contains a rotation $R_\theta^{\mathbf{p}}$. Then θ is of the form $2\pi/k$, where $k = 1, 2, 3, 4,$ or 6.*

Proof. Let $X = \mathrm{Orb}(\mathbf{p})$ be the orbit of $\mathbf{p}$ under the group G. Then given $\mathbf{q} \in X$, let $I \in G$ be such that $I\mathbf{p} = \mathbf{q}$, and observe that $IR_\theta^{\mathbf{p}}I^{-1} = R_\theta^{\mathbf{q}}$, and thus $R_\theta^{\mathbf{q}} \in G$. This shows that G contains the rotation by θ around every point in the lattice X.

Because $X = \mathrm{Orb}(\mathbf{p})$ is discrete, there exist $\mathbf{p} \neq \mathbf{q} \in X$ such that $d(\mathbf{p}, \mathbf{q})$ is minimal—that is, $d(\mathbf{x}, \mathbf{y}) \geq d(\mathbf{p}, \mathbf{q})$ for all $\mathbf{x} \neq \mathbf{y} \in X$. (This is where we use the requirement that G contains a translation, as this guarantees that the orbit of $\mathbf{p}$ contains a point besides $\mathbf{p}$ itself.)

By discreteness, we have $\theta = 2\pi a/b$, where a and b are relatively prime. Choosing n such that $na \equiv 1 \pmod{b}$, we see that G contains $(R_\theta^{\mathbf{p}})^n = R_{2\pi/b}^{\mathbf{p}}$. If $b > 6$, then we have

$$d(R_{2\pi/b}^{\mathbf{p}}\mathbf{q}, \mathbf{q}) < d(\mathbf{p}, \mathbf{q}),$$

which contradicts the definition of $\mathbf{p}$ and $\mathbf{q}$. Thus it only remains to eliminate the possibility that $b = 5$. This can be done by observing that in this case, we get

$$d(R_{2\pi/5}^{\mathbf{p}}\mathbf{q}, R_{-2\pi/5}^{\mathbf{q}}\mathbf{p}) < d(\mathbf{p}, \mathbf{q}),$$

which is again a contradiction. The result follows. □

Remark. If one considers square, triangular, and hexagonal lattices, it is not hard to see that each of the remaining possibilities in fact occurs. Using this observation together with Theorem 6.10, the 19th-century Russian crystallographer E. S. Fedorov showed that every discrete subgroup of $\mathrm{Isom}(\mathbb{R}^2)$ with a rank-2 translation subgroup is isomorphic to one of the 17 *crystallographic groups*.

Lecture 7. Wednesday, September 16

a. Remarks on Homework #3.

Remark. The group $Q = \{\pm 1, \pm i, \pm j, \pm k\}$ with multiplication rules given by $i^2 = j^2 = k^2 = ijk = -1$ is called the *quaternion group*. It is related to the *quaternions*, which are a sort of four-dimensional analogue of the complex numbers. To make this more precise, observe that the real numbers are a one-dimensional vector space with the structure of a field—that is, addition and multiplication satisfying commutativity, associativity, and distributive laws, such that every element has both additive and multiplicative inverses (except for 0). The complex numbers are a two-dimensional vector space with the structure of a field, and it is reasonable to ask if we can turn $\mathbb{R}^d$ into a field for $d \geq 3$. It turns out that this is impossible; however, $\mathbb{R}^4$ can be given a structure which makes it *almost* a field, by choosing basis elements $\{1, i, j, k\}$ and defining multiplication as above. The only thing that is missing is commutativity of multiplication (we get what is called a *skew field*). One can carry out a similar procedure in $\mathbb{R}^8$, but in this case we lose both commutativity *and* associativity.

Remark. The quaternion group is the last item in complete classification of groups of order 8 after three abelian groups and the dihedral group D_4. This is the smallest number n such that there are more than two non-isomorphic groups of order n. Recall that the possibility of a complicated internal structure of a group corresponds to a high degree of divisibility of its order; thus "interesting" orders to consider are numbers that are products of more than two primes: $8, 12, 16, 18, 20, 24, \dots$.

Remark. So far we have considered groups of isometries of objects X which live in Euclidean space $\mathbb{R}^2$ or $\mathbb{R}^3$. In the torus $\mathbb{R}^2/\mathbb{Z}^2$ and the elliptic plane $\mathbb{E}^2$, we have objects which do *not* live in Euclidean space (at least, not in $\mathbb{R}^3$), and so we are in some sense in a more general setting.

b. Classifying isometries of $\mathbb{R}^3$. In Lecture 4, on page 42, we defined the notion of a *right-handed basis* of $\mathbb{R}^3$, and characterised even isometries of $\mathbb{R}^3$ as those isometries which map right-handed bases to right-handed bases. An alternate definition is to observe that in two dimensions, there were just two possible configurations for the basis vectors $\mathbf{v}_1$ and $\mathbf{v}_2$; either $\mathbf{v}_1$ is $\pi/2$ clockwise from $\mathbf{v}_2$, or vice versa. In $\mathbb{R}^3$, there are six configurations, corresponding to the six permutations in S_3, and even and odd isometries correspond to even and odd permutations.

This can be stated most clearly in the language of linear algebra, about which we will have more to say later on. For now, observe that if I is an isometry which fixes a point $\mathbf{p}$, then in coordinates centred at $\mathbf{p}$, I defines a linear map, and hence a 3×3 matrix. One may show that I is orientation-preserving if the determinant of this matrix is positive, and orientation-reversing if it is negative.

We have analogues of Propositions 5.2 and 5.3 in three dimensions.

PROPOSITION 7.1. *Let* $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4 \in \mathbb{R}^2$ *be non-coplanar, and suppose* $\mathbf{x}'_i \in \mathbb{R}^2$ *are such that*

$$d(\mathbf{x}'_i, \mathbf{x}'_j) = d(\mathbf{x}_i, \mathbf{x}_j) \tag{7.1}$$

for every $1 \leq i, j \leq 4$. *Then there exists a unique isometry* $I\colon \mathbb{R}^2 \to \mathbb{R}^2$ *such that* $I\mathbf{x}_i = \mathbf{x}'_i$ *for every* $1 \leq i \leq 4$.

PROOF. Let P be the plane containing $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{x}_3$, and let P' be the plane containing $\mathbf{x}'_1$, $\mathbf{x}'_2$, and $\mathbf{x}'_3$. Then there exists an isometry $J \in \operatorname{Isom}(\mathbb{R}^3)$ such that $J(P) = P'$: if P and P' are parallel, take J to be a translation, and if they intersect, take J to be a rotation around their line of intersection. Furthermore, by Proposition 5.2, there exists a unique isometry $I \in \operatorname{Isom}(P')$ such that $I(J(\mathbf{x}_i)) = \mathbf{x}'_i$ for $i = 1, 2, 3$.

In fact, Proposition 5.2 shows that the isometry $I \circ J$ is uniquely defined on P. To determine the extension of $I \circ J$ to $\mathbb{R}^3$, it suffices to know which side of P and P' the points $\mathbf{x}_4$ and $\mathbf{x}'_4$ lie on; this determines the orientation of the isometry. To see that this determines the extension uniquely, observe that given any $\mathbf{y} \in \mathbb{R}^3$, the image of $\mathbf{y}$ must lie on each of the three spheres centred at $\mathbf{x}'_i$ with radius $d(\mathbf{y}, \mathbf{x}_i)$, for $i = 1, 2, 3$. Because the points $\mathbf{x}'_i$ are non-collinear, these three spheres intersect in at most two points (in fact, in exactly two points). One point corresponds to an even isometry, the other to an odd isometry. □

PROPOSITION 7.2. *Every isometry of* $\mathbb{R}^3$ *is a product of no more than four reflections.*

Every isometry of $\mathbb{R}^3$ *with a fixed point is a product of no more than three reflections.*

PROOF. The proof is similar to the proof of Proposition 5.3.

Since every isometry is determined by images of four points we will construct reflections matching additional point without moving those that have already been matched.

First notice that there is exactly one reflection that maps a given point $\mathbf{x}$ into another point $\mathbf{x}' \neq \mathbf{x}$, namely reflection in the plane $P_{\mathbf{x},\mathbf{x}'}$ perpendicular to the segment $[\mathbf{x}, \mathbf{x}']$ passing through the midpoint of that segment. This plane may be characterized as the locus of points equidistant from $\mathbf{x}$ and $\mathbf{x}'$. Thus if the points $\mathbf{x}$ and$\mathbf{x}'$ are equidistant from one, two or three other points the plane $P_{\mathbf{x},\mathbf{x}'}$ passes through those points.

These remarks provides an algorithm for constructing reflections as required. Given quadruples of points as is Proposition 7.1 we first construct the reflection L_1 in $P_{\mathbf{x}_1,\mathbf{x}'_1}$, then, if necessary, reflection L_2 in the plane $P_{L_1\mathbf{x}_2,\mathbf{x}'_2}$ (that fixes $\mathbf{x}'_1$), then, if necessary, reflection L_3 in $P_{L_2 \circ L_1 \mathbf{x}_3,\mathbf{x}'_3}$ (that fixes $\mathbf{x}'_1$ and $\mathbf{x}'_2$) and finally, if necessary, reflection L_4 in $P_{L_3 \circ L_2 \circ L_1 \mathbf{x}_4,\mathbf{x}'_4}$. Any step can be missed if two points in the corresponding pair coincide. In particular, if $\mathbf{x}_1 = \mathbf{x}'_1$ the first step is not needed. □

Since reflections are odd isometries and the product of reflections in two parallel planes is a translation, we immediately have

Corollary 7.3. *Every even isometry with a fixed point is a product of two reflections in two non-parallel planes and hence is a rotation whose axis is the line of their intersection.*

Every odd isometry other than a reflection is a product of three reflections.

Remark. One can see that above arguments extend to higher dimensions in a straightforward way: in $\mathbb{R}^n$ at most $n+1$ reflections are needed. Nevertheless synthetic approach to classification of isometries in higher dimension becomes rather cumbersome; we will resort to linear algebra to accomplish that goal.

As mentioned in Lecture 5d, there are isometries of $\mathbb{R}^3$ which are qualitatively different from anything found in $\mathbb{R}^2$. Since every isometry is a product of reflections, and each reflection is determined by a plane, this corresponds to the fact that there are arrangments of planes in $\mathbb{R}^3$ for which there are no analogous arrangements of lines in $\mathbb{R}^2$.

We saw in Figure 2.5 that given two reflections L_1 and L_2, there are many other pairs of reflections L_1' and L_2' such that $L_1 \circ L_2 = L_1' \circ L_2'$. In particular, this allowed us to classify all products of three reflections in $\mathbb{R}^2$ by assuming without loss of generality that two of the corresponding lines are parallel.

No such general result is available in $\mathbb{R}^3$. To see this, let P_1 be the xy-plane, P_2 the xz-plane, and P_3 the yz-plane. For $i = 1, 2, 3$, let L_i be the reflection in P_i, and for $i \neq j$, let $\ell_{ij} = P_i \cap P_j$ be the line of intersection of P_i and P_j. Observe that ℓ_{12} is orthogonal to P_3, and similarly for P_1 and P_2.

The product $L_1 \circ L_2$ is a rotation by π around the line ℓ_{12}. This rotation can be decomposed as the product $L_1' \circ L_2'$ if and only if the corresponding planes P_1' and P_2' meet at a right angle and have intersection ℓ_{12}. Since ℓ_{12} is orthogonal to P_3, we see that neither P_1' nor P_2' can be made parallel to P_3. It follows that *any* decomposition of $C = L_1 \circ L_2 \circ L_3$ as a product of three reflections must use three pairwise orthogonal planes.

The isometry $C\colon \mathbf{x} \mapsto -\mathbf{x}$ is known as the *central symmetry*. The analogously defined map in $\mathbb{R}^2$ was an even isometry (rotation by π around $\mathbf{0}$); this is not the case here, and we have obtained an isometry which does not fit into any of the four classes that categorise $\text{Isom}(\mathbb{R}^2)$.

In fact, every odd isometry of $\mathbb{R}^3$ is either a reflection, a glide reflection, or a *rotatory reflection*—that is, a composition of reflection in a plane P with rotation around a line ℓ orthogonal to P.

Passing to even isometries, we must categorise the possible configurations of *four* planes. This will allow us to completely classify even isometries of $\mathbb{R}^3$.

c. Isometries of the sphere. The first subgroup of $\mathrm{Isom}(\mathbb{R}^2)$ that we considered was $G_{\mathbf{p}}$, the group of isometries which fix a given point $\mathbf{p}$. We saw that this group has a normal subgroup of index two which comprises all rotations around $\mathbf{p}$; this subgroup is abelian, and so $G_{\mathbf{p}}$ is not a particularly complicated group. In particular, it is solvable.

It turns out that the situation in $\mathbb{R}^3$ is a little more subtle. We can use the results in the previous section to describe all the isometries which fix the origin $\mathbf{0} \in \mathbb{R}^3$:[4]

$$O(3) = O(3,\mathbb{R}) = \{I \in \mathrm{Isom}(\mathbb{R}^3) \mid I\mathbf{0} = \mathbf{0}\}. \tag{7.2}$$

The group $O(3)$ is called the *orthogonal group*, for reasons that will become clear when we consider matrix representations of isometries. Observe that every isometry $I \in O(3)$ is also an isometry of the sphere

$$S^2 = \{\mathbf{x} \in \mathbb{R}^3 \mid \|\mathbf{x}\| = 1\}.$$

Conversely, every isometry $I \in \mathrm{Isom}(S^2)$ can be extended to an isometry of $\mathbb{R}^3$ using Proposition 7.1, and it follows that $O(3)$ is isomorphic to $\mathrm{Isom}(S^2)$.

We will also be interested in the *special orthogonal group*

$$SO(3) = SO(3,\mathbb{R}) = \{I \in \mathrm{Isom}^+(\mathbb{R}^3) \mid I\mathbf{0} = \mathbf{0}\}. \tag{7.3}$$

Since orientation on the surface of the sphere can be defined similarly to that in the plane we can speak about even isometries of the sphere. Then $SO(3) = \mathrm{Isom}^+(S^2)$.

What even isometries of $\mathbb{R}^3$ have a fixed point? That is, what sorts of isometries are in $SO(3)$? One immediately sees that if ℓ is a line through $\mathbf{0}$, then any rotation R_θ^ℓ around ℓ by an arbitrary angle θ fixes $\mathbf{0}$, and so $R_\theta^\ell \in SO(3)$. By Corollary 7.3 we immediately see that those are the only elements of $SO(3)$.

As a cautionary note, we observe that we must actually fix a *directed* line ℓ in order to determine the rotation R_θ^ℓ—that is, we must fix a positive and negative direction along ℓ so that we know which direction of rotation is positive and which is negative. If we denote by $-\ell$ the line ℓ with the opposite orientation, then $R_\theta^{-\ell} = R_{-\theta}^\ell$.

d. The structure of $SO(3)$. Significance of the group $SO(3)$ goes far beyond geometry, and we seek to understand its algebraic structure. One way of doing this is to find the conjugacy classes of $SO(3)$; how do we tell when two rotations in $SO(3)$ are conjugate?

Given $R_\theta^\ell \in SO(3)$ and an arbitrary isometry I, we see that $IR_\theta^\ell I^{-1}$ fixes the line $\ell' = I\ell$, and indeed, that

$$IR_\theta^\ell I^{-1} = R_\theta^{\ell'}.$$

If ℓ and ℓ' are any two lines through the origin, then there exists $I \in SO(3)$ such that $I\ell = \ell'$, and it follows that R_θ^ℓ and $R_\theta^{\ell'}$ are conjugate. Thus

[4] If we consider isometries fixing an arbitrary point $\mathbf{p}$, we will obtain a subgroup that is conjugate to the one we now consider.

the conjugacy classes in $SO(3)$ stand in one-to-one correspondence with the interval $[0,\pi]$; each conjugacy class is of the form

$$\mathcal{R}_\theta = \{R_\theta^\ell \mid \ell \text{ is a line through } \mathbf{0}\}$$

for some $\theta \in [0,\pi]$. Observe that we do not need to consider $\theta \in (\pi, 2\pi)$, since $R_\theta^\ell = R_{2\pi-\theta}^{-\ell}$. In particular, every rotation in $SO(3)$ is conjugate to its inverse.

The two-dimensional analogue of $SO(3)$ is the group of rotations of the plane around $\mathbf{0}$, which is abelian, and hence has conjugacy classes which are single elements. The much larger conjugacy classes in $SO(3)$ correspond to the fact that $SO(3)$ is significantly less commutative than its two-dimensional counterpart. Indeed, we have the following result.

Theorem 7.4. *$SO(3)$ is simple.*

Proof. We must show that $SO(3)$ has no normal subgroups besides the trivial subgroup and $SO(3)$ itself. Recall that a subgroup is normal if and only if it is a union of conjugacy classes. Thus we must show that any subgroup G which contains an entire conjugacy class $\mathcal{R}_\theta$ is in fact the entire group.

Geometrically, this means that given $\theta \neq 0$, we can obtain *any* rotation as the product of rotations by θ around different axes. The idea of the proof is made clear by considering the particular case $\theta = \pi$.

Observe that given any two orthogonal planes which contain $\mathbf{0}$, the product of the corresponding reflections is a rotation by π, and hence lies in $\mathcal{R}_\pi$. Let P and P' be arbitrary planes which contain $\mathbf{0}$, and let P'' be another plane which contains $\mathbf{0}$ and is orthogonal to both P and P' (this can be accomplished by taking P'' orthogonal to the line $\ell = P \cap P'$). Let L, L', L'' be the corresponding reflections. Then $L \circ L''$ and $L'' \circ L'$ are both in $\mathcal{R}_\pi$, and their product is

$$(L \circ L'') \circ (L'' \circ L') = L \circ L' \in \langle \mathcal{R}_\pi \rangle.$$

L and L' were arbitrary, and hence every rotation is in $\langle \mathcal{R}_\pi \rangle$. It follows that $G = SO(3)$.

The same technique works for $\theta \neq \pi$; all that needs to be modified is that P'' should meet the line ℓ at an angle of $\theta/2$. □

This is the first complete proof we have seen that a group is simple; we encountered the alternating group A_5 earlier, but did not prove its simplicity. In fact, A_5 is the group of even isometries of the dodecahedron (or icosahedron), and so can be realised as a subgroup of $SO(3)$.

Theorem 7.4 shows that $\mathrm{Isom}^+(\mathbb{R}^3)$ is not solvable, and a similar result holds in higher dimensions. Thus the result of Proposition 5.5 shows that the two-dimensional case is somehow exceptional.

e. The structure of $O(3)$ and odd isometries. Finally, we observe that the orthogonal group $O(3)$ is generated by $SO(3)$ and the central symmetry C that commutes with every reflection in a plane passing through the origin and hence with every element of $O(3)$.

Definition 7.5. Given a group G, the *centre* of G is

$$Z(G) = \{g \in G \mid gh = hg \text{ for all } h \in G\}.$$

The center of any group is a subgroup since If $g_1h = hg_1$ and $G_2h = hg_2$ then $g_1g_2h = g_1hg_2 = hg_1g_2$

Proposition 7.6. *$Z(O(3)) = \{\mathrm{Id}, C\}$ and $O(3)$ is isomorphic to the direct product of $SO(3)$ and $\mathbb{Z}/2\mathbb{Z}$.*

Proof. We saw that $C \in Z(O(3))$. If the center contains other elements it has to contain even isometries other than identity, hence rotations. But this contradicts simplicity of $SO(3)$.

Now any odd isometry $I \in O(3)$ can be written in a unique way as the product of C and an element of $SO(3)$. Since those commute this provides desired isomorphism. □

This representation allows us to finish classification of odd isometries of $\mathbb{R}^3$ with fixed points. Every such isometry is conjugate to an element I of $O(3) \setminus SO(3)$, i.e the product of the central symmetry with a rotation R The former is the product of reflections in *any* three mutually orthogonal planes. One can pick the first two so that their product is rotation by π around the axis of R. Thus I is the product of a rotation and reflection in the plane perpendicular to its axis (that commutes with the rotation), what we called *rotatory reflection.*

Lecture 8. Friday, September 18

a. Odd isometries with no fixed points. Having classified all the isometries of $\mathbb{R}^3$ which have fixed points, we turn our attention to isometries without fixed points. First we consider odd isometries; any odd isometry with no fixed point is the product of three reflections,

$$I = L_1 \circ L_2 \circ L_3.$$

Let P_1, P_2, P_3 be the corresponding planes. If P_1 and P_2 are parallel, then $T = L_1 \circ L_2$ is a translation and $I = T \circ L_3$; otherwise consider the line $\ell = P_1 \cap P_2$. If $\ell \cap P_3 \neq \emptyset$, then there exists $\mathbf{p} \in P_1 \cap P_2 \cap P_3$, and we see that $I\mathbf{p} = \mathbf{p}$. Since I has no fixed point, we conclude that ℓ and P_3 are parallel. Observe that $L_1 \circ L_2 = L_1' = L_2'$ whenever the corresponding planes P_1' and P_2' intersect at the appropriate angle in the line ℓ; in particular, we may take P_2' parallel to P_3 and obtain

$$I = L_1' \circ L_2' \circ L_3 = L_1' \circ T$$

for some translation T. Thus every odd isometry with no fixed point is the product of a translation and a reflection.

As in Lecture 5(c), we may decompose the translation T into parts T' and T'' which are respectively parallel to and orthogonal to the plane P_1, and hence we obtain $L_1' \circ T = L_1'' \circ T' = T' \circ L_1'$, where L_1'' is reflection in a plane P_1'' parallel to P_1'. A similar decomposition may be done when $I = T \circ L_3$, and it follows that every odd isometry of $\mathbb{R}^3$ with no fixed point is a glide reflection.

We see now that every odd isometry is one of three things:

(1) a reflection;
(2) the product of a reflection L and a rotation R, where L and R commute;
(3) the product of a reflection L and a translation T, where L and T commute.

Similarly, we will later see that every even isometry is one of three things:

(1) a rotation;
(2) a translation;
(3) the product of a rotation R and a translation T, where R and T commute.

b. Remarks from Thursday tutorial: Solution to Homework #2, Exercise 11(1).

PROPOSITION 8.1. *If $G \subset \mathrm{Isom}(\mathbb{R}^2)$ is discrete, then G contains a translation subgroup of finite index.*

PROOF. Let $G^+ = G \cap \mathrm{Isom}^+(\mathbb{R}^2)$ be the subgroup of even isometries in G, and observe that either $G^+ = G$ or G^+ is of index 2. Let $G_T = G \cap \mathcal{T} \subset G^+$ be the translation subgroup of G (the group of all translations in G). Then G_T has finite index in G if and only if it has finite index in G^+.

Next notice that if G (and hence G^+ is infinite) it contains a translation. For, the commutator of any two elements of G^+ is a translation and since rotations around different point do not commute there are non-triviual commutators. (Remember that any group consisting of rotations around one center is either finite or not discrete).

Now notice that the conjugate of a translation by a rotation is a translation by the vector of the same length turned by the angle of rotation. Hence either G contains two linearly independent translations or it can only contain rotations by π. In the latter case the subgroup of translations in G^+ has index two.

In the former case we can apply the Crystallographic restrictions Theorem 6.10

Let $\alpha\colon G^+ \to S^1 = \mathbb{R}/2\pi\mathbb{Z}$ be the homomorphism defined in the proof of Proposition 5.4, so that $I \in G^+$ is a translation if $\alpha(I) = 0$, and otherwise is a rotation of the form $R^{\mathbf{p}}_{\alpha I}$ for some $\mathbf{p} \in \mathbb{R}^2$. Consider the subgroup

$$\Theta = \operatorname{Im}(\alpha) = \{\theta + 2\pi\mathbb{Z} \mid R^{\mathbf{p}}_{\theta} \in G \text{ for some } \mathbf{p} \in \mathbb{R}^2\},$$

and observe that G_T has finite index in G^+ if and only if Θ is finite.

Indeed, every coset of G_T in G^+ is of the form $R^{\mathbf{p}}_{\theta} G_T$ for some $\theta \in [0, 2\pi)$ and $\mathbf{p} \in \mathbb{R}^2$, and thus determines a unique $\theta \in \Theta$. To show that $|\Theta|$ is equal to the index of G_T in G^+, it suffices to show that two rotations by the same angle around different points determine the same coset, and so the map $R^{\mathbf{p}}_{\theta} \mapsto \theta + 2\pi\mathbb{Z}$ is one-to-one. (It is onto by the definition of Θ.)

This is quite simple. In the first place, $R^{\mathbf{p}}_{\theta}$ and $R^{\mathbf{q}}_{\theta}$ determine the same coset if and only if $R^{\mathbf{p}}_{\theta} G_T = R^{\mathbf{q}}_{\theta} G_T$—that is, if and only if $I = (R^{\mathbf{p}}_{\theta})^{-1} R^{\mathbf{q}}_{\theta} \in G_T$. We see immediately that $\alpha(I) = 0$, so I is a translation, and since both the rotations are in G^+, so is I. It follows that $I \in G_T$; hence any two rotations by the same angle determine the same coset of G_T.

By Theorem 6.10 Θ is finite and in fact contains at most six elements. □

c. Finite subgroups of $\operatorname{Isom}(\mathbb{R}^3)$**.** We continue to ask the same sorts of questions we did for isometries of the plane, and seek a classification of all finite subgroups of $\operatorname{Isom}(\mathbb{R}^3)$.

Let $G \subset \operatorname{Isom}(\mathbb{R}^3)$ be a finite subgroup, and observe that Lemma 6.2 works in three dimensions as well, so there exists $\mathbf{p} \in \mathbb{R}^3$ such that $I\mathbf{p} = \mathbf{p}$ for all $I \in G$. It follows that G is conjugate (by translation) to a subgroup of $O(3)$. Furthermore, since every odd isometry in $O(3)$ is the product of the central symmetry C and an even isometry in $SO(3)$, we may consider $G^+ = G \cap SO(3)$ and observe that either $G^+ = G$ or G contains the central symmetry C and then isomorphic to the direct product of G^+ and $\mathbb{Z}/2\mathbb{Z}$, or G is isomorphic to a subgroup of $SO(3)$ that consists of G^+ and $C(G \setminus G^+$. Thus in order to classify the finite subgroups of $\operatorname{Isom}(\mathbb{R}^3)$, it suffices to classify the finite subgroups of $SO(3)$.

We have already encountered a number of finite subgroups of $SO(3)$. For example, given any line ℓ through $\mathbf{0}$, we have the cyclic group

$$C_n = \{R^{\ell}_{2\pi k/n} \mid 0 \leq k < n\} \subset SO(3).$$

A somewhat surprising fact is that we also have the dihedral group D_n, as a subgroup of $SO(3)$, which in $\text{Isom}(\mathbb{R}^2)$ contains odd as well as even isometries. The difference in $\mathbb{R}^3$ is as follows. Let $L \in \text{Isom}(\mathbb{R}^2)$ be a reflection in the line ℓ, and embed $\mathbb{R}^2$ in $\mathbb{R}^3$ as a plane P. Consider the isometry R^{ℓ}_{π}, which is rotation around the line ℓ by an angle π; this isometry maps P to itself, and its restriction to that plane agrees with the reflection in ℓ.[5]

Thus we may take any line ℓ through $\mathbf{0}$, and let $\ell_1, \dots, \ell_n$ be lines through $\mathbf{0}$ which are orthogonal to ℓ and whose angles with each other are multiples of π/n. Then we can realise the dihedral group as

$$D_n = C_n \cup \{R^{\ell_k}_{\pi} \mid 1 \leq k \leq n\} \subset SO(3).$$

Notice that this is the group of isometries of a regular n-gon in $\mathbb{R}^3$ or of the rectangular prism build over that polygon.

In Lecture 4(c), we investigated the symmetry groups of the five platonic solids. We return to these groups now and complete the investigations begun there.

The tetrahedron. Let X be a regular tetrahedron with vertices $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$ lying on the sphere S^2. For any $i \neq j \in \{1, 2, 3, 4\}$, let P_{ij} be the plane which is the perpendicular bisector of the line segment from $\mathbf{x}_i$ to $\mathbf{x}_j$, and let L_{ij} be reflection in P_{ij}. Then P_{ij} contains the other two vertices of X, and hence L_{ij} permutes the vertices of X according to the transposition $(i\ j)$; it interchanges i and j and leaves the other two vertices fixed. By taking products of such reflections, we obtain every permutation in S_4 as an element of $\text{Isom}(X)$, and since an isometry is determined by its action on four points, the action on the vertices determines the isometry. Thus $\text{Isom}(X)$ is isomorphic to S_4. Even isometries correspond exactly to even permutation of vertices. Thus the $Isom^+(X)$ is isomorphic to A_4.

The cube/octahedron. We saw that there are 24 even isometries of the cube (or of the octahedron, its dual), and we mentioned they form the symmetric group S_4. To see this more carefully, fix a face of the cube, and label its vertices as $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4$. For each $1 \leq i \leq 4$, let ℓ_i be the line passing through $\mathbf{x}_i$ and $-\mathbf{x}_i$ (and hence $\mathbf{0}$ as well).

Now given $i \neq j$, there are two possibilities; either $\mathbf{x}_i$ and $\mathbf{x}_j$ are adjacent, or they are not. If they are adjacent, then let $\mathbf{y}$ be the midpoint of the edge of the cube that runs from one to the other; if they are not adjacent, then $\mathbf{x}_i$ and $-\mathbf{x}_j$ are adjacent, and we take $\mathbf{y}$ to be the midpoint of the edge running from $\mathbf{x}_i$ to $-\mathbf{x}_j$. Let ℓ_{ij} be the line through $\mathbf{y}$ and $-\mathbf{y}$, and observe that if k

[5] Another way of thinking of this is that if we rotate a clock around the line ℓ, then we flip the clock over in the plane P, but we also move our point of observation from one side of P to the other, so the clock still appears to be running clockwise.

is neither i nor j, then ℓ_k is perpendicular to ℓ_{ij}, and hence I_{ij} acts on the vertices as the transposition $(i\ j)$. It follows that every permutation of the diagonals ℓ_i can be realised by an isometry I of the cube, and if I is even, it is unique.

The dodecahedron/icosahedron. We saw that there are 60 even isometries of the dodecahedron (or of the icosahedron, its dual). To see that they form the alternating group A_5, one partitions the 20 vertices of the dodecahedron into five classes V_1, V_2, V_3, V_4, V_5 of four vertices each, with the property that for each $1 \leq i \leq 5$, each of the 12 faces of the dodecahedron is adjacent to exactly one vertex from V_i. One then observes that for each i, the four points in V_i are the vertices of a regular tetrahedron X_i. Finally, one shows that every even permutation of the five tetrahedra X_1, X_2, X_3, X_4, X_5 can be realised by a rotation in $SO(3)$.

Thus we found finite subgroups of $SO(3)$ which are isomorphic to A_4, S_4, and A_5. We observe that the subgroups of *all* isometries of the tetrahedron, cube, and dodecahedron are isomorphic to S_4, $S_4 \times \mathbb{Z}/2\mathbb{Z}$, and $A_5 \times \mathbb{Z}/2\mathbb{Z}$, respectively. The last two of these have the same sort of structure as $O(3)$, in that they are generated by their even subgroup and their centre $\{\mathrm{Id}, C\}$, where $C\colon \mathbf{x} \mapsto -\mathbf{x}$ is the central symmetry, which is a symmetry of the cube and dodecahedron (and octahedron and icosahedron), but not the tetrahedron.

THEOREM 8.2. *Every non-trivial finite subgroup of $SO(3)$ is isomorphic to C_n, D_n, A_4, S_4, or A_5. In fact, this isomorphism can be realised by a conjugacy within $SO(3)$: if $G \subset SO(3)$ is a non-trivial finite subgroup, then there exists $R \in SO(3)$ such that $RGR^{-1} \subset SO(3)$ is* equal *to one of the subgroups described above.*

PROOF. The proof comes in two stages. First, we use a combinatorial argument to determine the possible numbers of various types of rotations in G. Having done this, we then show that the group G "is what we think it is"—that is, that it corresponds to the group from the list C_n, D_n, A_4, S_4, A_5 which has the same number of rotations of various orders.

The combinatorial part. Let $F \subset S^2$ be the set of points $\mathbf{x}$ on the sphere such that the line $\ell(\mathbf{x})$ through $\mathbf{x}$ and $\mathbf{0}$ is an axis of rotation for some element of G—that is,

$$F = \bigcup_{R_\theta^\ell \in G} \ell \cap S^2.$$

Observe that F may also be characterised as

$$F = \{\mathbf{x} \in S^2 \mid I\mathbf{x} = \mathbf{x} \text{ for some non-trivial } I \in G\}.$$

Because G is finite, F is finite as well, and given $\mathbf{x} \in F$, every rotation around $\mathbf{x}$ has finite order. In particular, there exists $p = p(\mathbf{x})$ such that the rotations in G which have $\ell = \ell(\mathbf{x})$ as their axis are exactly the rotations

$R^{\ell(\mathbf{x})}_{2\pi k/p(\mathbf{x})}$ for $0 \leq k < p$. To simplify notation, we will write

$$R_{\mathbf{x}} = R^{\ell(\mathbf{x})}_{2\pi/p(\mathbf{x})},$$

so that every rotation around the line through $\mathbf{0}$ and $\mathbf{x}$ can be written as $R^k_{\mathbf{x}}$ for some $0 \leq k < p(\mathbf{x})$.

Define an equivalence relation on F as follows: $\mathbf{x} \sim \mathbf{y}$ if and only if there exists $I \in G$ such that $I\mathbf{x} = \mathbf{y}$. Recall that in this case $I \circ R^{\ell(\mathbf{x})}_\theta \circ I^{-1} = R^{\ell(\mathbf{y})}_\theta$, and so $\mathbf{x} \sim \mathbf{y}$ implies that $p(\mathbf{x}) = p(\mathbf{y})$.

Now choose a point $\mathbf{z} \in S^2$ which is very close to $\mathbf{x}$ (but not equal to $\mathbf{x}$ or any other element of F). In particular, suppose that $\gamma = d(\mathbf{z}, \mathbf{x}) < d(\mathbf{z}, \mathbf{y})/2$ for every $\mathbf{y} \in F$, $\mathbf{y} \neq \mathbf{x}$. Observe that $\text{Orb}(\mathbf{z}) = \{I\mathbf{z} \mid I \in G\}$ is a set of n points, where $n = |G|$; this is because $I_1\mathbf{z} = I_2\mathbf{z}$ for $I_1 \neq I_2$ would imply $I_2^{-1} \circ I_1\mathbf{z} = \mathbf{z}$, and hence $\mathbf{z} \in F$, since $I = I_2^{-1} \circ I_1$ is a non-trivial element in G.

It follows that for each $\mathbf{z}' \in \text{Orb}(\mathbf{z})$ there exists a unique $I \in G$ such that $I\mathbf{z} = \mathbf{z}$. Let $\sigma(\mathbf{z}') = I\mathbf{x}$, and observe that $\sigma(\mathbf{z}') \sim \mathbf{x}$ for all $\mathbf{z}' \in \text{Orb}(\mathbf{z})$. We show that the map $\sigma\colon \text{Orb}(\mathbf{z}) \to \{\mathbf{y} \in S^2 \mid \mathbf{y} \sim \mathbf{x}\}$ is p-to-1, i.e., that for each $\mathbf{y} \sim \mathbf{x}$ there are exactly $p(\mathbf{x})$ different points $\mathbf{z}' \in \text{Orb}(\mathbf{z})$ such that $\sigma(\mathbf{z}') = \mathbf{y}$. This is easy to see by observing that if $I\mathbf{x} = \mathbf{y}$, then $\sigma(\mathbf{z}') = \mathbf{y}$ if and only if $\sigma(I^{-1}\mathbf{z}') = \mathbf{x}$; that is, if $I^{-1}\mathbf{z}' = R^{\ell(\mathbf{x})}_{2\pi k/p(\mathbf{x})}$ for some $0 \leq k < p(\mathbf{x})$.

This counting argument boils down to the following: the orbit of $\mathbf{z}$ contains n points, and each point $\mathbf{y} \sim \mathbf{x}$ has $p = p(\mathbf{x})$ of these points in its immediate vicinity. Thus the number of such points $\mathbf{y}$ is n/p. (We could also have observed that the set of rotations in G which fix $\mathbf{x}$ is a subgroup of order p, and that its index n/p—that is, the number of cosets—is exactly the number of points $\mathbf{y} \sim \mathbf{x}$.)

We now proceed with a counting argument. G contains $n-1$ non-trivial rotations, each of which is determined by its axis ℓ and angle of rotation θ. Each axis ℓ corresponds to two points $\mathbf{x}, -\mathbf{x} \in F$, and for each such ℓ there are $p(\mathbf{x}) - 1$ non-zero angles of rotation available; it follows that the total number of non-trivial rotations is

$$n - 1 = \frac{1}{2}\sum_{\mathbf{x} \in F}(p(\mathbf{x}) - 1). \tag{8.1}$$

Group the points in F by equivalence classes under $\sim$. That is, fix a subset $\{\mathbf{x}_1, \dots \mathbf{x}_k\} \subset F$ such that every $\mathbf{y} \in F$ is equivalent to exactly one of the $\mathbf{x}_i$, and let $p_i = p(\mathbf{x}_i)$. Then (8.1) may be written as

$$\begin{aligned}2(n-1) &= \sum_{i=1}^k |\{\mathbf{y} \in F \mid \mathbf{y} \sim \mathbf{x}_i\}|(p(\mathbf{x}_i) - 1) \\ &= \sum_{i=1}^k \frac{n}{p_i}(p_i - 1).\end{aligned}$$

Thus we have reduced the combinatorial part of things to the question of finding integers $p_1, \dots, p_k$ such that

$$2 - \frac{2}{n} = \sum_{i=1}^{k} \left(1 - \frac{1}{p_i}\right). \tag{8.2}$$

Observe that since the left-hand side of (8.2) is strictly less than 2 and since $1 - 1/p \geq 1/2$ for all $p \geq 2$, we must have $k \leq 3$. Furthermore, since $n = 1$ corresponds to the trivial group, we may assume that $n \geq 2$, and hence the left-hand side is at least 1, whence we have $k \geq 2$.

Case one: $k = 2$. In this case (8.2) becomes

$$2 - \frac{2}{n} = 1 - \frac{1}{p_1} + 1 - \frac{1}{p_2} = 2 - \frac{1}{p_1} - \frac{1}{p_2},$$

and hence

$$\frac{n}{p_1} + \frac{n}{p_2} = 2.$$

Since $n/p(\mathbf{x})$ is a positive integer for every $\mathbf{x} \in F$, we conclude that $p_1 = p_2 = n$. Thus F contains just two points $\mathbf{x}_1$ and $\mathbf{x}_2$, which must be antipodal ($\mathbf{x}_2 = -\mathbf{x}_1$), and is the cyclic group C_n of rotations around the line through $\mathbf{x}_1$ and $\mathbf{x}_2$ (and $\mathbf{0}$).

Case two: $k = 3$. In this case we have

$$2 - \frac{2}{n} = 1 - \frac{1}{p_1} + 1 - \frac{1}{p_2} + 1 - \frac{1}{p_3} = 3 - \frac{1}{p_1} - \frac{1}{p_2} - \frac{1}{p_3},$$

which yields

$$\frac{1}{p_1} + \frac{1}{p_2} + \frac{1}{p_3} = 1 + \frac{2}{n} > 1. \tag{8.3}$$

Observe that if $p_i \geq 3$ for each $i = 1, 2, 3$, then the left-hand side is at most 1, a contradiction. Thus without loss of generality we may assume that $p_3 = 2$, and (8.3) becomes

$$\frac{1}{p_1} + \frac{1}{p_2} = \frac{1}{2} + \frac{2}{n}. \tag{8.4}$$

Multiplying through by $2p_1p_2$, we obtain

$$2p_2 + 2p_1 = p_1p_2 + \frac{4p_1p_2}{n},$$

which may be rewritten as

$$4 - \frac{4p_1p_2}{n} = p_1p_2 - 2p_1 - 2p_2 + 4 = (p_1 - 2)(p_2 - 2).$$

If $p_1 \geq 4$ and $p_2 \geq 4$, then the right-hand side is at least 4, a contradiction since the left-hand side is clearly less than 4. Thus without loss of generality, we have $p_2 = 2$, in which case $p_1 = n/2$, or $p_2 = 3$, in which case $p_1 = 3, 4$, or 5.

Recall that since $p_3 = 2$ in each of these cases, (8.4) yields

$$n = 2\left(\frac{1}{p_1} + \frac{1}{p_2} - \frac{1}{2}\right)^{-1}.$$

We can now list all the possible solutions of (8.2) with $k = 3$.

p_1	p_2	p_3	n
$n/2$	2	2	arbitrary even
3	3	2	12
4	3	2	24
5	3	2	60

At this point it is obvious what groups we think we have our hands on: the first line appears to correspond to $D_{n/2}$, the second to A_4, the third to S_4, and the fourth to A_5; in all cases not only the abstract isomorphism types of the groups are fixed but also conjugacy classes of their embeddings into $SO(3)$; namely, they are isometry groups of a regular $n/2$-gon, regular tetrahedron, regular octahedron/cube and regular icosahedron/dodecahedron correspondingly. That these embeddings actually exist will be proved next time. □

Lecture 9. Monday, September 21

a. Completion of the proof of Theorem 8.2. Having completed the combinatorial part of the proof, we turn now to the task of showing that the combinatorial properties enumerated last time (see table on p. 73) do in fact determine the groups we think they do. We use all the notation from last time.

In each of the four cases which we must examine, there are three equivalence classes of points in F. Thus we may decompose $F = X_1 \cup X_2 \cup X_3$, where the sets X_i are disjoint from each other, and where any two points in the same X_i are equivalent: for every $\mathbf{x}, \mathbf{y} \in X_i$ there exists $I \in G$ such that $I\mathbf{x} = \mathbf{y}$. In particular, every rotation $I \in G$ preserves X_i: $I(X_i) = X_i$.

Recall that we write

$$R_{\mathbf{x}} = R_{2\pi/p(\mathbf{x})}^{\ell(\mathbf{x})}.$$

Given $i = 1, 2, 3$, let G_i denote the set of rotations in G of the form $R_{\mathbf{x}}^k$ for $0 \leq k < p_i$ and $\mathbf{x} \in X_i$, and observe that G is the disjoint union of the sets G_i.

Recall that the number of points in X_i is given by $|X_i| = n/p$, where n is the order of G.

Case one: $p_1 = n/2$, $p_2 = p_3 = 2$. In this case X_1 has just 2 points, and G contains rotations of order $n/2$ around each of the corresponding axes. X_2 and X_3 both have $n/2$ points, and G contains rotations of order 2 around each one.

Fix $\mathbf{p} \in X_1$, and observe that as long as $n > 4$, we have $p_1 \neq p_i$ for $i \neq 1$; consequently, since $p(-\mathbf{p}) = p(\mathbf{p})$, we must have $X_1 = \{\mathbf{p}, -\mathbf{p}\}$. Since every rotation in G preserves X_1, and since all rotations in both G_2 and G_3 are of order 2, we see that all points in X_2 and X_3 must lie on the "equator" between the poles $\mathbf{p}$ and $-\mathbf{p}$—that is, the great circle in which the sphere intersects the plane through $\mathbf{0}$ perpendicular to $\ell(\mathbf{p})$.

Let $\mathbf{q} \in X_2$ be any such point, and observe that since X_2 is preserved by the action of G_1, we have

$$X_2 = \{R_{\mathbf{p}}^k \mathbf{q} \mid 0 \leq k < n/2\}.$$

Thus the points in X_2 form the vertices of a regular $n/2$-gon; call it Q. Furthermore, if $\mathbf{q}' \in X_3$, then a similar argument implies

$$X_3 = \{R_{\mathbf{p}}^k \mathbf{q}' \mid 0 \leq k < n\}. \tag{9.1}$$

Since X_3 is preserved by the rotations in G_2, we have

$$R_{\mathbf{q}} \mathbf{q}' \in X_3.$$

Now $R_{\mathbf{q}}^2 = \text{Id}$, and so it follows from (9.1) that $\ell(\mathbf{q}')$ contains the midpoint of one of the edges of Q. It follows that G is the group of symmetries of a regular $n/2$-gon—that is, the dihedral group $D_{n/2}$.

Case two: $p_1 = 3, p_2 = 3, p_3 = 2, n = 12$. In this case X_1 and X_2 each have 4 points, and G_1 and G_2 contain rotations of order 3 around each of the corresponding axes. X_3 has 6 points, and G_3 contains rotations of order 2 around each one.

Fix $\mathbf{p} \in X_1$, and observe that since X_1 has 4 points, there exists $\mathbf{q} \in X_1$ such that $\mathbf{q} \neq -\mathbf{p}, \mathbf{p}$. Furthermore, we have $R_\mathbf{p}\mathbf{q}, R_\mathbf{p}^2\mathbf{q} \in X_1$. Since $\mathbf{q} \notin \ell(\mathbf{p})$, the four points $\{\mathbf{p}, \mathbf{q}, R_\mathbf{p}\mathbf{q}, R_\mathbf{p}^2\mathbf{q}\}$ are all distinct and lie in the same equivalence class; hence $X_1 = \{\mathbf{p}, \mathbf{q}, R_\mathbf{p}\mathbf{q}, R_\mathbf{p}^2\mathbf{q}\}$.

Now observe that $R_\mathbf{q}(R_\mathbf{p}\mathbf{q}) \in X_1$, and that this point is not equal to $\mathbf{q}$, $R_\mathbf{p}\mathbf{q}$, or $R_\mathbf{p}^2\mathbf{q}$. It follows that $\mathbf{p} = R_\mathbf{q}(R_\mathbf{p}\mathbf{q})$, or equivalently, the points $\mathbf{p}, \mathbf{q}, R_\mathbf{p}\mathbf{q}$ form the vertices of an equilateral triangle. A similar argument applies to any three points in X_1, and hence the four points in X_1 are the vertices of a regular tetrahedron; call it Q.

Given $\mathbf{x} \in X_3$ and $R_\mathbf{x} \in G_3$, observe that $R_\mathbf{x}\mathbf{p} \in \{\mathbf{q}, R_\mathbf{p}\mathbf{q}, R_\mathbf{p}^2\mathbf{q}\}$. Since $R_\mathbf{x}$ is rotation by π around the axis $\ell = \ell(\mathbf{x})$, it follows that ℓ must pass through the midpoint of the line segment from $\mathbf{p}$ to $R_\mathbf{x}\mathbf{p}$, which is one of the edges of Q. Thus the points in X_3 are precisely the points on the sphere which one obtains by taking the radial projection $\mathbf{z} \mapsto \mathbf{z}/\|\mathbf{z}\|$ of the midpoints of the edges of Q.

We also see that since X_1 does not contain any antipodal pairs $\mathbf{y}, -\mathbf{y}$, we must have $X_2 = \{-\mathbf{p}, -\mathbf{q}, -R_\mathbf{p}\mathbf{q}, -R_\mathbf{p}^2\mathbf{q}\}$. This gives a complete description of all elements of G, and we see that G is exactly the group of isometries of the regular tetrahedron Q—that is, S_4.

Case three: $p_1 = 4, p_2 = 3, p_3 = 2, n = 24$. In this case X_1 has 6 points and G_1 contains rotations of order 4 around each of the corresponding axes. X_2 has 8 points and G_2 has rotations of order 3; X_3 has 12 points and G_3 has rotations of order 2.

We will show that the points in X_1 are the vertices of an octahedron, for which points in X_2 correspond to centres of faces, and points in X_3 to midpoints of edges.

Choose $\mathbf{p}, \mathbf{q} \in X_1$ such that $\mathbf{q} \notin \ell(\mathbf{p})$. Then as before, $R_\mathbf{p}\mathbf{q}$, $R_\mathbf{p}^2\mathbf{q}$, and $R_\mathbf{p}^3\mathbf{q}$ are in X_1 and are not equal to either $\mathbf{p}$ or $-\mathbf{p}$; furthermore, $p(-\mathbf{p}) = p(\mathbf{p}) = 4$, and hence $-\mathbf{p} \in X_1$. It follows that

$$X_1 = \{\mathbf{p}, -\mathbf{p}, \mathbf{q}, R_\mathbf{p}\mathbf{q}, R_\mathbf{p}^2\mathbf{q}, R_\mathbf{p}^3\mathbf{q}\}.$$

Observing that $R_\mathbf{q}^k\mathbf{p} \in X_1$ for $0 \leq k < 4$, we see that the points in X_1 are the vertices of an octahedron; call it Q. The argument from the previous case shows that given any $\mathbf{x} \in X_3$, the line $\ell(\mathbf{x})$ contains the midpoint of an edge of Q; there are 12 such points and 12 such edges.

What about X_2? Label the 6 vertices of Q as follows: $\mathbf{p}$ is "1"; $\mathbf{q}$, $R_\mathbf{p}\mathbf{q}$, $R_\mathbf{p}^2\mathbf{q}$, $R_\mathbf{p}^3\mathbf{q}$ are "2", "3", "4", and "5", respectively; and $-\mathbf{p}$ is "6". Observe that every isometry of the octahedron corresponds to a permutation of the set $\{1, 2, 3, 4, 5, 6\}$, and thus we can identify rotations in G_1 and G_3 with elements of S_6. It only remains to show that rotations in G_2 can be so identified.

Let R_1 be the rotation around $\ell(R_{\mathbf{p}}\mathbf{q})$ which accomplishes the permutation $(1\ 2\ 6\ 4)$, and let R_2 be the rotation around $\ell(\mathbf{q})$ which accomplishes the permutation $(1\ 5\ 6\ 3)$. Then we see that $R = R_2 \circ R_1$ is a rotation which accomplishes the permutation

$$(1\ 2\ 6\ 4)(1\ 5\ 6\ 3) = (1\ 2\ 3)(4\ 5\ 6).$$

(We multiply elements in S_6 from left to right.) It follows that R is a rotation of order 3, and so $R \in G_2$; furthermore, the axis of rotation of R passes through the centre of a face of Q. A similar argument shows that every line through $\mathbf{0}$ and the centre of a face of Q is the axis of a rotation in G_2, and since such lines intersect the sphere in 8 points, we have found X_2.

Observe that the points in X_2 form the vertices of a cube which is dual to the octahedron just constructed.

Case four: $p_1 = 5, p_2 = 3, p_3 = 2, n = 60$. In this case X_1 has 12 points and G_1 contains rotations of order 5 around each of the corresponding axes. X_2 has 20 points and G_2 has rotations of order 3; X_3 has 30 points and G_3 has rotations of order 2.

The proof here follows the same lines as in the previous case—the points in X_1 form the vertices of an icosahedron, and the points in X_2 form the vertices of a dodecahedron. G can be shown to be the set of isometries of either of these polyhedra. Details are left as an exercise for the reader. □

b. Regular polyhedra. In fact, in the course of classifying the finite subgroups of $\mathrm{Isom}^+(\mathbb{R}^3)$, we have also done virtually all the work needed to classify the convex regular polyhedra in $\mathbb{R}^3$. Of course, in order to make this precise we need to define the terms involved.

It is helpful to first consider the two-dimensional case, in order to see how the definitions work. In two dimensions, each line ℓ divides the plane into two regions; if we associate a direction to ℓ, then we may say that one of these half-planes lies to the left of ℓ, and the other lies to the right. Now given a collection of lines $\ell_1, \dots, \ell_n$, we may consider the set of points in $\mathbb{R}^2$ which lie to the right of every ℓ_i. Call this set X; if X is bounded and non-empty, we say that the boundary of X is a *convex polygon.* The segments of the lines ℓ_i which intersect the boundary of X are called the *edges* of X, and the points of intersection $\ell_i \cap \ell_j$ which lie on the boundary of X are the *vertices* of X.

Passing to three dimensions, we replace lines with planes. A *convex polyhedron* is defined by a collection of planes $P_1, \dots, P_n$ in $\mathbb{R}^3$. Giving each plane an orientation—say, painting one side red and the other side blue—we may consider the set X of points in $\mathbb{R}^3$ which see the red side of every P_i. If X is bounded and non-empty, its boundary is a convex polyhedron. Each P_i intersects the polyhedron in a convex polygon; these are the *faces* of the polyhedron. The edges are the segments of the lines of intersection $P_i \cap P_j$ which lie on the boundary—that is, the intersection of two neighbouring faces—and the vertices are the points of intersection of three or more faces.

A polyhedron is *regular* if "all faces look the same". This can be made precise in at least two different (non-equivalent) ways.

DEFINITION 9.1. A convex polyhedron Q is *combinatorially regular* if there exist integers p and q such that every face of Q has p edges, and every vertex of Q touches q edges and q faces. The pair (p, q) is known as the *Schläfli symbol* of Q.

THEOREM 9.2. *Every combinatorially regular convex polyhedron has one of the following five Schläfli symbols:* $(3,3)$*,* $(4,3)$*,* $(3,4)$*,* $(5,3)$*, or* $(3,5)$.

PROOF. Let F be the number of faces, E the number of edges, and V the number of vertices. These numbers are related by the *Euler characteristic* of a polyhedron:

$$F - E + V = 2. \tag{9.2}$$

We will prove this equality (called the *Euler theorem*) shortly. Since every edge meets two faces, we have $2E = pF$. Similarly, every vertex meets q faces, and so $qV = pF$. Write $n = pF = 2E = qV$, so $F = n/p$, $E = n/2$, and $V = n/q$. Then (9.2) becomes

$$\frac{n}{p} - \frac{n}{2} + \frac{n}{q} = 2. \tag{9.3}$$

Observe that this is exactly the equation we obtained as (8.4) in the proof of Theorem 8.2, where we showed that the only solutions are exactly the ones listed above.

Now let us prove Euler theorem. [6] Let $\Sigma\alpha$ be the sum of all angles of all faces of a polyhedron with V vertices, E edges and F faces. Compute $\Sigma\alpha$ in two different ways:

- by computing the sum of angles of each face and adding them up to obtain $\Sigma\alpha = 2\pi(E - F)$;
- by deforming the polyhedron and projecting it to one face to obtain $\Sigma\alpha = 2\pi V - 4\pi$.

Let faces of a polyhedron have $n_1, n_2, \dots, n_F$ sides. The sum of angles of the $i - th$ face is $\pi(n_i - 2)$, and the total sum $\Sigma\alpha = \sum_{i=1}^{F} \pi(n_i - 2) = 2\pi(E - F)$ since each edge was counted twice. To compute $\Sigma\alpha$ in the other way, we notice that by deforming the polyhedron and projecting it to one face we do not change the sum of the angle of each face. Let the face we were projecting to be an n-gon. Then the remaining $V - n$ vertices are inside this face, and the total sum of the angles at these vertices is $2\pi(V - n)$. The sum of angles of the n-gon should be counted twice, and it is $2\pi(n - 2)$. Then $\Sigma\alpha = 2\pi(V - n) + 2\pi(n - 2) = 2\pi V - 4\pi$. Comparing the two expressions obtain $E - F = V - 2$.

□

[6] This elegant proof that did not appear in the lecture was shown to us by Svetlana Katok.

The five possibilities in Theorem 9.2 correspond respectively to the tetrahedron, the cube, the octahedron, the dodecahedron, and the icosahedron.

Observe that *any* tetrahedron is a combinatorially regular convex polyhedron, whether or not the faces are all equilateral triangles. A similar statement applies to each of the five polyhedra listed above; although we are most familiar with their highly symmetric versions, we can deform those versions without destroying combinatorial regularity.

Notice that all polyhedra with the same Schläfli symbol are combinatorially equivalent, i.e one can associate faces with those of the standard model in such a way that this correspondence naturally extend to that between edges (intersections of adjacent faces) and vertices (intersections of adjacent edges). This is obvious in the $(3,3)$ case since all polyhedra with four faces are tetrahedra. In other cases one proceeds as in glueing of a paper model of a polyhedron from a flat cutting. For example, for $(4,3)$ one starts with a quadrilateral face, attaches four more to its edges, notices that edges coming from vertices of the original quadrilateral must match and four remaining edges are the sides of the only remaining face producing a combinatorial cube. For $(3,4)$ one proceed in the dual way, starting with four triangles attached to a vertex, then taking the only vertex that does not appear among vertices of these triangle, attach four triangles to it and notice that two “pyramids without bases” must be attached to each other producing a combinatorial octahedron. The argument for $(5,3)$ is similar to the one for cube: Attach five pentagons to a pentagon obtaining a figure with ten free edges connecting cyclically ten vertices. At five alternating vertices two pentagons have already been attached; remaining five are free. Attaching five pentagons to the vertices of the first set one sees that edges attached to free vertices must match leaving only five free edges that bound the only remaining face thus producing a combinatorial dodecahedron. We leave the $(5,3)$ case as an exercise.

Definition 9.3. A convex polyhedron is *geometrically regular* if the isometry group acts transitively on faces edges and vertices—that is, given any two faces there is an isometry of the polyhedron which takes one to the other, and similarly for edges and vertices.

Any geometrically regular polyhedron is combinatorially regular. This would *not* be the case if we weakened the above definition by only requiring transitivity on faces or on vertices. To see this, observe that if Q is a triangular prism, then $\operatorname{Isom}(Q)$ acts transitively on vertices but not on faces, and Q is not combinatorially regular, since two faces are triangles and three are rectangles. Similarly, if Q' is the dual solid to Q—that is, a “double tetrahedron” with six faces which are isosceles triangles, then $\operatorname{Isom}(Q')$ acts transitively on faces but not on vertices, and once again Q' is not combinatorially regular, since two vertices have three adjacent faces, while three vertices have four such faces.

Every combinatorial type in Theorem 9.2 has a unique up to a similarity (isometry and homothety) geometrically regular realization; these are the five Platonic solids with which we are already familiar.

While existence is self-evident uniqueness requires a proof. First we need to identify subgroups of full isometry groups of Platonic solids that act transitively on faces, vertices and edges. Only the full group and the subgroups of all even isometries have this property. This follows from the fact that the order of such a subgroup must divide the numbers of faces, edges and vertices. This implies that the isometry group of a geometrically regular polyhedron contains subgroups fixing a vertex and cyclically interchanging faces attached to that vertex. Hence all angles of all faces are equal that those are regular polygons. After that uniqueness follows as in the argument for uniqueness of combinatorial type above. Alternatively, one can take a vertex that must lie on the axis of a rotational symmetry, apply the isometry group to obtain all other vertices and see by direct inspection that they form one of the Platonic bodies.

A peculiar fact is that in the case of tetrahedron transitivity on *both* faces and edges does not guarantee regularity. Examples are tetrahedra with vertices (x, y, z), $(x, -y, -z)$, $(-x, -y, z)$ and $(-x, y, -z)$ for any triple of non-zero numbers x, y, z.

However, for the remaining Schläfli symbols transitivity on vertices and faces is sufficient. This distinction is due to the fact that A_4 contains a normal subgroup of four elements (rotations of order two plus identity) that acts transitively on the vertices and faces but on the edges of a tetrahedron. In the remaining cases the order of such a subgroup must be a multiple of 6 and 8 (in the $(4, 3)$ and $(3, 4)$ cases) and of 12 and 20 (in the $(5, 3)$ and $(3, 5)$ cases) and hence be a either the full isometry group (of order 48 and 120 correspondingly) or its index two subgroup of even isometries.

One could generalize the notion of a geometrically regular convex polyhedron in various ways. For example, we could weaken the regularity slightly by allowing the sets of vertices and faces to comprise not one but two orbits under the action of the symmetry group, and so consider *semi-regular polyhedra.* Or we could consider analogous constructions in higher dimensions, and study the *regular polytopes.* We could also do away with the requirement of convexity and study non-convex polyhedra. . .

c. Completion of classification of isometries of $\mathbb{R}^3$. In Lecture 8(a), we completed the classification of odd isometries of $\mathbb{R}^3$. We have already classified even isometries with fixed points; it only remains to describe even isometries without fixed points.

Aside from translations, we have already seen one such isometry—the product of a rotation R and a translation T, where R and T commute. Geometrically, this means that the axis ℓ of the rotation R is parallel to the direction of the vector $\mathbf{v}$ which specifies the translation. The product $T \circ R$ may be called a *screw motion*, a *twist*, or a *glide rotation.*

PROPOSITION 9.4. *Every element $I \in \mathrm{Isom}^+(\mathbb{R}^3)$ is a screw motion $T \circ R$ (if the rotation or translation part is trivial, then I is a translation or a rotation, respectively).*

PROOF. Given $I \in \mathrm{Isom}^+(\mathbb{R}^3)$, we can decompose I as the product of four reflections by Proposition 7.2:

$$I = L_1 \circ L_2 \circ L_3 \circ L_4 \tag{9.4}$$

Write P_i for the plane corresponding to L_i, and write ℓ_{ij} for the line of intersection of P_i and P_j. Then we have $L_1 \circ L_2 = L_1' \circ L_2'$ whenever the corresponding planes P_1' and P_2' intersect at the appropriate angle in the line ℓ_{12}, and similarly for $L_3 \circ L_4$.

Let P_2' be the plane parallel to ℓ_{34} that contains ℓ_{12}, and let P_3' be the plane parallel to ℓ_{12} that contains ℓ_{34} (if ℓ_{12} and ℓ_{34} are parallel, rather than skew, then P_2' and P_3' are not uniquely determined). Then P_2' and P_3' are parallel, and we have

$$I = (L_1 \circ L_2) \circ (L_3 \circ L_4) = (L_1' \circ L_2') \circ (L_3' \circ L_4') = L_1' \circ T \circ L_4',$$

where $T = L_2' \circ L_3'$ is a translation. As in Lecture 5(c), we may decompose the translation T into parts which are parallel to and orthogonal to the plane P_1', and hence obtain $L_1' \circ T = L_1'' \circ T' = T' \circ L_1''$, where L_1'' is reflection in a plane P_1'' parallel to P_1', and T' is translation by a vector parallel to P_1''.

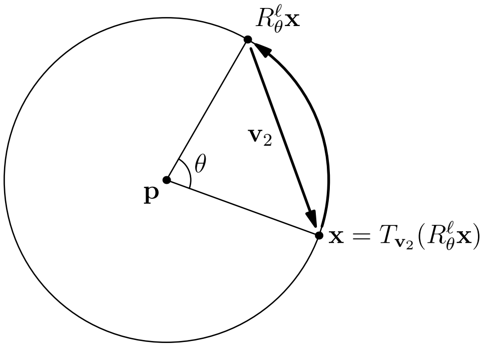

$R_\theta^\ell \mathbf{x}$
$\mathbf{v}_2$
θ
$\mathbf{p}$
$\mathbf{x} = T_{\mathbf{v}_2}(R_\theta^\ell \mathbf{x})$

FIGURE 2.9. The composition of a translation and a rotation is a rotation.

It follows that $I = T' \circ L_1'' \circ L_4'$ is either the product of two translations (if P_1'' and P_4' are parallel) or of a translation and a rotation (if they are not). In the former case, I is a translation. In the latter, let ℓ denote the line of intersection of P_1'' and P_4', so $I = T' \circ R_\theta^\ell$ for some θ.

Let $\mathbf{v}$ be the translation vector for T', and decompose $\mathbf{v}$ as $\mathbf{v}_1 + \mathbf{v}_2$, where $\mathbf{v}_1$ is parallel to ℓ and $\mathbf{v}_2$ is perpendicular to ℓ. Then $T_{\mathbf{v}_2} \circ R_\theta^\ell$ is rotation by θ in a line parallel to ℓ. To see this, consider Figure 2.9, which shows a plane P perpendicular to ℓ. The centre of rotation $\mathbf{p}$ is the point of intersection of ℓ with P. The radius of the circle shown is uniquely determined by the requirement that a chord of length $\|\mathbf{v}_2\|$ subtend an angle θ, and the position of $\mathbf{x}$ on this circle is uniquely determined by the requirement that the vector from $\mathbf{x}$ to $R_\theta^\ell \mathbf{x}$ be equal to $-\mathbf{v}_2$. As an isometry on the plane P, we see that $T_{\mathbf{v}_2} \circ R_\theta^\ell$ is an even isometry with a fixed point $\mathbf{x}$; hence it is rotation around $\mathbf{x}$ by an angle $\alpha(T_{\mathbf{v}_2} \circ R_\theta^\ell) = \theta$.

Let ℓ' be the line through $\mathbf{x}$ perpendicular to P; it follows that

$$T_{\mathbf{v}_2} \circ R_\theta^\ell = R_\theta^{\ell'},$$

and hence

$$I = T_{\mathbf{v}_1} \circ T_{\mathbf{v}_2} \circ R_\theta^\ell = T_{\mathbf{v}_1} \circ R_\theta^{\ell'}$$

is a screw motion. □

Remark. Even though Proposition 9.4 deals with orientation-preserving isometries, the proof relies on the decomposition of every such isometry into reflections, which are orientation-reversing. The fundamental role played by reflections is analogous to the role played by transpositions in S_n. In both cases the group is generated by odd involutions; we will see this phenomenon repeated a little later on, when we consider conformal geometry.

d. From synthetic to algebraic: Scalar products. The synthetic approach we have been pursuing becomes more and more cumbersome when we pass to higher dimensions. In order to deal with this more general setting, we will take a more algebraic approach, using tools from linear algebra to describe and study isometries.

Recall that a *linear map* from one real vector space to another is a map A such that $A(\lambda\mathbf{v} + \mathbf{w}) = \lambda A(\mathbf{v}) + A(\mathbf{w})$ for every real number λ and every pair of vectors $\mathbf{v}, \mathbf{w}$. Equivalently, a linear map is a map which fixes the origin ($A\mathbf{0} = \mathbf{0}$) and takes lines to lines (for every line ℓ, the image $A(\ell)$ is also a line). An *affine map* is a map with the latter property—lines are mapped to lines—but not necessarily the former.

Proposition 9.5. *Isometries of $\mathbb{R}^d$ are affine.*

Proof. Given $I \in \operatorname{Isom}(\mathbb{R}^d)$, it suffices to show that the points $I\mathbf{x}, I\mathbf{y}, I\mathbf{z}$ are collinear whenever $\mathbf{x}, \mathbf{y}, \mathbf{z}$ are. Now $\mathbf{y}$ lies on the line segment from $\mathbf{x}$ to $\mathbf{z}$ if and only if equality holds in the triangle inequality—that is, if and only if

$$d(\mathbf{x}, \mathbf{z}) = d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z}). \tag{9.5}$$

Thus if $\mathbf{x}, \mathbf{y}, \mathbf{z}$ are collinear, one of them lies on the line segment between the other two; without loss of generality, suppose $\mathbf{y}$ lies between $\mathbf{x}$ and $\mathbf{z}$, so that (9.5) holds. Then since I is an isometry, we have

$$d(I\mathbf{x}, I\mathbf{z}) = d(I\mathbf{x}, I\mathbf{y}) + d(I\mathbf{y}, I\mathbf{z}),$$

and hence $I\mathbf{x}, I\mathbf{y}, I\mathbf{z}$ are collinear. □

Isometries have a property that affine maps do not have—they preserve angles. In fact, they can be characterised using the *scalar product*, which for any two vectors $\mathbf{v}, \mathbf{w} \in \mathbb{R}^d$ is the real number

$$\langle \mathbf{v}, \mathbf{w} \rangle = \sum_{i=1}^{d} v_i w_i. \tag{9.6}$$

The length of a vector is related to the scalar product by the formula $\|\mathbf{v}\|^2 = (\mathbf{v}, \mathbf{v})$, and the angle between two vectors is given by the scalar product using the formula $\langle \mathbf{v}, \mathbf{w} \rangle = \|\mathbf{v}\| \cdot \|\mathbf{w}\| \cos \theta$, where θ is the angle between $\mathbf{v}$ and $\mathbf{w}$.

Thus if $I \colon \mathbb{R}^d \to \mathbb{R}^d$ preserves scalar products—if $\langle I\mathbf{v}, I\mathbf{w} \rangle = \langle \mathbf{v}, \mathbf{w} \rangle$ for every $\mathbf{v}, \mathbf{w}$—then I preserves lengths as well, and hence is an isometry. In fact, this is a two-way street; every isometry preserves not only lengths, but scalar products (and hence angles as well). This is a consequence of the *polarisation identity*

$$\langle \mathbf{v}, \mathbf{w} \rangle = \frac{1}{2}(\|\mathbf{v} + \mathbf{w}\|^2 - \|\mathbf{v}\|^2 - \|\mathbf{w}\|^2), \tag{9.7}$$

which can easily be proved by observing that

$$\langle \mathbf{v} + \mathbf{w}, \mathbf{v} + \mathbf{w} \rangle = \langle \mathbf{v}, \mathbf{v} \rangle + \langle \mathbf{w}, \mathbf{w} \rangle + 2 \langle \mathbf{v}, \mathbf{w} \rangle .$$

The definition of the scalar product in (9.6) relies on the choice of coordinate system in $\mathbb{R}^d$—that is, on the choice of basis. In another coordinate system, we would obtain a different scalar product. However, certain basic properties would still go through, which are encoded in the following definition.

Definition 9.6. A *scalar product* (or *inner product*, or *dot product*) on $\mathbb{R}^d$ is a function $\langle \cdot, \cdot \rangle : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}$ such that the following properties hold for all $\mathbf{u}, \mathbf{v}, \mathbf{w} \in \mathbb{R}^d$ and $\lambda \in \mathbb{R}$.

(1) $\langle \mathbf{v}, \mathbf{v} \rangle \geq 0$, with equality if and only if $\mathbf{v} = \mathbf{0}$.
(2) *Symmetry*: $\langle \mathbf{v}, \mathbf{w} \rangle = \langle \mathbf{w}, \mathbf{v} \rangle$.
(3) *Linearity*: $\langle \mathbf{u} + \mathbf{v}, \mathbf{w} \rangle = \langle \mathbf{u}, \mathbf{w} \rangle + \langle \mathbf{v}, \mathbf{w} \rangle$.
(4) $\langle \lambda \mathbf{v}, \mathbf{w} \rangle = \lambda \langle \mathbf{v}, \mathbf{w} \rangle$.

It follows from symmetry and linearity that the scalar product is actually *bilinear*—that is, it is linear in its second argument as well. The fourth condition follows from the third if we require that the scalar product be a continuous function of its arguments.

Every function $\mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}$ which satisfies properties (2)–(4) is of the form

$$\langle \mathbf{v}, \mathbf{w} \rangle = \sum_{i=1}^{d} \sum_{j=1}^{d} a_{ij} v_i w_j, \tag{9.8}$$

where $A = [a_{ij}]$ is a symmetric $d \times d$ matrix ($a_{ij} = a_{ji}$). This does not yet guarantee property (1); for example, we might define a bilinear form on $\mathbb{R}^2$ by $(\mathbf{v}, \mathbf{w}) = u_1 v_1 - u_2 v_2$. This satisfies properties (2)–(4), but for $\mathbf{v} = (0, 1)$ we have $\langle \mathbf{v}, \mathbf{v} \rangle = -1$, so (1) fails. In the next lecture, we will see what properties A must have in order to define a genuine scalar product.

Lecture 10. Wednesday, September 23

a. Convex polytopes. In Lecture 9(b), we defined a convex polyhedron in $\mathbb{R}^3$ using a finite collection of planes—that is, affine subspaces of codimension 1 and taking intersection of certain half-space into which each plane divides the space. This definition generalizes to higher dimensions.

DEFINITION 10.1. A subset $X \subset \mathbb{R}^n$ is an *affine subspace* if there exists $\mathbf{v} \in \mathbb{R}^n$ such that

$$\mathbf{v} + X = \{\mathbf{v} + \mathbf{x} \mid \mathbf{x} \in X\}$$

is a linear subspace.[7] In this case we see that $\mathbf{w} + X$ is a linear subspace if and only if $-\mathbf{w} \in X$.

If X is an affine subspace, the dimension of X is the dimension of the linear subspace $\mathbf{v} + X$. If $X \subset \mathbb{R}^n$ is an affine subspace of dimension d, we say that $n - d$ is the *codimension* of X.

The affine subspaces of $\mathbb{R}^2$ are lines (which have codimension 1) and points (which have codimension 2). In $\mathbb{R}^3$, an affine subspace of codimension 1 is a plane; codimensions 2 and 3 yield lines and points, respectively.

DEFINITION 10.2. A *convex n-polytope* is a region $Q \subset \mathbb{R}^n$ which is defined by m affine subspaces of codimension 1 also called *hyperplanes* in $\mathbb{R}^n$ as follows: there exist m such subspaces $X_1, \dots, X_m$ be m such that the boundary ∂Q is a subset of $\bigcup_{i=1}^m X_i$, and for which $Q \cap X_i \subset \partial Q$ for every i.

Now we will define k-dimensional faces or k-faces for $k = 0, 1 \dots, n-1$ of an n-polytope. While intuitively this notion looks evident, certain care is needed for a rigorous definition.

First we assume that all hyperplanes $X_1, \dots, X_m$ are essential, i.e. that removing any one of those increase the intersection.

In this case an $n-1$-*face* of Q is the intersection $Q \cap X_i$ for some i.

Now we proceed by induction. Assume that k-faces are defined for polytopes up to the dimension $n-1$. Any hyperplane in $\mathbb{R}^n$ is an $n-1$ affine space that is of course can be identified with $\mathbb{R}^{n-1}$. An $n-1$-face of an n-polytope is an $n-1$ polytope in the corresponding hyperplane so that its k-faces for $k = 0, 1, \dots, n-2$ have been defined. we define a k-face of Q as a k-face of one of its $n-1$ faces.

One need to prove coherence of this definition: If $F \subset X - i \cap X_j$ is a k-face of X_i, it is also a k-face of X_j. To see that notice that any k-face can be represented as in intersection

$$F = X_{i_1} \cap X_{i_2} \cap \dots \cap X_{i_{n-k}} \cap Q$$

for some set of indices $i_1, \dots i_{n-k}$. Furthermore, such an intersection is a k-face if and only if it does not belong to a $k-1$-dimensional affine subspace.

[7] Observe the similarity between the geometric notion of an affine subspace and the algebraic notion of a coset.

We see that in the case $n = 3$, this reduces to our earlier definition of a convex polyhedron. For example, if X_1, X_2, X_3, X_4 are the four planes in $\mathbb{R}^3$ which contain the faces of a tetrahedron Q, then the 1-dimensional faces of Q are the edges of the tetrahedron, which have the form $X_i \cap X_j \cap \partial Q$. Similarly, the 0-dimensional faces are the vertices, which can be written as $X_i \cap X_j \cap X_k \cap \partial Q$.

An alternate (dual) definition may be given using general notions of convexity.

Definition 10.3. A set $X \subset \mathbb{R}^n$ is *convex* if $[\mathbf{x}, \mathbf{y}] \subset X$ for every $\mathbf{x}, \mathbf{y} \in X$, where $[\mathbf{x}, \mathbf{y}]$ is the line segment

$$[\mathbf{x}, \mathbf{y}] = \{t\mathbf{x} + (1-t)\mathbf{y} \mid 0 \leq t \leq 1\}$$

which comprises all *convex combinations* of $\mathbf{x}$ and $\mathbf{y}$. Given an arbitrary set $X \subset \mathbb{R}^n$ (which may or may not be convex), the *convex hull* of X is the intersection of all convex sets which contain X. If $X = \{\mathbf{x}_1, \dots, \mathbf{x}_k\}$ is finite, the convex hull is

$$\left\{ t_1\mathbf{x}_1 + \dots + t_k\mathbf{x}_k \,\middle|\, t_i \geq 0, \sum_{i=1}^{k} t_i = 1 \right\}.$$

If X is convex, then an *extreme point* of X is a point $\mathbf{x} \in X$ such that $\mathbf{x}$ is not in the convex hull of $X \setminus \{\mathbf{x}\}$—that is, $\mathbf{x}$ is not the convex combination of any two distinct points $\mathbf{y} \neq \mathbf{z} \in X$.

Extreme points play the role of essential planes in the earlier definition. Now one takes a finite set S of points in $\mathbb{R}^n$ that does not lie in a hyperplane and such that no point belongs to convex hull of the rest and the points. The convex hull Q of S is a convex n-polytope. Then elements of S are the vertices of Q and k-faces can be defined as follows. Let $K \subset S$ is such that all points in K lie in ak-dimensional affine subspace but not in any $k-1$-dimensional affine subspace. Furthermore, assume that convex hulls of L and $S \setminus L$ are disjoint. Then convex hull of L is a k-face of Q.

b. Transformation groups and symmetries of polytopes. The notion of transformation group has appeared already in several contexts: as permutations groups, isometry groups and matrix groups. Let us formalize this and some related notions.

We say that a group G *acts on* a set X if every $g \in G$ defines a bijection $\tilde{g}\colon X \to X$ with the property that $\widetilde{gh} = \tilde{g} \circ \tilde{h}$. Thus a group action on a set X may be thought of as a homomorphism from G into $S(X)$, the group of bijections of X. We will usually write gx, $g(x)$, or $g \cdot x$ in place of $\tilde{g}(x)$.

Given $x \in X$, the *orbit* of x under the action of G is

$$\mathrm{Orb}(x) = Gx = \{g \cdot x \mid g \in G\}.$$

We say that G acts *transitively* on X is $\mathrm{Orb}(x) = X$ for every $x \in X$; equivalently, for every $x, y \in X$ there exists $g \in G$ such that $g \cdot x = y$.

The *stationary subgroup* $S(x)$ of x is the subgroup of elements that leave x fixed. For a transitive action stationary subgroups of all elements of X are conjugate.

Example 10.4. Fix a convex polytope Q generated by the codimension one affine subspaces $X_1, \dots, X_m$, and observe that given any isometry $I \in \text{Isom}(Q)$ and $1 \leq i \leq m$, there exists $1 \leq \pi_I(i) \leq m$ such that $I(X_i) = X_{\pi_I(i)}$. Thus $\text{Isom}(Q)$ acts on the set $\{X_1, \dots, X_m\}$.

Continuing in this vein, let $\mathcal{F}_k$ be the set of k-dimensional faces of Q. Given a face $F \in \mathcal{F}_k$, we have

$$F = X_{i_1} \cap X_{i_2} \cap \dots \cap X_{i_{n-k}} \cap Q$$

for some set $\{i_1, \dots, i_{n-k}\}$, and hence

$$I(F) = X_{\pi_I(i_1)} \cap \dots \cap X_{\pi_I(i_{n-k})} \cap Q$$

is also a k-dimensional face of Q. It follows that $\text{Isom}(Q)$ acts on $\mathcal{F}_k$ for each $0 \leq k \leq n-1$. This is nothing more than the fact that any symmetry of the polytope Q maps vertices to vertices, edges to edges, and so on.

c. Regular polytopes.

Definition 10.5. A convex polytope Q is *regular* if $\text{Isom}(Q)$ acts transitively on $\mathcal{F}_k$ for every $0 \leq k \leq n-1$.

One might reasonably ask if it suffices to have a transitive action on $\mathcal{F}_k$ for *some* values of k. Indeed, there are a number of polyhedra for which transitivity on vertices and faces implies transitivity on edges. However, this is not the case in full generality, as we saw in the previous lecture; if Q is the tetrahedron with vertices (x, y, z), $(-x, -y, z)$, $(-x, y, -z)$, and $(x, -y, -z)$, where x, y, z are not all equal, then $\text{Isom}(Q)$ acts transitively on $\mathcal{F}_0$ (vertices) and $\mathcal{F}_2$ (faces), but not on $\mathcal{F}_1$ (edges).

Now let us look at regular polytopes in various dimensions.

In $\mathbb{R}^2$, the regular polytopes are just the regular n-gons.

In $\mathbb{R}^3$, the regular polytopes are the five Platonic solids. Three of those, the tetrahedron, the cube, and the octahedron, have analogues in any dimension, which we may denote by S_n, I_n, and O_n, respectively. These can be constructed either explicitly or inductively.

The *n-simplex* S_n has $n+1$ vertices (faces of minimal dimension) and $n+1$ faces of maximal dimension. S_2 is an equilateral triangle and S_3 is a tetrahedron. S_n can be constructed inductively by taking $S_{n-1} \subset \mathbb{R}^{n-1} \subset \mathbb{R}^n$ and adding one of the two points in $\mathbb{R}^n$ which is the same distance from every vertex of S_{n-1} that these vertices are from each other. Alternately, it can be explicitly given as a subset of $\mathbb{R}^{n+1}$; the vertices are the tips of the standard basis vectors taken from the origin.

The *n-cube* I_n has 2^n vertices and $2n$ faces of maximal dimension. It can be constructed inductively as $I_n = I_{n-1} \times [0,1]$ by considering $I_{n-1} \subset \mathbb{R}^{n-1} \subset \mathbb{R}^n$ and adding an extra copy of each vertex in the plane $x_n = 1$.

It can also be explicitly given by taking as the 2^n vertices all points $\mathbf{x} = (x_1, \dots, x_n)$ for which $x_i = \pm 1$ for every i.

The dual of the n-cube is the analogue of the octahedron, denoted O_n, and has $2n$ vertices and 2^n faces of maximal dimension. Inductively, one takes $O_{n-1} \subset \mathbb{R}^{n-1} \subset \mathbb{R}^n$ and adds two more vertices—the two points which are the same distance from every vertex of O_{n-1} as neighbouring vertices in O_{n-1} are from each other. Explicitly, one can take as the $2n$ vertices every point in $\mathbb{R}^n$ which lies on a coordinate axis at a distance of 1 from the origin.

The four-dimensional case is special. In addition to three standard polyhedra there are two rather big ones, dual to each other. One of them have dodecahedra for its three-dimensional faces, the other icosahedra for the *vertex figures*, convex hulls of points on all edges attached to vertex at a fixed small distance form the vertex. It is reasonable then to think of those bodies as "big cousins" of the dodecahedron and the icosahedron.

There is also a sixth, the *octacube*, not as big as the last two but bigger that the cube abd the octahedron. It has 24 vertices: 16 of these are of the form $(\pm 1, \pm 1, \pm 1, \pm 1)$, and are the vertices of a four-dimensional cube; the other 8 lie on the 4 coordinate axes at a distance of 2 from the origin, and are the vertices of the four-dimensional analogue of the octahedron. An imaginative three-dimensional representation of this highly symmetric object occupies pride of place on the main lobby of McAllister building.

This list is complete turns out to be complete: for $n \geq 4$, there are no regular convex polytopes besides S_n, I_n, and O_n. The proofs (both for dimension 4 and for higher dimension) although not exceeding difficult, lie beyond the scope of theses lectures.

This outcome is representative of a number of classification results that one finds in algebra: one wishes to classify every occurrence of a particular structure (in this case regular convex polytopes), and then finds that there are certain series or families which include infinitely many "regular" examples of that structure (in this case S_n, I_n, and O_n), and then a finite list of exceptional cases which include all other examples (in this case the dodecahedron, the icosahedron, their four-dimensional counterparts, and the octacube). A similar phenomenon occurs in the classification of complex simple Lie groups (four series of classical groups and five exceptional groups), and in the more formidable way, for finite simple groups where the exceptions include the famous (or infamous) "monster".

d. Back to scalar products. In the previous lecture, we introduced the general notion of a scalar product, and showed that in the standard coordinates on $\mathbb{R}^n$, every scalar product can be written in the form (9.8). We now show that by taking an appropriate choice of basis, every scalar product can actually be written in the canonical form (9.6). The matrix A from (9.8) turns out to be the change of basis matrix, as we will see.

Let $\langle \cdot, \cdot \rangle$ be an arbitrary scalar product on $\mathbb{R}^n$, which may or may not be the standard one. A basis is called *orthonormal* with respect to a scalar product if the scalar product of any two different elements is zero and the scalar product of any element with itself is one. Let $\langle\cdot,\cdot\rangle_0$ be the standard scalar product, and let $\mathcal{E} = \{\mathbf{e}^1,\dots,\mathbf{e}^n\}$ be the standard basis. $\mathcal{E}$ is orthonormal with respect to $\langle\cdot,\cdot\rangle_0$, but not with respect to $\langle\cdot,\cdot\rangle$. We construct an orthonormal basis for $\langle\cdot,\cdot\rangle$ using the following lemma.

Lemma 10.6. *Let $\mathbf{u}^1,\mathbf{u}^2,\dots,\mathbf{u}^k$ be linearly independent vectors in $\mathbb{R}^n$. Then the set*

$$L_k = \{\mathbf{v}\in\mathbb{R}^n \mid \langle \mathbf{u}^i,\mathbf{v}\rangle = 0 \textit{ for all } 1\le i\le k\} \tag{10.1}$$

is a $(n-k)$-dimensional subspace of $\mathbb{R}^n$.

Proof. For each $1\le i\le k$, the set $K_i = \{\mathbf{v}\in\mathbb{R}^n \mid \langle\mathbf{u}^i,\mathbf{v}\rangle = 0\}$ is the set of all roots of the linear equation

$$\sum_{j,k} a_{jk}u_j^i v_k = 0, \tag{10.2}$$

where u_j^i and v_j are the coordinates of $\mathbf{u}^i$ and $\mathbf{v}$ in the basis $\mathcal{E}$. It follows that K_i is a linear subspace of codimension one in $\mathbb{R}^n$, and the set L_k in (10.1) is the intersection $\bigcap_{i=1}^k K_i$. Observe that for every $1\le i<k$, we have

$$L_{i+1} = K_1\cap\cdots\cap K_{i+1} \subseteqq K_1\cap\cdots\cap K_i = L_i,$$

with equality if and only if $\mathbf{u}^{i+1}$ lies in the span of $\{\mathbf{u}^1,\dots,\mathbf{u}^k\}$. By linear independence, this never happens, and so the intersection with each new subspace decreases the dimension by one. The result follows. □

Let $\mathbf{u}^1\in\mathbb{R}^n$ be any unit vector—that is, $\langle\mathbf{u}^1,\mathbf{u}^1\rangle = 1$. Let L_1 be as in Lemma 10.6, so L_1 is a linear subspace of dimension $n-1$ which comprises all vectors orthogonal to $\mathbf{u}^1$. Choose an arbitrary unit vector $\mathbf{u}^2\in L_1\subset\mathbb{R}^n$, and again let L_2 be as in the lemma. Continuing in this manner, we obtain a basis $\mathcal{U} = \{\mathbf{u}^1,\dots,\mathbf{u}^n\}$ for which

$$\langle\mathbf{u}^i,\mathbf{u}^j\rangle = \begin{cases} 0 & i\ne j,\\ 1 & i=j.\end{cases}$$

$\mathcal{U}$ is the orthonormal basis we promised.

Now given any vector $\mathbf{x}\in\mathbb{R}^n$, we can write $x'_j = \langle\mathbf{x},\mathbf{u}^j\rangle$ and obtain

$$\mathbf{x} = \sum_{j=1}^n x'_j\mathbf{u}^j. \tag{10.3}$$

Using bilinearity of the scalar product, we see that

$$\begin{aligned}\langle \mathbf{x}, \mathbf{y} \rangle &= \left\langle \sum_{i=1}^{n} x_i' \mathbf{u}^i, \sum_{j=1}^{n} y_j' \mathbf{u}^j \right\rangle \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} x_i' y_j' \left\langle \mathbf{u}^i, \mathbf{u}^j \right\rangle \qquad (10.4) \\ &= \sum_{i=1}^{n} x_i' y_i',\end{aligned}$$

and hence $\langle \cdot, \cdot \rangle$ has the standard form (9.6) in the basis $\mathcal{U}$.

It follows that the particular form of any given scalar product (9.8) is merely a matter of what basis we choose; there is no intrinsic difference between different scalar products.

To see the relationship between (9.8) and (10.4), let $c_{ij} = \left\langle \mathbf{e}^i, \mathbf{u}^j \right\rangle$ for $1 \leq i, j \leq n$, so that $\mathbf{e}^i = \sum_{j=1}^{n} c_{ij} \mathbf{u}^j$, and hence

$$\mathbf{x} = \sum_{i=1}^{n} x_i \mathbf{e}^i = \sum_{i=1}^{n} \sum_{j=1}^{n} x_i c_{ij} \mathbf{u}^j. \qquad (10.5)$$

Comparing this with (10.3), we get the change of coordinates formula

$$x_j' = \sum_{i=1}^{n} c_{ij} x_i,$$

and it follows that in terms of the coordinates with respect to $\mathcal{E}$, the inner product $\langle \cdot, \cdot \rangle$ may be written

$$\begin{aligned}\langle \mathbf{x}, \mathbf{y} \rangle &= \sum_{j=1}^{n} \left(\sum_{i=1}^{n} c_{ij} x_i \right) \left(\sum_{k=1}^{n} c_{kj} y_k \right) \\ &= \sum_{i=1}^{n} \sum_{k=1}^{n} \left(\sum_{j=1}^{n} c_{ij} c_{kj} \right) x_i y_k.\end{aligned}$$

Writing C for the $n \times n$ matrix $[c_{ij}]$ and A for the matrix $[a_{ij}]$ in (9.8), we see that $A = CC^T$, where C^T is the transpose of C.

CHAPTER 3

Groups of matrices: Linear algebra and symmetry in various geometries

Lecture 11. Friday, September 25

a. Orthogonal matrices. Returning to the standard scalar product, let I be an isometry of $\mathbb{R}^n$ which fixes $\mathbf{0}$; thus I is a linear map which preserves the standard scalar product. In particular, the set $I\mathcal{E} = \{I\mathbf{e}^1, \dots, I\mathbf{e}^n\}$ is still an orthonormal basis, since $\left\langle I\mathbf{e}^i, I\mathbf{e}^j \right\rangle = \left\langle \mathbf{e}^i, \mathbf{e}^j \right\rangle$.

Let B be the $n \times n$ matrix which represents the linear transformation I in the basis $\mathcal{E}$—that is, $b_{ij} = \left\langle I\mathbf{e}^i, \mathbf{e}^j \right\rangle$, so

$$I\mathbf{e}^i = \sum_{j=1}^{n} b_{ij}\mathbf{e}^j.$$

Then the statement that $I\mathcal{E}$ is an orthonormal basis is equivalent to the statement that the row vectors of B are orthonormal, because in this case

$$\left\langle I\mathbf{e}^i, I\mathbf{e}^j \right\rangle = \sum_{k=1}^{n} b_{ik}b_{jk} = 0$$

for $i \neq j$, and

$$\left\langle I\mathbf{e}^i, I\mathbf{e}^i \right\rangle = \sum_{j=1}^{n} b_{ij}^2 = 1.$$

Recall from the rules for matrix multiplication that this is equivalent to the condition $BB^T = \mathrm{Id}$, or $B^T = B^{-1}$. This is in turn equivalent to $B^TB = \mathrm{Id}$, which is the statement that the column vectors of B are orthonormal.

Alternately, one may observe that if we let $\mathbf{x}$ denote a column vector and $\mathbf{x}^T$ a row vector, then the standard form of the scalar product (9.6) becomes $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T\mathbf{y}$, and so we have the following general relationship:

$$\langle \mathbf{x}, A\mathbf{y} \rangle = \mathbf{x}^T A\mathbf{y} = (A^T\mathbf{x})^T\mathbf{y} = \left\langle A^T\mathbf{x}, \mathbf{y} \right\rangle. \tag{11.1}$$

Thus if B is the matrix of I, which preserves scalar products, we have

$$\langle \mathbf{x}, \mathbf{y} \rangle = \langle B\mathbf{x}, B\mathbf{y} \rangle = \left\langle B^TB\mathbf{x}, \mathbf{y} \right\rangle$$

for every $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, which implies $B^TB = \mathrm{Id}$, as above.

Definition 11.1. A matrix B such that $B^T = B^{-1}$ is called *orthogonal*. The group of orthogonal $n \times n$ matrices is denoted $O(n)$.

It should come as no surprise by now that the group of orthogonal matrices is identified with the group of isometries which fix the origin. Furthermore, since $\det B^T = \det B$, we see that any orthogonal matrix has

$$1 = \det \mathrm{Id} = \det(B^T B) = (\det B^T)(\det B) = (\det B)^2,$$

and hence $\det B = \pm 1$. Matrices with determinant 1 correspond to even isometries fixing the origin and compose the special orthogonal group $SO(n)$; matrices with determinant -1 correspond to odd isometries fixing the origin.

In $SO(3)$, we saw that the conjugacy class of a rotation contained all rotations through the same angle θ. In the next lecture, we will sketch a proof of the analogous result in higher dimensions, which states that given any $B \in O(n)$, there exists $A \in O(n)$ such that the matrix $A^{-1}BA$ has the form

$$\begin{pmatrix} 1 &&&&&&&& \\ & \ddots &&&&&&& \\ && 1 &&&&&& \\ &&& -1 &&&&& \\ &&&& \ddots &&&& \\ &&&&& -1 &&& \\ &&&&&& R_{\theta_1} && \\ &&&&&&& \ddots & \\ &&&&&&&& R_{\theta_k} \end{pmatrix}, \tag{11.2}$$

where all entries not shown are taken to be 0, and where R_θ is the 2×2 rotation matrix

$$R_\theta = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix},$$

and the θ_i may be the same or may be different but not equal to 0 or π.

One can also combines pairs of 1's into rotations by angle 0 and pairs of -1's into rotations by π. Then no more than one "loose" diagonal element 1 and -1 is left. In the even dimension n for an $SO(n)$ matrix no loose elements remain and for a matrix with determinant -1 there is one 1 and one -1. in the odd dimension exactly one loose element remains and it is 1 or -1 according tot he sign of the determinant.

Geometrically, this means that $\mathbb{R}^n$ can be decomposed into the orthogonal direct sum of a number of one-dimensional subspaces X_i which are fixed by B, a number of one-dimensional subspaces Y_i on which B acts as the map $x \mapsto -x$ (that is, a reflection), and a number of two-dimensional subspaces Z_i on which B acts as a rotation by θ_i. Since the rotation is a product of two reflections this also gives a representation of isometry as the product of at most n reflections.

The isometry determined by the matrix B can be written as the product of commuting reflections in the orthogonal complements of Y_i (reflection is always around an affine subspace of codimension one) together with commuting rotations in the orthogonal complements of Z_i (rotation is always around an affine subspace of codimension two). The number of subspaces Y_i—that is, the number of times -1 occurs on the diagonal—determines whether the isometry given by B is even or odd.

The three dimensional case is particularly easy then: there is one rotation block (possibly the identity) and either 1 on the diagonal (resulting in a rotation or the identity map) or -1 (resulting in a rotatory reflection or a pure reflection).

b. Eigenvalues, eigenvectors, and diagonalizable matrices. We stated in the previous lecture that every orthogonal matrix $A \in O(n)$ can be put in the form (11.2) by a suitable change of coordinates—that is, a transformation of the form $A \mapsto CAC^{-1}$, where $C \in O(n)$ is the change of basis matrix. This is related to perhaps the most important result in linear algebra, *Jordan normal form*. In this lecture, we will review the relevant concepts from linear algebra and show why every orthogonal transformation can be so represented. Along the way we will learn importance of *complexification*, when objects defined over the field of real numbers (in our case, linear spaces, linear transformations and scalar products) are extended to the complex field.

Before diving into the details, we observe that our mission can be described both geometrically and algebraically. Geometrically, the story is this: we are given a linear transformation $L\colon \mathbb{R}^n \to \mathbb{R}^n$, and we wish to find a basis in which the matrix of L takes on as simple a form as possible. In algebraic terms, we are given a matrix $L \in GL(n, \mathbb{R})$, and we wish to describe the conjugacy class of L—that is, we want to characterise all matrices L' such that $L' = CLC^{-1}$ for some $C \in GL(n, \mathbb{R})$.[1] Ideally, we would like to select a good representative from each conjugacy class, which will be the *normal form* of L.

Definition 11.2. Let L be an $n \times n$ matrix with real entries. An *eigenvalue* of L is a number λ such that

$$L\mathbf{v} = \lambda\mathbf{v} \tag{11.3}$$

for some vector $\mathbf{v} \in \mathbb{R}^n$, called an *eigenvector* of L. The set of all eigenvectors of λ is a subspace of $\mathbb{R}^n$, called the *eigenspace* of λ. The *multiplicity* of λ is the dimension of this subspace.

Although this definition only allows real eigenvalues, we will soon see that complex eigenvalues can also exist, and are quite important.

[1] If $L \in O(n)$, then we would like to take the conjugating matrix C to be orthogonal as well. In this case there is no difference between conjugacy in the group $GL(n, \mathbb{R})$ and conjugacy in the subgroup $O(n)$, but this is not always the case; recall that rotations $R_\theta^{\mathbf{x}}$ and $R_{-\theta}^{\mathbf{x}}$ are conjugate in $\mathrm{Isom}(\mathbb{R}^2)$, but not in $\mathrm{Isom}^+(\mathbb{R}^2)$.

Exercise 11.1. Let $\mathbf{v}_1, \dots, \mathbf{v}_k$ be eigenvectors of L, and let $\lambda_1, \dots, \lambda_k$ be the corresponding eigenvalues. Suppose that $\lambda_i \neq \lambda_j$ for $i \neq j$, and show that the eigenvectors $\mathbf{v}_i$ are linearly independent.

It follows from Exercise 11.1 that there are only finitely many eigenvalues for any matrix. But why should we be interested in eigenvalues and eigenvectors? What purpose does (11.3) serve?

One important (algebraic) reason is that the set of eigenvalues of a matrix is invariant under conjugacy.

An important geometric reason is that (11.3) shows that on the subspace containing $\mathbf{v}$, the action of the linear map $L \colon \mathbb{R}^n \to \mathbb{R}^n$ is particularly simple—multiplication by λ! If we can decompose $\mathbb{R}^n$ into a direct product of such subspaces, then we can legitimately claim to have understood the action of L.

Definition 11.3. L is *diagonalizable* (over $\mathbb{R}$) if there exists a basis $\mathbf{v}_1, \dots, \mathbf{v}_n \in \mathbb{R}^n$ such that each $\mathbf{v}_i$ is an eigenvector of L.

Suppose $\{\mathbf{v}_i\}$ is a basis of eigenvectors with eigenvalues $\{\lambda_i\}$, and let $C \in GL(n, \mathbb{R})$ be the linear map such that $C\mathbf{v}_i = \mathbf{e}_i$ for each $1 \leq i \leq n$. Observe that

$$CLC^{-1}\mathbf{e}_i = CL\mathbf{v}_i = C(\lambda_i \mathbf{v}_i) = \lambda_i \mathbf{e}_i;$$

hence the matrix of CLC^{-1} is

$$\operatorname{diag}(\lambda_1, \dots, \lambda_n) = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix}. \tag{11.4}$$

It follows from Exercise 11.1 that L has no more than n eigenvalues. So far, though, nothing we have said prevents it from having fewer than n eigenvalues, even if we count each eigenvalue according to its multiplicity. Indeed, one immediately sees that any rotation of the plane by an angle not equal to 0 or π is a linear map with no real eigenvalues. Thus we cannot expect to diagonalise every matrix, and must look to more general forms for our classification.

The eigenvalue equation (11.3) characterises eigenvectors (and hence eigenvalues) geometrically: $\mathbf{v}$ is an eigenvector if and only if it is parallel to its image $L\mathbf{v}$. An algebraic description of eigenvalues can be obtained by recalling that given an $n \times n$ matrix A, the existence of a vector $\mathbf{v}$ such that $A\mathbf{v} = \mathbf{0}$ is equivalent to the condition that $\det A = 0$. We can rewrite (11.3) as $(L - \lambda \operatorname{Id})\mathbf{v} = \mathbf{0}$, and so we see that λ is an eigenvalue of L if and only if $\det(L - \lambda \operatorname{Id}) = 0$.

The determinant of an $n \times n$ matrix is the sum of $n!$ terms, each of which is a product of n entries of the matrix, one from each row and column. It follows that $p(\lambda)$ is a polynomial of degree n, called the *characteristic* *polynomial* of the matrix L, and that the coefficients of p are polynomial expressions in the entries of the matrix.

The upshot of all this is that the eigenvalues of a matrix are the roots of its characteristic polynomial, and now we see the price we pay for working with the real numbers—$\mathbb{R}$ is not algebraically closed, and hence the characteristic polynomial may not factor completely over $\mathbb{R}$! Indeed, it may not have any roots at all; for example the characteristic polynomial of the rotation matrix $\left(\begin{smallmatrix}\cos\theta & -\sin\theta \\ \sin\theta & \cos\theta\end{smallmatrix}\right)$ is $p(\lambda) = \lambda^2 - 2\cos\theta + 1$.

We can resolve this difficulty and ensure that L has "enough eigenvalues" by passing to the complex numbers, over which every polynomial factors completely, and declaring any *complex* root of $p(\lambda) = 0$ to be an eigenvalue of L. Then the Fundamental Theorem of Algebra gives us

$$p(\lambda) = \det(L - \lambda \operatorname{Id}) = \prod_{i=1}^{n} (\lambda - \lambda_i), \tag{11.5}$$

where $\{\lambda_1, \dots, \lambda_n\} \subset \mathbb{C}$ are the eigenvalues of L.

The set of all eigenvalues of L is called the *spectrum* of L.

EXERCISE 11.2. Given an $n \times n$ matrix L and a change of coordinates $C \in GL(n, \mathbb{R})$, show that L and $L' = CLC^{-1}$ have the same spectrum, and that C takes eigenvectors of L into eigenvectors of L'.

At this point, it is not at all clear what geometric significance a complex eigenvalue has, if any. After all, if $\lambda \in \mathbb{C} \setminus \mathbb{R}$ is an eigenvalue of L and $\mathbf{v}$ is a vector in $\mathbb{R}^n$, what does the expression $\lambda\mathbf{v}$ even *mean*?

c. Complexification, complex eigenvectors and rotations. The difficulty in interpreting the expression $\lambda\mathbf{v}$ for $\lambda \in \mathbb{C}$ and $\mathbf{v} \in \mathbb{R}^n$ is that vectors in $\mathbb{R}^n$ must have real coordinates. We can solve this problem in a rather simple-minded way—just let the coordinates be complex! If we consider vectors $\mathbf{v} \in \mathbb{C}^n$, the n-dimensional *complex* vector space, then $\lambda\mathbf{v}$ makes perfect sense for any $\lambda \in \mathbb{C}$; thus (11.3) may still be used as the definition of an eigenvalue and eigenvector, and agrees with the definition in terms of the characteristic polynomial.

The same procedure can be put more formally: $\mathbb{C}^n$ is the *complexification* of the real vector space $\mathbb{R}^n$, and is equal as a real vector space to the direct sum of two copies of $\mathbb{R}^n$. We call these two copies V_R and V_I (for real and imaginary); given vectors $\mathbf{x} \in V_R$ and $\mathbf{y} \in V_I$, we intertwine the coordinates and write

$$\mathbf{z} = (\mathbf{x}_1, \mathbf{y}_1, \mathbf{x}_2, \mathbf{y}_2, \dots, \mathbf{x}_n, \mathbf{y}_n) \in \mathbb{R}^{2n} \tag{11.6}$$

for the vector with real part $\mathbf{x}$ and imaginary part $\mathbf{y}$. As a vector with n complex coordinates, we write $\mathbf{z}$ as

$$\mathbf{z} = (\mathbf{x}_1 + i\mathbf{y}_1, \mathbf{x}_2 + i\mathbf{y}_2, \dots, \mathbf{x}_n + i\mathbf{y}_n). \tag{11.7}$$

In order to go from the formulation (11.6) to the complex vector space (11.7), we must observe that multiplication by i acts on $\mathbb{R}^{2n}$ as the linear operator

$$J\colon (\mathbf{x}_1, \mathbf{y}_1, \dots, \mathbf{x}_n, \mathbf{y}_n) \mapsto (-\mathbf{y}_1, \mathbf{x}_1, \dots, -\mathbf{y}_n, \mathbf{x}_n).$$

That is, if we decompose $\mathbb{R}^{2n}$ as the direct sum of n copies of $\mathbb{R}^2$, the action of J rotates each copy of $\mathbb{R}^2$ by $\pi/2$ counterclockwise, which is exactly the effect multiplication by i has on the complex plane.[2]

Having defined $\mathbb{C}^n$, we observe that since L and J commute, L extends uniquely to a linear operator $L^{\mathbb{C}}\colon \mathbb{C}^n \to \mathbb{C}^n$. All the definitions from the previous section go through for $L^{\mathbb{C}}$, and now the fundamental theorem of algebra guarantees that (11.5) holds and the characteristic polynomial factors completely over $\mathbb{C}$. We refer to any eigenvalue of $L^{\mathbb{C}}$ as an eigenvalue of L itself, and this justifies our definition of spectrum of L as a subset of $\mathbb{C}$. But now we must ask: What do the (complex-valued) eigenvalues and eigenvectors of $L^{\mathbb{C}}$ have to do with the geometric action of L on $\mathbb{R}^n$?

To answer this, we consider an eigenvalue $\lambda \in \mathbb{C} \setminus \mathbb{R}$ and the corresponding eigenvector $\mathbf{z} \in \mathbb{C}^n$. Obviously since $\lambda \notin \mathbb{R}$ we have $\mathbf{z} \notin \mathbb{R}^n$; how do we extract a real-valued vector from $\mathbf{z}$ on which the action of L is related to λ?

Observe that since the entries of the matrix for L are real-valued, the coefficients of the characteristic polynomial $p(\lambda)$ are real-valued. It follows that (11.5) is invariant under the involution $\lambda \mapsto \bar{\lambda}$, and hence if $\lambda \in \mathbb{C} \setminus \mathbb{R}$ is an eigenvalue of $L^{\mathbb{C}}$, so is $\bar{\lambda}$. Furthermore, one may easily verify that $L^{\mathbb{C}}\bar{\mathbf{z}} = \bar{\lambda}\bar{\mathbf{z}}$, where $\bar{\mathbf{z}}$ is defined in the obvious way as

$$\bar{\mathbf{z}} = (\bar{z}_1, \bar{z}_2, \dots, \bar{z}_n) = \mathbf{x} - i\mathbf{y},$$

where $\mathbf{z} = \mathbf{x} + i\mathbf{y}$ for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. Observe that $\mathbf{x} = (\mathbf{z} + \bar{\mathbf{z}})/2$ and $\mathbf{y} = i(\mathbf{z} - \bar{\mathbf{z}})/2$; thus the two-dimensional *complex* subspace of $\mathbb{C}^n$ spanned by $\mathbf{z}$ and $\bar{\mathbf{z}}$ intersects $V_R = \mathbb{R}^n$ in the two-dimensional *real* subspace spanned by $\mathbf{x}$ and $\mathbf{y}$.

To see how L acts on this subspace, write $\lambda = \rho e^{i\theta}$, where $\rho > 0$ and $\theta \in [0, 2\pi)$. Then we have

$$\begin{aligned} L\mathbf{x} + iL\mathbf{y} = L^{\mathbb{C}}\mathbf{z} &= \lambda\mathbf{z} \\ &= \rho(\cos\theta + i\sin\theta)(\mathbf{x} + i\mathbf{y}) \\ &= \rho(\cos\theta\mathbf{x} - \sin\theta\mathbf{y}) + i\rho(\cos\theta\mathbf{y} + \sin\theta\mathbf{x}), \end{aligned}$$

and so L acts on the two-dimensional subspace spanned by $\mathbf{x}$ and $\mathbf{y}$ as a spiral motion—rotation by θ scaled by ρ, with matrix

$$\rho R_\theta = \begin{pmatrix} \rho\cos\theta & -\rho\sin\theta \\ \rho\sin\theta & \rho\cos\theta \end{pmatrix}.$$

[2] It turns out that there are other settings, beyond that of linear spaces, in which one can go from a real structure to a complex structure with the help of a linear operator J with the property that $J^2 = -\operatorname{Id}$. The most accessible and of the most important instance, is the theory of *Riemann surfaces*.

Now suppose $L^{\mathbb{C}}$ is diagonalisable over $\mathbb{C}$—that is, there exists $C \in GL(n, \mathbb{C})$ such that

$$CL^{\mathbb{C}}C^{-1} = \operatorname{diag}\left(\lambda_1, \dots, \lambda_j, \rho_1 e^{i\theta_1}, \rho_1 e^{-i\theta_1}, \dots, \rho_k e^{i\theta_k}, \rho_k e^{-i\theta_k}\right),$$

where $\lambda_i \in \mathbb{R}$, $\rho_i > 0$, $\theta_i \in (0, \pi)$, and $j + 2k = n$. Then using the above procedure, one obtains a basis for $\mathbb{R}^n$ in which the matrix of L is

$$\operatorname{diag}\left(\lambda_1, \dots, \lambda_j, \rho_1 R_{\theta_1}, \dots, \rho_k R_{\theta_k}\right). \tag{11.8}$$

Thus while L cannot be diagonalised over $\mathbb{R}$, it can at least be put into block diagonal form, provided $L^{\mathbb{C}}$ can be diagonalised over $\mathbb{C}$. But is even this much always possible?

d. Differing multiplicities and Jordan blocks. Observe that since the determinant of any upper-triangular matrix is the product of the diagonal entries, the characteristic polynomial of an upper-triangular matrix L is

$$\det(L - \lambda \operatorname{Id}) = \prod_{i=1}^{n} (L_{ii} - \lambda).$$

Thus the eigenvalues of L are simply the diagonal entries.

Example 11.4. Consider the matrix $L = \left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$. Its only eigenvalue is 1, and it has $(1, 0)$ as an eigenvector. In fact, this is the *only* eigenvector (up to scalar multiples); this fact can be shown directly, or one can observe that if L were diagonalisable, then we would have $CLC^{-1} = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$ for some $C \in GL(n, \mathbb{R})$, which would then imply $L = \operatorname{Id}$, a contradiction.

This example shows that not every matrix is diagonalisable over $\mathbb{C}$, and hence not every matrix can be put in block diagonal form over $\mathbb{R}$. In general, this occurs whenever L has an eigenvalue λ for which the geometric multiplicity (the number of linearly independent eigenvectors) is strictly less than the algebraic multiplicity (the number of times λ appears as a root of the characteristic polynomial). In this case the eigenspace corresponding to λ is not as big as it "should" be. A notion of *generalised eigenspace* can be introduced, and it can be shown that every matrix can be put in *Jordan normal form.*

We shall not go through the details of this here; rather, we observe that the non-existence of a basis of eigenvectors is a result of the fact that as we select eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \dots,$ we reach a point where there is no L-invariant subspace transverse to the subspace spanned by $\mathbf{v}_1, \dots, \mathbf{v}_k$, and thus no further eigenvectors can be found. For *orthogonal* matrices, we avoid this problem, as follows.

Let $V \subset \mathbb{R}^n$ be an invariant subspace for L—that is, $L(V) = V$—and let $V^{\perp}$ be the orthogonal complement of $\mathbb{R}^n$,

$$V^{\perp} = \{\mathbf{v} \in \mathbb{R}^n \mid \langle \mathbf{v}, \mathbf{w} \rangle = 0 \text{ for all } \mathbf{w} \in V\}.$$

Given $\mathbf{v} \in V^{\perp}$, we have $\langle L\mathbf{v}, L\mathbf{w} \rangle = \langle \mathbf{v}, \mathbf{w} \rangle$ for all $\mathbf{w} \in V$, and hence $L\mathbf{v} \in V^{\perp}$. It follows that $V^{\perp}$ is invariant, and so there exists an eigenvector of L in $V^{\perp}$ (or perhaps a two-dimensional space on which L acts as ρR_{θ}). Continuing in this way, we can diagonalise $L^{\mathbb{C}}$, and hence put L in the form (11.8).

Finally, we observe that any eigenvalue of an orthogonal matrix must have absolute value one. This follows since the determinant of L restricted to any invariant subspace is equal to 1. It follows that (11.8) reduces to the form given at the end of the previous lecture, and we thus completed the proof of reducibily of any orthogonal matrix to the form (11.2) by an orthogonal transformation.

Lecture 12. Monday, September 28

a. Hermitian product and unitary matrices. One can extend the scalar product on $\mathbb{R}^n$ to a *Hermitian product* on $\mathbb{C}^n$ by

$$\langle \mathbf{z}, \mathbf{w} \rangle = \sum_{j=1}^{n} z_j \overline{w_j}. \tag{12.1}$$

The Hermitian product satisfies similar properties to the scalar product:

(1) $\langle \mathbf{w}, \mathbf{w} \rangle \geq 0$, with equality if and only if $\mathbf{w} = \mathbf{0}$.
(2) $\langle \mathbf{v}, \mathbf{w} \rangle = \overline{\langle \mathbf{w}, \mathbf{v} \rangle}$.
(3) *Linearity*: $\langle \mathbf{u} + \mathbf{v}, \mathbf{w} \rangle = \langle \mathbf{u}, \mathbf{w} \rangle + \langle \mathbf{v}, \mathbf{w} \rangle$.
(4) $\langle \lambda \mathbf{v}, \mathbf{w} \rangle = \lambda \langle \mathbf{v}, \mathbf{w} \rangle$, $\langle \mathbf{v}, \lambda \mathbf{w} \rangle = \bar{\lambda} \langle \mathbf{v}, \mathbf{w} \rangle$.

This devise will allow to find a natural extension of the theory of orthogonal matrices to the complex domain.

It may not be immediately apparent why we should use (12.1) instead of the more natural-looking extension $\sum_{j=1}^{n} z_j w_j$. One could define a scalar product on $\mathbb{C}^n$ using the latter formula; however, one would obtain a totally different sort of beast than the one we now consider. In particular, the Hermitian product defined in (12.1) has the following property: If $\mathbf{z} = \mathbf{x} + i\mathbf{y}$ and $\mathbf{w} = \mathbf{u} + i\mathbf{v}$ for real vectors $\mathbf{x}, \mathbf{y}, \mathbf{u}, \mathbf{v}$, then

$$\begin{aligned} \langle \mathbf{z}, \mathbf{w} \rangle &= \sum_{j=1}^{n} (x_j + i y_j)(u_j - i v_j) \\ &= \sum_{j=1}^{n} (x_j u_j + y_j v_j) + i(y_j u_j - x_j v_j). \end{aligned} \tag{12.2}$$

In particular, the real part of $\langle \mathbf{z}, \mathbf{w} \rangle$ is the *real* scalar product of the vectors $(x_1, y_1, \dots x_n, y_n)$ and $(u_1, v_1, \dots, u_n, v_n)$ in $\mathbb{R}^{2n}$. Thus the Hermitian product is a natural generalisation of the real scalar product, and we see that the complex conjugate $\overline{w_j}$ must be used in order to avoid a negative sign in front of the term $y_j v_j$ in (12.2).

Furthermore, the presence of the complex conjugate in (12.1) is crucial in order to guarantee that

$$\langle \mathbf{z}, \mathbf{z} \rangle = \sum_{j=1}^{n} z_j \overline{z_j} = \sum_{j=1}^{n} |z_j|^2$$

is a non-negative real number, which vanishes if and only if $\mathbf{z} = \mathbf{0}$. In particular, the Hermitian product defines a *norm* on $\mathbb{C}^n$ by $\|\mathbf{z}\|^2 = \langle \mathbf{z}, \mathbf{z} \rangle$, with the following properties.

(1) $\|\mathbf{z}\| \geq 0$, with equality if and only if $\mathbf{z} = \mathbf{0}$.
(2) $\|\lambda \mathbf{z}\| = |\lambda| \, \|\mathbf{z}\|$ for all $\lambda \in \mathbb{C}$.
(3) $\|\mathbf{z} + \mathbf{w}\| \leq \|\mathbf{z}\| + \|\mathbf{w}\|$ for all $\mathbf{z}, \mathbf{w} \in \mathbb{C}^n$.

The norm provides a notion of length, and the Hermitian product provides a notion of orthogonality: as in the real case, two vectors $\mathbf{w}, \mathbf{z} \in \mathbb{C}^n$ are orthogonal if $\langle \mathbf{w}, \mathbf{z} \rangle = 0$. Thus we once again have a notion of an *orthonormal basis*—that is, a basis $\{\mathbf{z}^1, \dots, \mathbf{z}^n\}$ of $\mathbb{C}^n$ such that

$$\left\langle \mathbf{z}^j, \mathbf{z}^k \right\rangle = \delta_{jk},$$

where δ_{jk} is the Dirac delta function, which takes the value 1 if $j = k$ and 0 otherwise.

As in $\mathbb{R}^n$, we have a standard orthonormal basis $\mathcal{E} = \{\mathbf{e}^1, \dots, \mathbf{e}^n\}$:

$$\mathbf{e}^j = (0, \dots, 0, 1, 0, \dots, 0),$$

where the 1 appears in the jth position. An orthonormal basis corresponds to a decomposition of the vector space into one-dimensional subspaces which are pairwise orthogonal. In both $\mathbb{R}^n$ and $\mathbb{C}^n$, we can generate other orthonormal bases from $\mathcal{E}$ without changing the subspaces in the decomposition: simply replace $\mathbf{e}^j$ with a parallel unit vector. In $\mathbb{R}^n$, the only parallel unit vector to $\mathbf{e}^j$ is $-\mathbf{e}^j$; in $\mathbb{C}^n$, we can replace $\mathbf{e}^j$ with $\lambda\mathbf{e}^j$, where $\lambda \in S^1$ is any complex number with $|\lambda| = 1$.

This distinction is related to a fundamental difference between $\mathbb{R}^n$ and $\mathbb{C}^n$. In the former case, replacing $\mathbf{e}^j$ with $-\mathbf{e}^j$ changes the orientation of the basis, and hence we can distinguish between even and odd orientations. In $\mathbb{C}^n$, this replacement can be done continuously by moving $\mathbf{e}^j$ to $e^{i\theta}\mathbf{e}^j$ for $0 \leq \theta \leq \pi$; consequently, there is no meaningful way to say where the "orientation" reverses. In fact, in $\mathbb{C}^n$ we must abandon the notion of orientation entirely, and can no longer speak of even and odd maps.

Definition 12.1. A linear map $A\colon \mathbb{C}^n \to \mathbb{C}^n$ is *unitary* if $\langle A\mathbf{z}, A\mathbf{w} \rangle = \langle \mathbf{z}, \mathbf{w} \rangle$ for all $\mathbf{z}, \mathbf{w} \in \mathbb{C}^n$. The group of unitary $n \times n$ complex matrices is denoted $U(n)$.

Observe that since the real part of the Hermitian product is just the usual real scalar product on $\mathbb{R}^{2n}$, every unitary map on $\mathbb{C}^n$ corresponds to an orthogonal map on $\mathbb{R}^{2n}$. The converse is not true; there are orthogonal maps on $\mathbb{R}^{2n}$ which are not unitary maps on $\mathbb{C}^n$. Indeed, such a map may not even be linear on $\mathbb{C}^n$; it must behave properly with respect to multiplication by i.

However, unitary maps *are* a generalisation of orthogonal maps in the following sense: given an orthogonal linear map $L\colon \mathbb{R}^n \to \mathbb{R}^n$, the complexification $L^{\mathbb{C}}\colon \mathbb{C}^n \to \mathbb{C}^n$ is unitary.

Proposition 12.2. *If $A\colon \mathbb{C}^n \to \mathbb{C}^n$ is unitary and λ is an eigenvalue of A, then $|\lambda| = 1$.*

Proof. Let $\mathbf{z} \in \mathbb{C}^n$ be an eigenvector for λ, and observe that

$$\langle \mathbf{z}, \mathbf{z} \rangle = \langle A\mathbf{z}, A\mathbf{z} \rangle = \langle \lambda\mathbf{z}, \lambda\mathbf{z} \rangle = \lambda\overline{\lambda} \langle \mathbf{z}, \mathbf{z} \rangle,$$

and hence

$$\lambda\overline{\lambda} = |\lambda|^2 = 1. \qquad \square$$

Because $\mathbb{C}$ is algebraically closed, the general normal form for (complex) unitary matrices is simpler than the result in the previous lectures for (real) orthogonal matrices. The proof, however, is basically the same, and relies on the fact that preservation of the (real or complex) scalar product guarantees the existence of invariant transverse subspaces.

Lemma 12.3. *Every linear map $L\colon \mathbb{C}^k \to \mathbb{C}^k$ has an eigenvector.*

Proof. Because $\mathbb{C}$ is algebraically complete, the characteristic polynomial $p(\lambda) = \det(L - \lambda \operatorname{Id})$ has a root λ_0. Thus $\det(L - \lambda_0 \operatorname{Id}) = 0$, and it follows that there exists $\mathbf{w} \in \mathbb{C}^k$ such that $(L - \lambda_0 \operatorname{Id})\mathbf{w} = \mathbf{0}$. This $\mathbf{w}$ is an eigenvector of L. □

Recall that given a linear map $L\colon V \to V$, a subspace W is *invariant* if $L(W) \subset W$. If $W \subset \mathbb{C}^n$ is an invariant subspace of L, then we may apply Lemma 12.3 to $\mathbb{C}^k = W$ and obtain the existence of an eigenvector in W.

The relationship between eigenvectors and invariant subspaces may be made even more explicit by the observation that an eigenvector is precisely a vector which spans a one-dimensional invariant subspace.

Definition 12.4. Let V be a vector space and $W \subset V$ a subspace. A subspace $W' \subset V$ is *transversal* to W if $W \cap W = \{\mathbf{0}\}$ and if $V = W + W'$. Equivalently, W and W' are transversal if for any $\mathbf{v} \in V$, there exist *unique* vectors $\mathbf{w} \in W$ and $\mathbf{w}' \in W'$ such that $\mathbf{v} = \mathbf{w} + \mathbf{w}'$.

If $\langle \cdot, \cdot \rangle$ is a Hermitian product on $\mathbb{C}^n$ and $W \subset \mathbb{C}^n$ is a subspace, then the *orthogonal complement* of W is

$$W^\perp = \{\mathbf{z} \in \mathbb{C}^n \mid \langle \mathbf{z}, \mathbf{w} \rangle = 0 \text{ for all } \mathbf{w} \in W\}.$$

Proposition 12.5. *Let $A\colon \mathbb{C}^n \to \mathbb{C}^n$ be unitary and $W \subset \mathbb{C}^n$ be invariant. Then $W^\perp$ is invariant as well.*

Proof. Observe that since A is unitary, A^{-1} is as well. Thus given $\mathbf{z} \in W^\perp$ and $\mathbf{w} \in W$, we have

$$\langle A\mathbf{z}, \mathbf{w} \rangle = \left\langle A^{-1}A\mathbf{z}, A^{-1}\mathbf{w} \right\rangle = \left\langle \mathbf{z}, A^{-1}\mathbf{w} \right\rangle. \tag{12.3}$$

Furthermore, since A is invertible and W is finite-dimensional, we have $A^{-1}(W) = W$, and hence the quantity in (12.3) vanishes. Since $\mathbf{w} \in W$ was arbitrary, it follows that $A\mathbf{z} \in W^\perp$. □

Proposition 12.6. *Given a linear map $L\colon \mathbb{C}^n \to \mathbb{C}^n$, the following are equivalent:*

(1) L is unitary.
(2) If $\mathcal{U} = \{\mathbf{u}^1, \dots, \mathbf{u}^n\}$ is any *orthonormal basis for $\mathbb{C}^n$, then $L(\mathcal{U})$ is again an orthonormal basis.*
(3) There exists an orthonormal basis $\mathcal{U}$ such that $L(\mathcal{U})$ is again an orthonormal basis.

Proof. That (1) implies (2) is immediate from the definition of unitary, and (2) is *a priori* stronger than (3). Finally, if (3) holds, then for any $\mathbf{w}, \mathbf{z} \in \mathbb{C}^n$ we may decompose $\mathbf{w} = \sum_j w_j \mathbf{u}^j$ and $\mathbf{z} = \sum_k z_k \mathbf{u}^k$, obtaining

$$\begin{aligned}\langle L\mathbf{w}, L\mathbf{z}\rangle &= \sum_{j,k} w_j z_k \left\langle L\mathbf{u}^j, L\mathbf{u}^k \right\rangle \\ &= \sum_{j,k} w_j z_k \delta_{jk} = \sum_{j,k} w_j z_k \left\langle \mathbf{u}^j, \mathbf{u}^k \right\rangle = \langle \mathbf{w}, \mathbf{z}\rangle . \qquad \square\end{aligned}$$

Now we can state the fundamental theorem on classification of unitary matrices.

Theorem 12.7. *For every $A \in U(n)$ there exists $C \in U(n)$ such that*

$$CAC^{-1} = \operatorname{diag}(\lambda_1, \dots, \lambda_n), \tag{12.4}$$

where $|\lambda_j| = 1$ for $1 \leq j \leq n$.

Proof. We apply Lemma 12.3 and Proposition 12.5 repeatedly. First let $\mathbf{u}^1 \in \mathbb{C}^n$ be any unit eigenvector of A, and let W_1 be the subspace spanned by $\mathbf{u}^1$. Then $W_1^\perp$ is invariant, and so there exists a unit eigenvector $\mathbf{u}^2 \in W_1^\perp$. Let W_2 be the subspace spanned by $\mathbf{u}^1$ and $\mathbf{u}^2$, and continue in this manner.

Thus we obtain an orthonormal basis $\{\mathbf{u}^1, \dots, \mathbf{u}^n\}$ such that $A\mathbf{u}^j = \lambda_j \mathbf{u}^j$ for $1 \leq j \leq n$. By Proposition 12.2, we have $|\lambda_j| = 1$ for every j. Furthermore, if we let C be the $n \times n$ complex matrix such that $C\mathbf{u}^j = \mathbf{e}^j$, then it follows from Proposition 12.6 that C is unitary, and furthermore,

$$CAC^{-1}\mathbf{e}_j = CA\mathbf{u}^j = C(\lambda_j \mathbf{u}^j) = \lambda_j \mathbf{e}^j,$$

which is enough to establish (12.4). □

For real matrices, we considered the special orthogonal group $SO(n)$ within the orthogonal group $O(n)$. We can do the same here and consider the *special unitary group*

$$SU(n) = \{A \in U(n) \mid \det A = 1\}.$$

However, in the complex case, the relationship between $SU(n)$ and $U(n)$ is much closer to the relationship between $SL(n, \mathbb{R})$ and $GL(n, \mathbb{R})$ than it is to the relationship between $SO(n)$ and $O(n)$. In particular, observe that $SO(n)$ is a subgroup of index 2 in $O(n)$, while $SL(n, \mathbb{R})$ and $SU(n)$ both have infinite index in their respective groups.

b. Normal matrices. We are interested in the class of matrices which can be diagonalised over $\mathbb{C}$, because such matrices have a simpler geometric meaning than matrices with no such diagonalisation. We have seen that this class does not include all matrices, thanks to the existence of matrices like $\left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$. Conversely, we have seen that this class *does* include all unitary matrices.

Of course, there are plenty of matrices which can be diagonalised but are not unitary; in particular, we may consider diagonal matrices $\operatorname{diag}(\lambda_1, \dots, \lambda_n)$ for which the eigenvalues λ_j do not lie on the unit circle—that is, $|\lambda| \neq 1$. Can we give a reasonable characterisation of the class of matrices which can be diagonalised over $\mathbb{C}$?

Remark. In the present setting, this question may seem somewhat academic, since any matrix can be put in Jordan normal form, which already gives us a complete understanding of its action on $\mathbb{C}^n$ (or $\mathbb{R}^n$). However, it turns out to be vital to understanding what happens in the infinite-dimensional situation, where $\mathbb{C}^n$ is replaced with the more general concept of a *Hilbert space*, and eigenvalues and eigenvectors give way to *spectral theory.* In this general setting there is no analogue of Jordan normal form, and the class of maps we examine here turns out to be very important.

Recall that given a real $n \times n$ matrix A (which may or may not be orthogonal), the *transpose* of A defined by $(A^T)_{ij} = A_{ji}$ has the property that

$$\langle \mathbf{x}, A\mathbf{y} \rangle = \left\langle A^T \mathbf{x}, \mathbf{y} \right\rangle$$

for every $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. For complex vectors and the Hermitian product, the analogous matrix is called the *adjoint* of A; it is denoted A^* and has the property that

$$\langle \mathbf{z}, A\mathbf{w} \rangle = \langle A^* \mathbf{z}, \mathbf{w} \rangle$$

for every $\mathbf{z}, \mathbf{w} \in \mathbb{C}^n$.

Exercise 12.1. Show that the matrix of A^* is the *conjugate transpose* of the matrix of A—that is, that

$$(A^*)_{ij} = \overline{A_{ji}}.$$

Recall that a matrix A is unitary if and only if $\langle A\mathbf{z}, A\mathbf{w} \rangle = \langle \mathbf{z}, \mathbf{w} \rangle$ for all $\mathbf{z}, \mathbf{w} \in \mathbb{C}^n$. This is equivalent to the condition that $\langle A^* A\mathbf{z}, \mathbf{w} \rangle = \langle \mathbf{z}, \mathbf{w} \rangle$ for all $\mathbf{z}$ and $\mathbf{w}$, which is in turn equivalent to the condition that $A^*A = \mathrm{Id}$. In particular, this implies that $A^* = A^{-1}$, and hence A and A^* commute.

Definition 12.8. $A \in M(n, \mathbb{C})$ is *normal* if $A^*A = AA^*$.

Every unitary matrix is normal, but there are normal matrices which are not unitary. This follows immediately from the fact that normality places no restrictions on the eigenvalues of A; in particular, every scalar multiple of the identity matrix is normal, but $\lambda \,\mathrm{Id}$ is only unitary if $|\lambda| = 1$.

It turns out that normality is precisely the condition we need in order to make the argument from the previous section go through (modulo the statement about the absolute values of the eigenvalues). In particular, we can prove an analogue of Proposition 12.5, after first making some general observations.

First we observe that given $A \in M(n, \mathbb{C})$ and $\lambda \in \mathbb{C}$, we have

$$\langle (A - \lambda \,\mathrm{Id})\mathbf{z}, \mathbf{w} \rangle = \langle A\mathbf{z}, \mathbf{w} \rangle - \lambda \langle \mathbf{z}, \mathbf{w} \rangle = \langle \mathbf{z}, A^* \mathbf{w} \rangle - \left\langle \mathbf{z}, \overline{\lambda} \mathbf{w} \right\rangle = \left\langle \mathbf{z}, (A^* - \overline{\lambda} \,\mathrm{Id})\mathbf{w} \right\rangle$$

for every $\mathbf{z}, \mathbf{w} \in \mathbb{C}^n$, and hence

$$(A - \lambda \operatorname{Id})^* = A^* - \overline{\lambda} \operatorname{Id}. \tag{12.5}$$

Proposition 12.9. *If $B \in M(n, \mathbb{C})$ is normal, then* $\ker B = \ker B^*$.

Proof. Suppose $B\mathbf{w} = \mathbf{0}$. Then we have

$$\|B^*\mathbf{w}\|^2 = \langle B^*\mathbf{w}, B^*\mathbf{w}\rangle = \langle BB^*\mathbf{w}, \mathbf{w}\rangle = \langle B^*B\mathbf{w}, \mathbf{w}\rangle = 0,$$

and it follows that $\ker B \subset \ker B^*$. Equality holds since $B = (B^*)^*$. □

Applying Proposition 12.9 to $B = A - \lambda \operatorname{Id}$ and using (12.5), we see that if $\mathbf{w}$ is an eigenvector of A with eigenvalue λ, then it is an eigenvector of A^* with eigenvalue $\overline{\lambda}$. In particular, if W is the subspace spanned by $\mathbf{u}_1, \dots, \mathbf{u}_k$, where each $\mathbf{u}^j$ is an eigenvector of A, then each $\mathbf{u}^j$ is an eigenvector of A^* as well, and hence $A^*W \subset W$.

Now we have the following analogue of Proposition 12.5.

Proposition 12.10. *Let $A \in M(n, \mathbb{C})$ be normal, and let $W \subset \mathbb{C}^n$ be an invariant subspace spanned by eigenvectors of A. Then $W^\perp$ is an invariant subspace as well.*

Proof. Given $\mathbf{z} \in W^\perp$ and $\mathbf{w} \in W$, observe that

$$\langle A\mathbf{z}, \mathbf{w}\rangle = \langle \mathbf{z}, A^*\mathbf{w}\rangle = 0,$$

where the last equality follows since $A^*\mathbf{w} \in W$ (by the above discussion). □

This lets us prove the following generalisation of Theorem 12.7.

Theorem 12.11. *An $n \times n$ complex matrix A is normal if and only if there exists $C \in U(n)$ and $\lambda_j \in \mathbb{C}$ such that*

$$CAC^{-1} = \operatorname{diag}(\lambda_1, \dots, \lambda_n). \tag{12.6}$$

Proof. One direction is easy; if (12.6) holds for some $C \in U(n)$ and $\lambda_j \in \mathbb{C}$, then we write $D = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$, and observe that $A = C^{-1}DC = C^*DC$. Thus we have

$$A^* = (C^*DC)^* = C^*D^*(C^*)^* = C^{-1}D^*C,$$

and we see that

$$AA^* = (C^{-1}DC)(C^{-1}D^*C) = C^{-1}DD^*C = C^{-1}D^*DC = A^*A,$$

where the third equality uses the fact that diagonal matrices commute.

The other direction is a word-for-word repetition of the proof of Theorem 12.7, using Proposition 12.10 in place of Proposition 12.5, and omitting the requirement that $|\lambda_j| = 1$. □

Remark. Normality characterises all matrices which can be diagonalised over $\mathbb{C}$ *with an orthonormal change of coordinates.* There are matrices that can be diagonalised with a change of coordinates which is *not* orthonormal; such matrices are not normal with respect to the standard Hermitian product. Recall that the definition of the adjoint A^* depends on the Hermitian product; if we choose a different Hermitian product on $\mathbb{C}^n$, we obtain a different adjoint, and hence a different class of normal matrices.

c. Symmetric matrices. We have settled the question of which matrices can be diagonalised over $\mathbb{C}$ via an orthonormal change of coordinates. What about the real numbers? There are plenty of matrices which can be diagonalised over $\mathbb{C}$ but which cannot be diagonalised over $\mathbb{R}$; any normal matrix with a non-real eigenvalue falls into this class.

Thus we see immediately that any matrix which can be put into the form (12.6) as a map on $\mathbb{R}^n$ must have only real eigenvalues. In particular, given $A \in M(n, \mathbb{R})$, let $A^{\mathbb{C}} \colon \mathbb{C}^n \to \mathbb{C}^n$ be the complexification of A, and observe that $A^T = (A^{\mathbb{C}})^*$. It follows from the remarks before Proposition 12.10 that if λ is an eigenvalue of A, then $\overline{\lambda}$ is an eigenvalue of A^T, with the same eigenvectors.

Definition 12.12. A real $n \times n$ matrix such that $A^T = A$ is called *symmetric*; a complex $n \times n$ matrix such that $A^* = A$ is called *Hermitian.*

If $A \in M(n, \mathbb{R})$ is symmetric, then for every eigenvalue λ and eigenvector $\mathbf{w} \in \mathbb{C}^n$ we have

$$\lambda \mathbf{w} = A\mathbf{w} = A^T \mathbf{w} = \overline{\lambda}\mathbf{w},$$

and hence $\lambda = \overline{\lambda}$. Thus symmetric matrices have only real eigenvalues. In particular, since real symmetric matrices are normal, every real symmetric matrix is orthogonally diagonalisable *over the real numbers.* Furthermore, the converse also holds: if C is a real orthogonal matrix such that $D = CAC^{-1}$ is a diagonal matrix with real entries, then

$$A^T = (C^T DC)^T = C^T D^T (C^T)^T = C^T DC = A,$$

and hence A is symmetric.

d. Linear representations of isometries and other classes of gtransformations. Our discussion of linear algebra began with a quest to understand the isometries of $\mathbb{R}^n$. We have seen various classes of matrices, but have not yet completed that quest—now we are in a position to do so.

We recall the following definition from Lecture 2.

Definition 12.13. A homomorphism $\varphi \colon G \to GL(n, \mathbb{R})$ is called a *linear representation* of G. If $\ker \varphi$ is trivial, we say that the representation is *faithful.*

Informally, a linear representation of a group G is a concrete realisation of the abstract group G as a set of matrices, and it is faithful if no two elements of G are represented by the same matrix. Linear representations are powerful tools, because the group of invertible matrices is general enough to allow us to embed many important abstract groups inside of it, and yet is concrete enough to put all the tools of linear algebra at our disposal in studying the group which is so embedded.

We were able to represent the group of all isometries of $\mathbb{R}^n$ *with a fixed point* as $O(n)$. In order to represent isometries with no fixed point, we must go one dimension higher and consider matrices acting on $\mathbb{R}^{n+1}$.

PROPOSITION 12.14. $\text{Isom}(\mathbb{R}^n)$ *has a linear representation in* $GL(n+1,\mathbb{R})$. *In particular,* $\text{Isom}^+(\mathbb{R}^n)$ *has a linear representation in* $SL(n+1,\mathbb{R})$.

PROOF. Given $I \in \text{Isom}(\mathbb{R}^n)$, let $\mathbf{b} = -I\mathbf{0}$; then $T_{-\mathbf{b}} \circ I\mathbf{0} = \mathbf{0}$, and hence $A = T_{-\mathbf{b}} \circ I \in O(n)$. Thus $I = T_{\mathbf{b}} \circ A$, and so for every $\mathbf{x} \in \mathbb{R}^n$ we have

$$I\mathbf{x} = T_{\mathbf{b}} \circ A\mathbf{x} = A\mathbf{x} + \mathbf{b}. \tag{12.7}$$

Embed $\mathbb{R}^n$ into $\mathbb{R}^{n+1}$ as the plane

$$P = \{\mathbf{x} \in \mathbb{R}^{n+1} \mid x_{n+1} = 1\}.$$

To the isometry I, associate the following block matrix:

$$\varphi(I) = \begin{pmatrix} A & \mathbf{b} \\ \mathbf{0} & 1 \end{pmatrix}. \tag{12.8}$$

Here $A \in O(n)$, $\mathbf{b}$ is an $n \times 1$ column vector, and $\mathbf{0}$ is a $1 \times n$ row vector. Observe that $\varphi(I) \in GL(n+1,\mathbb{R})$, and that $\varphi(I)$ maps P to itself; if $I \in \text{Isom}+(\mathbb{R}^n)$, then $\varphi(I) \in SL(n,\mathbb{R})$. Furthermore, the action of $\varphi(I)$ on P is exactly equal to the action of I on $\mathbb{R}^n$, and φ is a homomorphism: given $I_1, I_2 \in \text{Isom}(\mathbb{R}^n)$, we have

$$\varphi(I_2)\varphi(I_1) = \begin{pmatrix} A_2 & \mathbf{b}_2 \\ \mathbf{0} & 1 \end{pmatrix} \begin{pmatrix} A_1 & \mathbf{b}_1 \\ \mathbf{0} & 1 \end{pmatrix} = \begin{pmatrix} A_2A_1 & A_2\mathbf{b}_1 + \mathbf{b}_2 \\ \mathbf{0} & 1 \end{pmatrix},$$

which is equal to $\varphi(I_2 \circ I_1)$ since

$$I_2 \circ I_1\mathbf{x} = I_2(A_1\mathbf{x} + \mathbf{b}_1) = A_2(A_1\mathbf{x} + \mathbf{b}_1) + \mathbf{b}_2.$$

Finally, we observe that if I is an even isometry, then $\det \varphi(I) = 1$. $\square$

The technique exhibited in the proof of Proposition 12.14 embeds $\text{Isom}(\mathbb{R}^n)$ in $GL(n+1,\mathbb{R}^n)$ as

$$\begin{pmatrix} O(n) & \mathbb{R}^n \\ \mathbf{0} & 1 \end{pmatrix}.$$

Using a the same technique, we can represent the *affine group* $\text{Aff}(\mathbb{R}^n)$, which is the class of all maps which take lines to lines; every such map can again be written in the form (12.7), but here A may be *any* matrix, not necessarily orthogonal. Thus we embed $\text{Aff}(\mathbb{R}^n)$ into $GL(n+1,\mathbb{R}^n)$ as

$$\begin{pmatrix} M(n,\mathbb{R}) & \mathbb{R}^n \\ \mathbf{0} & 1 \end{pmatrix}.$$

We may also do this with the group of *similarity transformations*—maps of $\mathbb{R}^n$ which take lines to lines and preserve angles. Every such map may be written as $\mathbf{x} \mapsto \lambda R\mathbf{x} + \mathbf{b}$, where $\lambda \in \mathbb{R}$ and $R \in O(n)$. Thus the group embeds into the general linear group as

$$\begin{pmatrix} \mathbb{R} \cdot O(n) & \mathbb{R}^n \\ \mathbf{0} & 1 \end{pmatrix}.$$

The common thread in all these representations is that all the tools of linear algebra are now at our disposal. For example, suppose we wish to classify isometries of $\mathbb{R}^n$, and have forgotten all the synthetic geometry we ever knew. Then we observe that every isometry can be written as $I\mathbf{x} = A\mathbf{x} + \mathbf{b}$, and note that I has a fixed point if and only if

$$A\mathbf{x} + \mathbf{b} = \mathbf{x}$$

has a solution—that is, if and only if $\mathbf{b}$ lies in the range of $A - \mathrm{Id}$. If 1 is not an eigenvalue of A, then $A - \mathrm{Id}$ is invertible, and I has a fixed point. If 1 *is* an eigenvalue of A, then $\mathbf{b}$ may not lie in the range of $A - \mathrm{Id}$, that is the orthogonal complement to the eigenspace L_1 of vectors with eigenvalue 1 and I then has no fixed points. As before let us decompose $\mathbf{b} = \mathbf{b}_1 + \mathbf{b}_2$ where $\mathbf{b}_1 \in L_1$, i.e $A\mathbf{b}_1 = \mathbf{b}_1$, and $\mathbf{b}_2$ orthogonal to L_1, i.e. in the range of $A - \mathrm{Id}$. Then I is the composition of the isometry $I' : \mathbf{x} \to A\mathbf{x} + \mathbf{b}_2$ that has a fixed point and translation $T_{\mathbf{b}_1}$ by $\mathbf{b}_1$. I' is conjugate via a translation the to linear isometry A and I to the product of it with the translation $T_{\mathbf{b}_1}$. Notice that $AT_{\mathbf{b}_1}\mathbf{x} = A(\mathbf{x} + \mathbf{b}_1) = Ax + A\mathbf{b}_1 = A\mathbf{x} + \mathbf{b}_1 = T_{\mathbf{b}_1}A\mathbf{x}$.

Thus any isometry I without fixed points is is the product of a commuting pair of an isometry I_0 with fixed many points and a translation along the fixed set of that isometry. Depending on the dimension of the fixed set for I_0 we obtain different geometric types of fixed point free isometries.

Similar arguments provide for the classification of similarity transformations and affine transformations without fixed points.

Lecture 13. Wednesday, September 30

a. The projective line. In the previous lecture we saw that various linear groups correspond to various "geometries". Although we have spent most of our time studying the group of isometries of $\mathbb{R}^n$, we also saw that the group of affine transformations and the group of similarity transformations appear as subgroups of $GL(n+1,\mathbb{R})$. Thus we may go beyond the usual Euclidean structure of $\mathbb{R}^n$ and consider instead the affine structure of $\mathbb{R}^n$, or perhaps think about Euclidean geometry up to similarity.

There are other matrix groups which are of interest to us, and it turns out that they too correspond to certain "geometries". For example, each of the above examples arose from considering subgroups of the affine transformations on $\mathbb{R}^n$; that is, subgroups of $GL(n+1,\mathbb{R})$ of the form

$$\begin{pmatrix} G & \mathbb{R}^n \\ \mathbf{0} & 1 \end{pmatrix}, \tag{13.1}$$

where G is a subgroup of $GL(n,\mathbb{R})$. Such subgroups act on the n-dimensional subspace $P = \{\mathbf{x} \in \mathbb{R}^{n+1} \mid x_{n+1} = 1\}$.

In this lecture we will broaden our horizons beyond the groups (13.1), examining instead the action of all of $GL(n+1,\mathbb{R})$ on P. This will lead us in the end to *projective geometry.* We will take our time getting there, however, because the story is not quite as straightforward as it was before; for example, observe that most linear transformations of $\mathbb{R}^{n+1}$ do not preserve the subspace P, and so it is not at all obvious in what sense they are to "act" on P.

The fundamental fact that we *do* know about elements of $GL(n+1,\mathbb{R})$ is that they map lines to lines. Thus it makes sense to consider the action of $GL(n+1,\mathbb{R})$ on lines in $\mathbb{R}^{n+1}$; we begin in the simplest case, $n=1$. Here we have $GL(2,\mathbb{R})$ acting on $\mathbb{R}^2$, and in particular, on the following object.

Definition 13.1. The *real projective line* $\mathbb{R}P(1)$ is the set of all lines through the origin in $\mathbb{R}^2$.

Clearly if $A \in GL(2,\mathbb{R})$ and $\ell \in \mathbb{R}P(1)$ is a line, then $A\ell$ is a line as well, and so $A\ell \in \mathbb{R}P(1)$. Furthermore, multiplying A by a scalar does not change its action on $\mathbb{R}P(1)$, and so we may multiply A by $(\det A)^{-1}$ and deal only with matrices in $SL(2,\mathbb{R})$.

If one insists on thinking of geometric objects as being things whose fundamental building blocks are "points", rather than lines, the following construction is useful. Fix a line $\ell_0 \subset \mathbb{R}^2$ which does *not* pass through the origin, and then observe that every line $\ell \in \mathbb{R}P(1)$ intersects ℓ_0 in a unique point, with one exception: the line through $\mathbf{0}$ parallel to ℓ_0 never intersects ℓ_0. Associating this line to the "point at infinity", we obtain the following bijection between $\mathbb{R}P(1)$ and $\ell_0 \cup \{\infty\}$:

$$\ell \mapsto \begin{cases} \ell \cap \ell_0 & \ell \nparallel \ell_0, \\ \infty & \ell \parallel \ell_0. \end{cases}$$

Upon observing that the unit circle represents all directions that one might travel along a line from the origin, it is tempting to think of $\mathbb{R}P(1)$ as a circle—don't do it! Elements of $\mathbb{R}P(1)$ are *lines*, not rays, and do not carry an orientation; there is no "positive" direction along ℓ, and so ℓ and $-\ell$ are equivalent. Every element of $\mathbb{R}P(1)$ intersects the circle in not one, but *two* points; thus while we can study $\mathbb{R}P(1)$ using the unit circle, the proper model is the unit circle *with opposite points identified.*[3]

In particular, the linear map $-\operatorname{Id}\colon \mathbf{x} \mapsto -\mathbf{x}$ (which in $\mathbb{R}^2$ is rotation by π) fixes every element of $\mathbb{R}P(1)$, despite having determinant 1 and thus lying in $SL(2,\mathbb{R})$. Consequently, if we want to describe the possible maps on $\mathbb{R}P(1)$ which are induced by elements of $SL(2,\mathbb{R})$, we should factor out the two-element subgroup $\{\operatorname{Id}, -\operatorname{Id}\}$, which turns out to be the centre of $SL(2,\mathbb{R})$. This is analogous to the procedure by which we went from $O(3)$ to $SO(3)$; in this case we obtain the *projective special linear group*

$$PSL(2,\mathbb{R}) = SL(2,\mathbb{R})/\{id, -\operatorname{Id}\}.$$

Every element of $PSL(2,\mathbb{R})$ corresponds to a two-element equivalence class $\{A, -A\}$ in $SL(2,\mathbb{R})$—for the sake of convenience, we will denote this equivalence class simply by A. Each such element induces a *projective transformation* on $\mathbb{R}P(1)$. Furthermore, the action of $PSL(2,\mathbb{R})$ on $\mathbb{R}P(1)$ is *faithful*—no two elements of $PSL(2,\mathbb{R})$ induce the same projective transformation.

What do these transformations look like? Returning to the model of $\mathbb{R}P(1)$ as $\ell_0 \cup \infty$, let ℓ_0 be the line $\{(x,y) \in \mathbb{R}^2 \mid y = 1\}$. (This corresponds to the subspace P from the opening discussion.) Given $A = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in PSL(2,\mathbb{R})$ and $(x,1) \in \ell_0$, we have

$$A\begin{pmatrix} x \\ 1 \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} x \\ 1 \end{pmatrix} = \begin{pmatrix} ax+b \\ cx+d \end{pmatrix}.$$

Thus the line ℓ through $\mathbf{0}$ and $(x,1)$ is mapped by A to the line $A\ell$ through $\mathbf{0}$ and $(ax+b, cx+d)$. To find the point in which $A\ell$ intersects ℓ_0, we normalise by the y-coordinate and observe that

$$A\ell \cap \ell_0 = \left\{\left(\frac{ax+b}{cx+d}, 1\right)\right\}.$$

Thus A acts on $\mathbb{R} \cup \{\infty\}$ as the *fractional linear transformation*

$$f_A\colon x \mapsto \frac{ax+b}{cx+d}.$$

Observe that the point at infinity is essential to this picture; we see that

$$f_A(\infty) = \frac{a}{c}, \qquad f_A\left(-\frac{d}{c}\right) = \infty,$$

[3] It is in fact true that $\mathbb{R}P(1)$ and S^1 are topologically equivalent—this is a fluke which happens only in this lowest of dimensions. Already for $n = 2$, we will see that $\mathbb{R}P(2)$ and S^2 are very different.

and ∞ behaves just like any other point on $\mathbb{R}P(1)$. Furthermore, $f_A(\infty) = \infty$ (or equivalently, $f_A(\mathbb{R}) = \mathbb{R}$) if and only if $c = 0$, in which case the matrix A takes the form (13.1).

It is natural to ask if there is an intrinsic way to define projective transformations on $\mathbb{R} \cup \{\infty\}$ without resorting to a particular embedding in $\mathbb{R}^2$, which is somehow extrinsic. Obviously there are many maps on $\mathbb{R} \cup \{\infty\}$ which are *not* projective transformations—that is, which cannot be written as fractional linear transformations. What geometric property sets the fractional linear transformations apart?

The following seemingly capricious definition turns out to be useful.

DEFINITION 13.2. Given four points $x_1, x_2, x_3, x_4 \in \mathbb{R} \cup \{\infty\}$, the *cross-ratio* of the four points is

$$(x_1, x_2; x_3, x_4) = \frac{x_1 - x_3}{x_2 - x_3} \div \frac{x_1 - x_4}{x_2 - x_4}, \tag{13.2}$$

where expressions involving multiplication or division by ∞ or 0 are evaluated in the obvious way.

PROPOSITION 13.3. *A map $f \colon \mathbb{R} \cup \{\infty\} \to \mathbb{R} \cup \{\infty\}$ is projective if and only if it preserves the cross-ratio—that is,*

$$(f(x_1), f(x_2); f(x_3), f(x_4)) = (x_1, x_2; x_3, x_4) \tag{13.3}$$

for every $x_1, x_2, x_3, x_4 \in \mathbb{R} \cup \{\infty\}$.

PROOF. The fact that (13.3) holds whenever f is a fractional linear transformation is an easy exercise. To see the converse, observe that any map preserving the cross-ratio is determined by its action on three points. In particular, if $x_1, x_2, x_3, y_1, y_2, y_3 \in \mathbb{R} \cup \{\infty\}$ are arbitrary, then there is a unique map $f \colon \mathbb{R} \cup \{\infty\} \to \mathbb{R} \cup \{\infty\}$ such that $f(x_i) = y_i$ and (13.3) holds. This follows because the equation

$$\frac{x_1 - x_3}{x_2 - x_3} \div \frac{x_1 - x}{x_2 - x} = \frac{y_1 - y_3}{y_2 - y_3} \div \frac{y_1 - y}{y_2 - y}$$

can be solved for y as a function of x, and upon doing so one sees that f is a fractional linear transformation. □

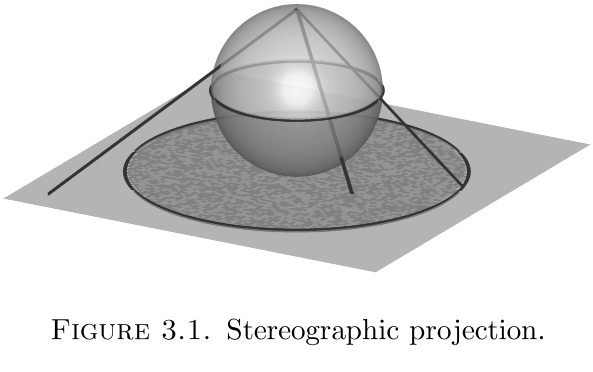


FIGURE 3.1. Stereographic projection.

b. The projective plane. What happens in higher dimensions? As we did with $\mathbb{R}P(1)$, we want to construct some geometric object which is a compactification of $\mathbb{R}^2$. One way to do this is to add a single point at infinity, as we did before; this corresponds to the *stereographic projection* shown in Figure 3.1, which "wraps" the plane around the sphere, covering every point except the north pole, which is identified with ∞.

We will come back to this construction later; for now we take a different course and generalise the construction in terms of lines in $\mathbb{R}^2$. In particular, denote by $\mathbb{R}P(2)$ the set of all lines $\ell \subset \mathbb{R}^3$ which pass through the origin; this is the *projective plane.* Let $P \subset \mathbb{R}^3$ be any plane which does not contain the origin; then every element $\ell \in \mathbb{R}P(2)$ which is not parallel to P intersects P in exactly one point.

This is where the difference arises. This time around there are many lines parallel to P, each of which corresponds to a "point at infinity". In fact, the set of lines parallel to P is just the set of lines in the plane through the origin parallel to P; but we have already characterised the set of lines in a plane as $\mathbb{R}P(1)$! Thus we see that $\mathbb{R}P(2)$ is the plane $\mathbb{R}^2$ together with a copy of the projective line at infinity:

$$\mathbb{R}P(2) = \mathbb{R}^2 \cup \mathbb{R}P(1) = \mathbb{R}^2 \cup \mathbb{R} \cup \{\infty\}.$$

As before, this is a purely set-theoretic description at this point; the geometric structure of the projective plane is encoded by the action of $SL(3,\mathbb{R})$ on $\mathbb{R}^3$. Taking $P = \{(x,y,z) \in \mathbb{R}^3 \mid z = 1\}$, a similar argument to the one in the previous section shows that a matrix $A = (a_{ij}) \in SL(3,\mathbb{R})$ induces the following action on $\mathbb{R}P(2)$:

$$f_A(x,y) = \left(\frac{a_{11}x + a_{12}y + a_{13}}{a_{31}x + a_{32}y + a_{33}}, \frac{a_{21}x + a_{22}y + a_{23}}{a_{31}x + a_{32}y + a_{33}}\right). \tag{13.4}$$

Observe that because $\mathbb{R}^3$ has odd dimension, the central symmetry $\mathbf{x} \mapsto -\mathbf{x}$ is *not* contained in $SL(3,\mathbb{R})$, and thus the action of $SL(3,\mathbb{R})$ on $\mathbb{R}P(2)$ is faithful.

We have described the points in $\mathbb{R}P(2)$: they are lines in $\mathbb{R}^3$ passing through the origin. But $\mathbb{R}P(2)$ is a two-dimensional object, and so it should have more internal structure than just points—it should have *lines.* What are the lines in $\mathbb{R}P(2)$?

Recall that $\mathbb{R}P(2)$ may be represented by the plane P together with the projective line at infinity. Let ℓ be a line lying in P, and let Q_ℓ be the plane in $\mathbb{R}^3$ which contains both $\mathbf{0}$ and ℓ. Then Q_ℓ is the union of all lines in $\mathbb{R}^3$ which pass through the origin and a point of ℓ; thus it may be thought of as a line in $\mathbb{R}P(2)$. That is, *lines* in $\mathbb{R}P(2)$ correspond to *planes* in $\mathbb{R}^3$.

Exercise 13.1. Which plane in $\mathbb{R}^3$ corresponds to the projective line at infinity?

As before, projective transformations of $\mathbb{R}P(2)$ are exactly those transformations which are induced by the action of $SL(3,\mathbb{R})$, and hence which have the form (13.4).

However there is an important difference with the one-dimensional case. Projective transformations can be characterized in purely geometric terms, somewhat similar to characterization of similarity transformations as those that preserve lines and angles. There is a whole range of such results characterizing different geometries

Theorem 13.4. *A map $f\colon \mathbb{R}P(2) \to \mathbb{R}P(2)$ is a projective transformation if and only if it maps projective lines to projective lines.*

As in Proposition 13.3, one direction is obvious. To see the converse we will show that a map which respects projective lines is a projective transformation, is determined by the images of four (projective) points not on the same (projective) line. This requires a certain geometric construction [4] that we will describe in the next lecture along with a similar construction in affine geometry.

The action of $SL(3, \mathbb{R})$ on $\mathbb{R}P(2)$ describes what is known as *projective geometry.* Observe that since every line in $\mathbb{R}P(2)$ intersects the unit sphere S^2 in exactly two antipodal points, we can write $\mathbb{R}P(2)$ as the factor space S^2/C, where $C\colon \mathbf{x} \to -\mathbf{x}$ is the central symmetry. This factor space (which is topologically quite different from the sphere itself) inherits a natural metric from the metric on S^2; the distance between two equivalence classes $\{\mathbf{x}, -\mathbf{x}\}$ and $\{\mathbf{y}, -\mathbf{y}\}$ is just the smaller of $d(\mathbf{x}, \mathbf{y})$ and $d(\mathbf{x}, -\mathbf{y})$, where d is distance along the surface of the sphere.[5] The distance may also be computed as the smaller of the two angles made by the lines through the origin which contain $\mathbf{x}$ and $\mathbf{y}$.

When we equip $\mathbb{R}P(2)$ with the metric just defined, we refer to it as the *elliptic plane* and denote it by $\mathbb{E}^2$. As we saw on the homework, the isometry group of $\mathbb{E}^2$ is $SO(3)$, the group of rotations of the sphere. This is a much smaller group than $SL(3, \mathbb{R})$, the symmetry group of $\mathbb{R}P(2)$, and so elliptic geometry is a more restrictive thing than projective geometry. The relationship between the two is analogous to the relationship between affine geometry and Euclidean geometry on $\mathbb{R}^n$; the former has a much larger symmetry group than the latter, which includes a notion of a metric.

Elliptic geometry can be thought of as spherical geometry with the non-uniqueness factored out. As mentioned above, projective lines correspond to real planes. A (real) plane through the origin intersects the sphere S^2 in a great circle, and so lines in the elliptic plane are great circles on the sphere (with antipodal points identified). Thus every pair of (elliptic) lines intersects in a unique (elliptic) point.

Remark. These constructions generalise to arbitrary dimensions. If we consider the set of all lines through the origin in $\mathbb{R}^{n+1}$, we obtain the

[4] That is in fact algebraic in nature since it can be carried out to projective spaces constructed over fields other than real numbers.

[5] Note that d is a different quantity from the usual distance in $\mathbb{R}^3$, which corresponds to being allowed to tunnel through the interior of the sphere. However, one quantity determines the other, and so they determine the same class of isometries.

projective space $\mathbb{R}P^n$. If we equip this space with a metric inherited from the sphere with antipodal points identified, we obtain the elliptic space $\mathbb{E}^n$. The relationship between the two spaces is just as it was in two dimensions.

c. The Riemann sphere. We return now to the one-point compactification of the plane, which seems to be a simpler object than the projective plane we wound up discussing, as it only adds a single point to the usual plane. It turns out that by viewing the plane as $\mathbb{C}$, rather than as $\mathbb{R}^2$, we can make sense of this.

In particular, recall our construction of the projective line $\mathbb{R}P(1)$, and now construct the *complex* projective line $\mathbb{C}P(1)$: this is the set of all *complex* lines (each of which is a real plane) through the origin in $\mathbb{C}^2$. Writing $Q = \{(z, w) \in \mathbb{C}^2 \mid w = 1\}$, we once again find that every complex line $az + bw = 0$ intersects Q in a single point, with the exception of the line $w = 0$, which again corresponds to the point at infinity.

Thus $\mathbb{C}P(1) = \mathbb{C} \cup \{\infty\}$, (called the *Riemann sphere*) and $SL(2, \mathbb{C})$ acts on $\mathbb{C}P(1)$ in the same way $SL(2, \mathbb{R})$ acted on $\mathbb{R}P(1)$—by fractional linear transformations. As before, these are characterised by the property of preserving the cross-ratio, which can be defined for complex numbers by the same formula (13.2), and the proof is exactly as it was in Proposition 13.3.

Geometrically, it can be shown that the relevant set of geometric objects in $\mathbb{C}P(1)$ is no longer the set of all lines, but the set of all lines and circles. If γ is a path in $\mathbb{C}P(1)$ which is either a line or a circle, then $f_A(\gamma)$ is either a line or a circle for every $A \in SL(2, \mathbb{C})$; furthermore, every f with this property is a fractional linear transformation, provided f preserves orientation.[6] It may be the case that f_A maps a line to a circle and vice versa; for example, consider the image of the unit circle $\{z \in \mathbb{C} \mid |z| = 1\}$ under the map f_A corresponding to the matrix $\left(\begin{smallmatrix} 0 & 1 \\ 1 & -1 \end{smallmatrix}\right)$. There are other transformations of the Riemann sphere that map lines and circles into lines and circles, for example the complex involution $z \to \bar{z}$ and hence its composition with any fractional linear transformation that has the form $z \to \frac{a\bar{z}+b}{c\bar{z}+d}$. It turns out that the property of mapping lines and circles into lines and circles characterizes exactly these two types of transformations. We will prove this in due time. Not surprisingly, the approach will be similar to the proofs of characterization of affine and projective transformations: we will show that images of *four points not belonging to a circle* uniquely determine a transformation that maps lines and circles into lines and circles.

To compare this with the geometry of the real projective plane, we recall that real projective transformations map lines to lines, and remark that circles may not be mapped to circles. A (projective) circle corresponds to a (real) cone, and the image of a cone under a linear transformation is again a cone, but one which may intersect the plane P in a different conic section.

[6] The map $z \mapsto \overline{z}$ has this property but reverses orientation, and in fact, any orientation-reversing map with this property may be written as $f(z) = (a\overline{z} + b)/(c\overline{z} + d)$ for some $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in SL(2, \mathbb{C})$.

Thus conics are mapped to conics under real projective transformations: the image of a circle may be either a circle, an ellipse, a hyperbola, or a parabola.

Of course, the group $SL(2,\mathbb{R})$ is a subgroup of $SL(2,\mathbb{C})$, and so it also acts on the Riemann sphere $\mathbb{C}P(1)$. However, it fixes the real line (together with the point at infinity), and preserves both the upper and lower half-planes. In fact, it can be shown to act transitively on each of these half-planes, and the geometric structure corresponding to this symmetry group is the *hyperbolic plane* $\mathbb{H}^2$.

While we do not yet see a notion of distance on $\mathbb{H}^2$, one will appear eventually. There are underlying themes here which are also present for $\mathbb{R}^2$ and $\mathbb{E}^2$. In the case of the Euclidean plane (a two-dimensional object), the group of isometries is

$$\begin{pmatrix} O(2) & \mathbb{R}^2 \\ \mathbf{0} & 1 \end{pmatrix},$$

which is three-dimensional since $O(2)$ has one degree of freedom. This group acts transitively on $\mathbb{R}^2$, and so $\mathbb{R}^2$ ought to possess some property which is invariant under the group action; this is exactly the notion of distance. A similar thing occurs for the elliptic plane (also a two-dimensional object), where the group of isometries is $SO(3)$ (again a three-dimensional group). The hyperbolic plane is two-dimensional and the symmetry group $SL(2,\mathbb{R})$ is three-dimensional, and so we will in the end find a notion of distance here as well.

Finally, notice that the construction of hyperbolic plane can be extended to higher dimensions similarly to that of Euclidean and elliptic planes. However, direct connection with complex numbers is lost: in dim three it appears in a different disguise since the groups of isometries turns out to be $SL(2,\mathbb{C})$, but in higher dimensions it disappears altogether.

Lecture 14. Friday, October 2

a. Review of various geometries. We have studied a number of two-dimensional geometries, each associated to a different group of transformations. Some of these underlying spaces of these geometries are topologically distinct, while others are homeomorphic to each other. In each case, we can give a synthetic characterisation of the elements of the transformation group; this information is summarised in Table 1.

Group	Dimension	Space	Preserves	n-dimensional
$SL(3,\mathbb{R})$	8	$\mathbb{R}P(2)$	lines	$SL(n+1,\mathbb{R})$, $\mathbb{R}P(n)$
$\mathrm{Aff}(\mathbb{R}^2)$	6	$\mathbb{R}^2$	lines	$\mathrm{Aff}(\mathbb{R}^n)$, $\mathbb{R}^n)$
$SL(2,\mathbb{C})$	6	$\mathbb{C}P(1)$	lines and circles	$SL(n,\mathbb{C})$, $\mathbb{C}P(n-1)$
$SO(3)$	3	$\mathbb{E}^2$	distances	$SO(n+1)$, $\mathbb{E}^n$
$\mathrm{Isom}(\mathbb{R}^2)$	3	$\mathbb{R}^2$	distances	$\mathrm{Isom}(\mathbb{R}^n)$, $\mathbb{R}^n$
$SL(2,\mathbb{R})$	3	$\mathbb{H}^2$	distances	$SO(n,1)$, $\mathbb{H}^n$

Table 1. Six different two-dimensional geometries and their generalizations.

For some of these six examples, we have already investigated some of the following algebraic properties of the group of transformation—conjugacy classes, finite and discrete subgroups, normal subgroups, centre of the group, etc. Fourth column contains geometric information that *characterizes* transformations from the corresponding group. So far we established this characterization only for $\mathbb{R}^2$ and $\mathbb{E}^2$ (the latter in a homework problem); we will shortly do that (with small caveats) for the first three groups in the list. As for the last one we have not yet defined the distance in the hyperbolic plane, let alone characterized isometries there. This will also be done in due time.

There are certain issues related to orientation:

- Groups $\mathrm{Isom}(\mathbb{R}^2)$ and $\mathrm{Aff}(\mathbb{R}^2)$ contain index two subgroups $\mathrm{Isom}^+(\mathbb{R}^2)$ and $\mathrm{Aff}^+(\mathbb{R}^2)$ of orientation preserving transformations. This subgroup and its other coset are separated connected components in the group. Notice drastic differences in the structure of conjugacy classes in the orientation preserving and orientation reversing cases.
- $SL(2,\mathbb{C})$ is the groups of transformations of the Riemann sphere preserving lines and circles and also orientation; the full groups of transformations preserving lines and circles is $SL(2,\mathbb{C})\times\mathbb{Z}/2\mathbb{Z}$ the second component generated by the complex conjugation.
- The same comment applies to the hyperbolic plane, where as we pointed out, the group $SL(2,\mathbb{R})$ does not act faithfully but its factor by the center $PSL(2,\mathbb{R}) = SL(2,\mathbb{R})/\pm\mathrm{Id}$ does and constitutes the groups of orientation preserving isometries.

There are natural embeddings among those six groups that have direct geometric meaning:

$\mathrm{Isom}^+(\mathbb{R}^2) \subset \mathrm{Aff}^+(\mathbb{R}^2) \subset SL(3,\mathbb{R})$ (isometries are affine transformations and affine transformations are exactly projective ones that preserve the line at infinity)

$SO(3) \subset SL(3,\mathbb{R})$ (isometries are projective transformations)

$SL(2,\mathbb{R}) \subset SL(2,\mathbb{C})$ (isometries of hyperbolic plane are fractional linear transformations of the Riemann sphere preserving the upper half-plane).

EXERCISE 14.1. Prove that the group $SL(2,\mathbb{C})$ cannot be embedded into $SL(3,\mathbb{R})$.

EXERCISE 14.2. Add the line for the group of similarity transformations of $\mathbb{R}^2$ and find natural embeddings involving that group.

Before proceeding to the proofs of geometric characterization of the first three groups in the table let us make brief comments about higher-dimensional generalizations.

For the first, second, fourth and fifth lines generalizations are straightforward and geometric characterizations stand. The only point is orientability of projective spaces. Since in $GN(n,\mathbb{R})$ $\det(-\mathrm{Id}) = (-1)^n$ odd-dimensional projective (and hence elliptic) spaces are orientable and their full isometry groups have two components similarly to $\mathrm{Isom}(\mathbb{R}^2)$ and $\mathrm{Aff}(\mathbb{R}^2)$.

Complex projective space has dimension $2n-2$ and for $n \geq 2$ is not even topologically equivalent to the sphere. Even though it has lots of complex "lines" that look like the Riemann sphere, its global structure is quite different.

Geometric construction of the hyperbolic plane as the half plane (the *Poincaré model*) extends directly to higher dimension. But the group of (orientation preserving) isometries of $\mathbb{H}^n$ turns out to be $SO(n,1)$, the group of linear transformations in $\mathbb{R}^{n+1}$, preserving the bilinear form

$$\sum_{i=1}^{n} x_i y_i - x_{n+1} y_{n+1},$$

that has nothing to do with either SL's or complex numbers. The form of this group becomes transparent from a different model, *the hyperboloid model*, that is closely related to the *Klein model*. Isomorphism between $SO(3,1)$ and $SL(2,\mathbb{R})$ is a pre-eminent example of the *low-dimensional phenomena* when early members of various series of matrix groups match, resulting in profound connections between fundamental structures of geometry and physics.

b. Affine geometry. Recall that the group of affine transformations can be faithfully represented in $GL(3,\mathbb{R})$ as

$$\begin{pmatrix} GL(2,\mathbb{R}) & \mathbb{R}^2 \\ \mathbf{0} & 1 \end{pmatrix}, \tag{14.1}$$

where $L \in GL(2,\mathbb{R})$ and $\mathbf{b} \in \mathbb{R}^2$ correspond to the affine transformation $A\colon \mathbf{x} \mapsto L\mathbf{x} + \mathbf{b}$. Thus we must specify 6 real numbers to determine an element of $\mathrm{Aff}(\mathbb{R}^2)$. Four of these determine the linear part L (there is a restriction on these four, namely that $\det L \neq 0$, but this does not remove a degree of freedom since we only remove one possible value of the fourth number to be chosen), while the other two determine the translation part $\mathbf{b}$. This explains the claim in the second column of Table 1 that $\mathrm{Aff}(\mathbb{R}^2)$ is a 6-dimensional group.

THEOREM 14.1. *A bijection $A\colon \mathbb{R}^2 \to \mathbb{R}^2$ is affine if and only if it maps lines to lines—that is, if and only if $A\ell \subset \mathbb{R}^2$ is a straight line whenever $\ell \subset \mathbb{R}^2$ is.*

PROOF. Let T be a bijection of $\mathbb{R}^2$ that takes lines into lines. Notice that it takes unique intersection point of two non-parallel line to a point, and hence it maps parallel lines to parallel lines.

LEMMA 14.2. *For any map T that maps lines into lines maps the midpoint of any segment $[p,q]$ into the midpoint of the segment $[Tp,Tq]$. For any point p, vector v and natural number n*

$$T(p+nv) = Tp + nTv. \tag{14.2}$$

PROOF. Take any parallelogram $\mathcal{P}$ whose one diagonal is $[p,q]$. The midpoint is the intersection of diagonals of $\mathcal{P}$. Hence $T(\frac{p+q}{2})$ is the point of intersection of the diagonals of the parallelogram. $T\mathcal{P}$, i.e. $\frac{Tp+Tq}{2}$.

Taking the image of the parallelogram with vertices at the origin, p, v and $p+v$ we see that $T(p+v) = Tp + Tv$. Now $p+v$ is the midpoint of the segment $[p, p+2v]$, hence its image is the midpoint of the segment $[Tp, T(p+2v)]$, so that $T(p+2v) = Tp + 2Tv$, $p+2v$ is the midpoint of the segment $[p+v, p+3v]$, and by induction in n we obtain (14.2). □

Since for any pairs of triples of non-collinear points there is a transformation $S \in \mathrm{Aff}^(\mathbb{R}^2)$ that matches the triples (and takes lines into lines) we may assume that T fixes three points: the origin, $(0,1)$ and $(1,0)$. Our goal is to show that such a map that still maps lines into lines is the identity. Notice furthermore that it is sufficient to show that it is the identity on the coordinate axes since any point in the plane is a midpoint of the segment whose endpoints lie on coordinate axes.

At this point we make an important comment. If we assume that the map is continuous (along the coordinate axes is sufficient) the proof can be easily completed. First notice that the integer vectors on the x-axis are preserved by Lemma 14.2. Then, by induction in n, $T((k/2^n, 0) = (k/2^n, 0)$ for all integers k and $n = 1, 2, \ldots$. Since those numbers are dense in $\mathbb{R}$, T is the identity on the x-axis and, by the same argument, on the y-axis. It would also be sufficient to assume that T preserves the order of points on the x axis.

But we do not assume continuity or monotonicity so we need another argument. A remarkable consequence of the proof to follow is its purely algebraic character until the last step. We use algebraic manipulations with real numbers that can be made with elements of an arbitrary field.

Since our map preserves the coordinate axes it takes lines parallel to those axes to parallel lines and thus can be written in the coordinate form:

$$T(x, y) = (f(x), g(y)),$$

where f and g are bijections of $\mathbb{R}$.

Since the map fixes the line $x+y=1$, it also maps each line $x+y = const$ into another such line. Hence we immediately see that $f = g$. As we already know $f(k/2^n) = k/2^n$. Furthermore, by Lemma 14.2 f is an *additive* map:

$$f(x + y) = f(x) + f(y).$$

Now take the line l through points $(x, 0)$ and $(0, y)$. Any point $p \in l$ has the form $(tx, (1 - t)y)$ for some $t \in \mathbb{R}$. The line Tl passes through $(f(x), 0)$ and $(0, f(y))$ and hence has the form $(sf(x), (1-s)f(y)) = (f(tx), f((1-t)y)$ for some $s \in \mathbb{R}$, Comparing first coordinates we obtain $s = \frac{f(tx)}{f(x)}$. Subssituting this into the equality for the second coordinate gives $(1 - \frac{f(tx)}{f(x)})f(y) = f((1 - t)y)$ or $f(x)f(y) = f(x)f((1 - t)y) + f(tx)f(y)$. By additivity $f((1 - t)y) = f(y) - f(ty)$. Substituting this into the last equation we obtain after cancellations

$$f(tx)f(y) = f(ty)f(x).$$

Since for $x = 1$, $f(x) = 1$ we deduce *multiplicativity* of the map f:

$$f(ty) = f(t)f(y).$$

Since f is also additive it is an *automorphism of the field of real numbers.* Up to here the argument works for any field.

But now if $t = s^2 > 0$ it follows that $f(t) = (f(s))^2 > 0$ hence by additivity f is monotone. But any monotone additive function is obviously linear since it is linear on a dense set. In out case $f = \mathrm{Id}$. □

The affine plane has natural analogues in higher dimensions: by replacing $GL(2, \mathbb{R})$ and $\mathbb{R}^2$ in (14.1) with $GL(n, \mathbb{R})$ and $\mathbb{R}^n$, we obtain the group of affine transformations $\mathrm{Aff}(\mathbb{R}^n)$ acting on $\mathbb{R}^n$.

The proof given above extends by a straightforward induction to arbitrary dimension after noticing that bijectivity and preservation of lines implies preservation of affine subspaces of any dimension.

c. Projective geometry. Returning to the first row of the column, we examine the projective plane $\mathbb{R}P(2)$. The group of transformations is $SL(3, \mathbb{R})$, which is 8-dimensional because the 9 entries of an element of $SL(3, \mathbb{R})$ can be any real numbers subject to a single constraint, that the determinant (which is a polynomial in the entries) be 1.

Projective transformations admit a similar characterisation to affine transformations.

Theorem 14.3. *A bijection of $\mathbb{R}P(2)$ is a projective transformation if and only if it maps (projective) lines to (projective) lines.*

Proof. Recall that as a set, the projective plane $\mathbb{R}P(2)$ is just the plane $\mathbb{R}^2$ together with a projective line $\mathbb{R}P(1)$ at infinity. With the exception of the line at infinity, which we shall denote ℓ^∞, every projective line is a real line in $\mathbb{R}^2$ together with a single point at infinity.

One direction is immediate. For the other direction, let $T\colon \mathbb{R}P(2) \to \mathbb{R}P(2)$ be a bijection which maps lines to lines; we prove that T is a projective transformation. Let $\ell = T\ell^\infty$ be the image of the line at infinity under the action of T; thus ℓ is again a projective line, and there exists a (non-unique) projective transformation $R \in SL(3,\mathbb{R})$ such that $R(\ell) = \ell^\infty$.

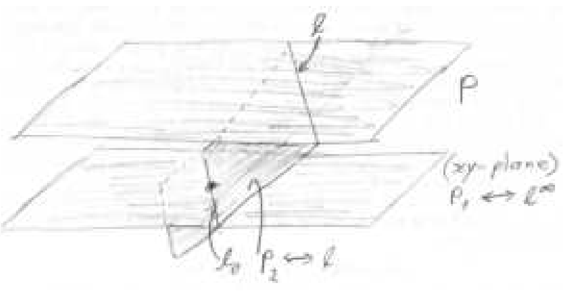


Figure 3.2. Sending a projective line to infinity.

To see this explicitly, observe that if we think of $\mathbb{R}P(2)$ as the set of all lines through the origin in $\mathbb{R}^3$, then projective lines correspond to real planes containing the origin. Taking the plane $P = \{\mathbf{x} \in \mathbb{R}^3 \mid x_3 = 1\}$ as our representative of $\mathbb{R}^2$, the line at infinity ℓ^∞ corresponds to the xy-plane P_1, and the line ℓ corresponds to the plane P_2 which contains ℓ and $\mathbf{0}$ (see Figure 3.2). The planes P_1 and P_2 both contain the origin, thus they intersect in a (real) line ℓ_0. Let $R \in SO(3) \subset SL(3,\mathbb{R})$ be the rotation around ℓ_0 such that $R(P_2) = P_1$. In projective terms, this means that $R(\ell) = \ell^\infty$.

Now we observe that $R' = R \circ T$ is again a bijection of $\mathbb{R}P(2)$ which takes projective lines to projective lines. Furthermore, $R'(\ell^\infty) = \ell^\infty$, and so R' also acts as a bijection of $\mathbb{R}^2$ which takes lines in $\mathbb{R}^2$ to lines in $\mathbb{R}^2$. It follows from Theorem 14.1 that R' is an affine map; but affine maps are also projective. Hence we have obtained to projective maps R and R' such that $T = R^{-1} \circ R'$, and it follows that T itself is projective. □

As with affine geometry, the generalisation of projective geometry to higher dimensions is straightforward. The projective space $\mathbb{R}P(n)$ is once again defined as the set of all lines through the origin in $\mathbb{R}^{n+1}$, and $SL(n+$

$1, \mathbb{R})$ once again gives all the projective transformations. The only slight complication is that if n is odd, then $SL(n+1, \mathbb{R})$ does not act faithfully since it contains the central symmetry; in this case the group of transformations is $PSL(n+1, \mathbb{R})$.

Notice that again projective space can be viewed as $\mathbb{R}^n$ with added $(n-1)$-dimensional hyperplane at infinity so Theorem 14.3 extends to higher dimensions.

d. The Riemann sphere and conformal geometry. An element of the group $SL(2, \mathbb{C})$ is specified by three of the (complex) entries of the matrix; the fourth is determined by these three. Thus the group has three (complex) dimensions, and as a real object is 6-dimensional.

To characterise fractional linear transformations of the Riemann sphere, we use the fact that these transformations (also called *Möbius transformations*) preserve the cross-ratio, together with the following exercise.

EXERCISE 14.3. Show that four points $z_1, z_2, z_3, z_4 \in \mathbb{C} \cup \{\infty\}$ lie on a single circle or line (where ∞ lies on every line) if and only if $(z_1, z_2; z_3, z_4) \in \mathbb{R}$.

The rather vague statement that "∞ lies on every line" may be compared with the (equally vague) statement that "lines are just circles that pass through ∞", which is illustrated in Figure 3.3 using stereographic projection. Or one can notice that the transformation $z \to 1/(z - w_0)$ takes all lines not passing through w_0 into bounded objects that hence must be circles.. We will use this fact soon.

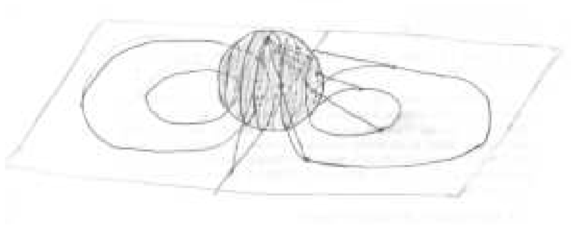


FIGURE 3.3. Circles through ∞ are lines.

THEOREM 14.4. *A bijection on the Riemann sphere is a fractional linear transformation (or the composition of such a transformation with the map $z \mapsto \overline{z}$ if and only if it maps lines and circles to lines and circles.*

PROOF. That any such transformation maps lines and circles to lines and circles follows from Proposition 13.3 (fractional linear transformations preserve cross-ratios) and Exercise 14.3.

Given such a bijection T, let $w_0 = T\infty$, and suppose $w_0 \in \mathbb{C}$ (that is, $w_0 \neq \infty$). Then the fractional linear transformation $F_1 \colon z \mapsto 1/(z - w_0)$ maps w_0 to ∞, and we see that $F_2 = F_1 \circ T$ is a bijection of the Riemann sphere which maps lines and circles to lines and circles, and which fixes ∞. Because it fixes ∞, it does not map any lines to circles or circles to lines; consequently, the restriction of F_2 to $\mathbb{R}^2$ is a bijection of the plane which maps lines to lines and circles to circles. We emphasise that this is a stronger property than what was originally assumed, since now lines and circles cannot be interchanged by the action of the map.

Since F_2 maps lines to lines, it must be affine by Theorem 14.1. Thus there exist $A \in GL(2, \mathbb{R})$ and $\mathbf{b} \in \mathbb{R}^2$ such that for every $\mathbf{x} \in \mathbb{R}^2 = \mathbb{C}$, we have $F_2\mathbf{x} = A\mathbf{x} + \mathbf{b}$. Since F_2 maps circles to circles, there exists $r \in \mathbb{R}$ such that $r^{-1}A \in O(2)$, otherwise F_2 would map the unit circle to an ellipse.

By composing F_2 with the map $z \mapsto \overline{z}$ if necessary, we may assume that $r^{-1}A \in SO(2)$, and hence A acts on $\mathbf{x} \in \mathbb{R}^2$ as rotation by some angle θ followed by dilation by r. Let $w_1 = re^{i\theta} \in \mathbb{C}$ and observe that identifying $\mathbf{x} \in \mathbb{R}^2$ with $z \in \mathbb{C}$, the action of A has the same effect as multiplication by w_1. Let $w_2 \in \mathbb{C}$ correspond to $\mathbf{b} \in \mathbb{R}^2$; then we see that F_2 acts on $\mathbb{C}$ as $F_2(z) = w_1 z + w_2$. In particular, F_2 is a fractional linear transformation, and hence $T = F_1^{-1} \circ F_2$ is a fractional linear transformation as well. □

The group of isometries is generated by reflections in lines; a similar result holds for transformations of the Riemann sphere using reflections in circles. More precisely, given a circle C of radius r centred at $\mathbf{x}$, the *circle inversion* in C is the map from $\mathbb{R}^2 \cup \{\infty\}$ to itself which takes the point $\mathbf{y} \in \mathbb{R}^2$ to the point $\mathbf{y}'$ such that:

(1) $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{y}'$ are collinear;
(2) $d(\mathbf{x}, \mathbf{y}) \cdot d(\mathbf{x}, \mathbf{y}') = r^2$.

If the circle of inversion converges to a circle through infinity—that is, a line—then the inversion converges to reflection in that line. This geometric characterization of fractions-linear and anti-fractional linear transformations (i.e. their compositions with inversions) extends to higher dimensions where no complex structure is available. Notice that the Riemann sphere generalizes striaghtforwardly to $\mathbb{R}^n \cup \{\infty\}$, that is indeed a sphere via stereographic projection, and has nothing to do with higher-dimensional projective spaces. The coincidence in real dimension two (and complex dimension one) is another instance of the low-dimensional phenomena. This construction is also a basis for a higher-dimensional generalization of the hyperbolic plane.

Exercise 14.4. Decompose the similarity transformation $\mathbf{x} \mapsto \lambda\mathbf{x}$ as a composition of circle inversions.

Remark. Any bijection of $\mathbb{R}^2$ which preserves angles is a similarity transformation (that is, an isometry composed with a homothety $\mathbf{x} \mapsto \lambda\mathbf{x}$). This is *not* true locally; given a domain $D \subset \mathbb{R}^2$, there are generally many holomorphic (hence conformal) maps from D to itself which are not fractional linear transformations. Writing the Taylor series expansions of these maps, we see that they are determined by infinitely many parameters, unlike the transformation groups we have seen so far, which only require finitely many parameters to specify. However, it follows from results of complex analysis that an entire bijection (that is, a holomorphic bijection from $\mathbb{C}$ to $\mathbb{C}$) is linear, and hence a similarity transformation.

An amazing fact is that in higher dimension preservation of angles forces transformations to be Möbius (up to an inversion). This is a first manifestation of a remarkable series of *rigidity* phenomena in higher-dimensional geometry and group theory.

Lecture 15. Friday, October 9

a. Comments on Theorem 14.1. A few remarks on the proof of Theorem 14.1 are in order. The proof of the theorem goes more or less as follows:

(1) Suppose that $F\colon \mathbb{R}^2 \to \mathbb{R}^2$ is a bijection which maps lines to lines, and without loss of generality, assume that F fixes the three points $(0, 0)$, $(1, 0)$, and $(0, 1)$, by composing F with an affine transformation if necessary.
(2) Use the fact that F respects parallel lines to show that $F(x, y) = (f(x), f(y))$ for some map $f\colon \mathbb{R} \to \mathbb{R}$.
(3) Show that $f(x + y) = f(x) + f(y)$ and $f(xy) = f(x)f(y)$—that is, f is a *field automorphism.*
(4) Show that the only field automorphism of $\mathbb{R}$ is the identity map.

One can consider the affine geometry of other fields besides $\mathbb{R}$. To do this, we emphasise not the geometric definition of a line in $\mathbb{R}^2$, but rather the *algebraic* definition as the set of points which satisfy a linear equation $ax + by = c$. In this way we can define "lines" in F^2 for an arbitrary field F, and then consider all bijections from F^2 to itself which map lines to lines.

This is the idea behind *algebraic geometry*: translate geometric notions into algebraic language, so that they can be worked with in the most general setting possible. The first two steps of the above proof go through in this more general setting; however, other fields may have non-trivial automorphisms. For example, the field $\mathbb{C}$ has the non-trivial automorphism $z \mapsto \overline{z}$. Thus for such fields, the analogue of Theorem 14.1 states that every bijection of F^2 which takes lines to lines has the form $\mathbf{x} \mapsto A\Phi(\mathbf{x}) + \mathbf{b}$, where $\mathbf{b} \in F^2$ and $A \in GL(2, F)$ are fixed, and $\Phi(x_1, x_2) = (\phi(x_1), \phi(x_2))$ for some automorphism ϕ of F.

We also point out that the difficult part of the proof of Theorem 14.1 is going from the statement that f is additive ($f(x + y) = f(x) + f(y)$) to the statement that f is linear, which also requires that $f(\lambda x) = \lambda f(x)$ for all $\lambda \in \mathbb{R}$. This implication does not hold in general; viewing $\mathbb{R}$ as a vector space over $\mathbb{Q}$ (of uncountably infinite dimension) and using the fact that every vector space has a basis (which is a consequence of the Axiom of Choice), one can construct an additive map of $\mathbb{R}$ which is not linear.

However, the implication *does* hold under any one of a number of relatively small additional assumptions; for example, one needs only to require that f is continuous at a single point, or is monotonic, or is measurable, etc. If the real line is considered only as a field (a purely algebraic object), none of these conditions can be formulated correctly, as they require additional structure on $\mathbb{R}$—a topology, an order, a σ-algebra, etc.

b. Hyperbolic geometry. Recall that $SL(2,\mathbb{C})$ acts on the Riemann sphere $\mathbb{C}\cup\{\infty\}$ via fractional linear transformations. Suppose that $a,b,c,d \in \mathbb{C}$ are such that the fractional linear transformation $f(z) = \frac{az+b}{cz+d}$ preserves the real line (together with ∞)—that is, $f(x) \in \mathbb{R}\cup\{\infty\}$ for all $x \in \mathbb{R}\cup\{\infty\}$. Then in particular, we have

$$f(0) = \frac{b}{d} \in \mathbb{R}\cup\{\infty\}, \qquad f(\infty) = \frac{a}{c} \in \mathbb{R}\cup\{\infty\}, \qquad f(1) = \frac{a+b}{c+d} \in \mathbb{R}\cup\{\infty\}.$$

Writing these three quantities as $\lambda_1, \lambda_2, \lambda_3$, respectively, we have $b = \lambda_1 d$, $a = \lambda_2 c$, and hence

$$a + b = \lambda_1 d + \lambda_2 c = \lambda_3 d + \lambda_3 c.$$

Rearranging, we obtain

$$(\lambda_1 - \lambda_3)d = (\lambda_3 - \lambda_2)c,$$

which together with the above equalities implies that

$$a = \lambda_2 c = \lambda_2 \frac{\lambda_1 - \lambda_3}{\lambda_3 - \lambda_2} d = \frac{\lambda_2}{\lambda_1} \frac{\lambda_1 - \lambda_3}{\lambda_3 - \lambda_2} b.$$

Thus writing $w = a/|a|$ and $a' = a/w$, $b' = b/w$, $c' = c/w$, and $d' = d/w$, we obtain $a', b', c', d' \in \mathbb{R} \cup \{\infty\}$, and furthermore, we can compute the determinant and see that $1 = ad - bc = w^2(a'd' - b'c')$, whence $w \in \mathbb{R}$ and so $a, b, c, d \in \mathbb{R} \cup \{\infty\}$.

This shows that any fractional linear transformation of the Riemann sphere which preserves the real line is determined by a matrix in $SL(2,\mathbb{R})$, and then acts on the upper half-plane $\mathbb{H}^2$. Finally, we observe that an element of $SL(2,\mathbb{R})$ is determined by three parameters. Thus, if we consider two points $w, z \in \mathbb{H}^2$, we have only three degrees of freedom in selecting their images $f(w), f(z)$ under some Möbius transformation f. However, the collection of *all* pairs of points has four degrees of freedom; and so there must be some constraint satisfied by the pair $(f(w), f(z))$. This constraint is precisely the notion of distance in the hyperbolic plane, which is preserved by any Möbius transformation, and which we will address in the next lecture.

Lecture 16. Monday, October 12

a. Ideal objects. Before introducting the notion of distance on the hyperbolic plane $\mathbb{H}^2$, we pause to make a few remarks about "ideal" objects. In a number of the two-dimensional geometric objects we have studied so far, one is concerned not only with points in the plane, but also with points which lie "at infinity" in some sense. Such points may be referred to as *ideal*.

If we begin with the (complex) plane and add a single ideal point, we obtain the Riemann sphere $\mathbb{C}P(1) = \mathbb{C} \cup \{\infty\}$. If we take the (real) plane and instead add an ideal *line*, we get the real projective plane $\mathbb{R}P(2) = \mathbb{R}^2 \cup \{\mathbb{R}P(1)\}$—in this case there are an uncountable number of points at infinity. In both these cases, the points at infinity may be obtained as limits of sequences of points in the plane.

Now consider the hyperbolic plane

$$\mathbb{H}^2 = \{z \in \mathbb{C} \mid \operatorname{Im} z > 0\} = \{(x, y) \in \mathbb{R}^2 \mid y > 0\}.$$

A convergent sequence of points in $\mathbb{H}^2$ (where convergence is understood in the sense of $\mathbb{C}P(1)$) may converge to one of two places: to a real number $x \in \mathbb{R}$, or to ∞. The set of points $\mathbb{R} \cup \{\infty\}$ is called the *ideal boundary* of $\mathbb{H}^2$.

At this point we see an important distinction between $\mathbb{H}^2$ and the first two examples. In the first two examples, the ideal points are part of the geometric object ($\mathbb{C}P(1)$ or $\mathbb{R}P(2)$), and furthermore, there is no intrinsic difference between ideal points and "finite" points—that is, one can find a Möbius transformation which takes ∞ to an arbitrary point in $\mathbb{C} = \mathbb{C}P(1) \setminus \infty$, and one can find a projective transformation which takes $\mathbb{R}P(1)$ to any projective line in $\mathbb{R}P(2)$.

The situation with $\mathbb{H}^2$ is quite different: here the ideal boundary is *not* part of the hyperbolic plane, and is preserved by Möbius transformations with real coefficients. Thus while it is an important part of descriptions of the geometry of the hyperbolic plane, it is not part of $\mathbb{H}^2$ itself. We point out, however, that while the ideal boundary $\mathbb{R} \cup \{\infty\}$ is distinct from the hyperbolic plane, any two points *within* the ideal boundary are intrinsically equivalent. In particular, given $x \in \mathbb{R}$, there exists a Möbius transformation f of $\mathbb{H}^2$ such that $f(\infty) = x$; thus the point ∞ is just like any other point on the ideal boundary as far as hyperbolic geometry is concerned.

b. Hyperbolic distance. Recall the definition of the cross-ratio in (13.2), and recall the following results regarding the action of $PSL(2, \mathbb{C})$ on $\mathbb{H}^2$ via fractional linear transformations.

(1) Proposition 13.3: A map of $\mathbb{C}P(1)$ is a Möbius transformation if and only if it preserves the cross-ratio.
(2) Theorem 14.4: A bijection of $\mathbb{C}P(1)$ is a Möbius transformation if and only if it preserves lines and circles.

(3) Previous lecture: A Möbius transformation $f(z) = (az + b)/(cz + d)$ preserves $\mathbb{H}^2$ if and only if $a, b, c, d \in \mathbb{R}$.

Putting all these together, we see that the following conditions are equivalent for a map $f\colon \mathbb{H}^2 \to \mathbb{H}^2$:

(1) f is a fractional linear transformation.
(2) f preserves the cross-ratio.
(3) f maps lines and circles to lines and circles.

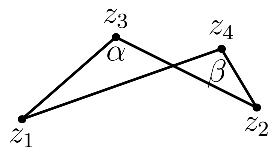


FIGURE 3.4. Interpreting the cross-ratio of four numbers.

We remark that Exercise 14.3, which is crucial to the above results, can be proved via the observation that the angles α and β in Figure 3.4 are the arguments of the complex numbers $\frac{z_2-z_3}{z_1-z_3}$ and $\frac{z_2-z_4}{z_1-z_4}$, respectively. Thus the argument of the cross-ratio $(z_1, z_2; z_3, z_4)$ given in (13.2) is exactly $\alpha - \beta$ (if the points are arranged as shown) or $\alpha + \beta$ (if z_4 lies on the other side of the line through z_1 and z_2), and the condition that the four points lie on a circle or a line is exactly the condition that this quantity be a multiple of π, which is thus equivalent to the cross-ratio being real.

Möbius transformations also have the important property of being *conformal*—that is, they preserve angles. This can be proved synthetically, by observing that every Möbius transformation is a composition of linear maps $z \mapsto az + b$ ($a, b \in \mathbb{C}$) with inversions $z \mapsto 1/z$, and that each of these maps preserves angles—this is immediate for linear maps, and for the inversion follows from the fact that circle inversion ($z \mapsto 1/\overline{z}$) and reflection in the real axis ($z \mapsto \overline{z}$) both preserve angles.

An analytic proof, which has the virtue of being given by a local argument, uses the fact that Möbius transformations are holomorphic, and thus conformal. Indeed, holomorphic functions are precisely those that can be locally approximated by linear functions: $f(z + w) = f(z) + L_z(w) + o(w)$, where L_z is a linear map which preserves angles, and $o(w)$ is of higher order.

We have now collected several significant features of the action of $PSL(2, \mathbb{R})$ on $\mathbb{H}^2$—in particular, we have four things that are preserved by this action, namely the cross-ratio, the ideal boundary, the collection of lines and circles, and the angle between any two curves. The definition of distance uses all four.

Given two points $z_1, z_2 \in \mathbb{H}^2$, we want to define a distance. This distance should be preserved by the action of $PSL(2, \mathbb{R})$. The two preserved *quantities* are the cross-ratio and angles. Of these two, the angle between any two curves is bounded, while $\mathbb{H}^2$ is unbounded, and so we expect our distance

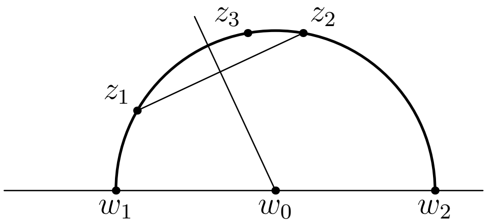


FIGURE 3.5. Defining distance in $\mathbb{H}^2$.

function to be unbounded. Thus we suspect that the distance between z_1 and z_2 ought to be defined in terms of the cross-ratio.

But the cross-ratio is defined in terms of *four* points, and we only have two! The other two points w_1 and w_2 will be chosen to lie on the ideal boundary, as shown in Figure 3.5. In particular, we choose $w_1, w_2 \in \mathbb{R}$ such that the four points w_1, w_2, z_1, z_2 all lie on the same circle.

Actually, we require slightly more than that. There are certain circles in $\mathbb{C} \cup \{\infty\}$ which are distinguished under the action of $PSL(2, \mathbb{R})$—namely, circles whose centre lies on the (extended) real line $\mathbb{R} \cup \{\infty\}$, and which intersect $\mathbb{R}$ orthogonally. If the centre is at ∞, such a circle in $\mathbb{C} \cup \{\infty\}$ is just a vertical line in $\mathbb{C}$ (parallel to the imaginary axis).

Since fractional linear transformations preserve angles and the (extended) real line, they also preserve the collection of lines and circles which intersect $\mathbb{R}$ orthogonally. Let us denote this class of curves in $\mathbb{H}^2$ by $\mathcal{G}$; the curves in $\mathcal{G}$ play a fundamental role in hyperbolic geometry. Given any two points $z_1, z_2 \in \mathbb{H}^2$, there exists a unique curve $\gamma \in \mathcal{G}$ which passes through z_1 and z_2. If z_1 and z_2 have the same real part, γ is just the vertical line passing through them. If they do *not* have the same real part, γ is the semi-circle constructed in Figure 3.5: the perpendicular bisector of the line segment from z_1 to z_2 comprises all centres of circles containing both z_1 and z_2, and it intersects the real line in a unique point w_0. The circle centred at w_0 that contains z_1 and z_2 lies in $\mathcal{G}$.

To define the distance between z_1 and z_2, we let w_1 and w_2 be the endpoints (on the ideal boundary $\mathbb{R} \cup \{\infty\}$) of the curve $\gamma \in \mathcal{G}$ that contains z_1 and z_2. Then the distance is given in terms of the cross-ratio

$$(z_1, z_2; w_1, w_2) = \frac{z_1 - w_1}{z_2 - w_1} \div \frac{z_1 - w_2}{z_2 - w_2}. \tag{16.1}$$

The distance cannot be the cross-ratio itself, for the following reason. A true distance function should be additive, in the sense that given a point z_3 lying "between" z_1 and z_2 (where in this case "between" means on the curve $\gamma \in \mathcal{G}$), we have $d(z_1, z_2) = d(z_1, z_3) + d(z_3, z_2)$. However, it may easily be seen from (16.1) that the cross-ratio is multiplicative:

$$(z_1, z_2; w_1, w_2) = (z_1, z_3; w_1, w_2)(z_3, z_2; w_1, w_2). \tag{16.2}$$

Thus in order to obtain a quantity which behaves like a distance function, we must take the *logarithm* of the cross-ratio, obtaining

$$d(z_1, z_2) = \log |(z_1, z_2; w_1, w_2)| = \left|\log \frac{(z_1 - w_1)(z_2 - w_2)}{(z_1 - w_2)(z_2 - w_1)}\right|. \tag{16.3}$$

If z_1 and z_2 have the same real part, then we have $z_1 = x + iy_1$, $z_2 = x + iy_2$, $w_1 = x$, $x_2 = \infty$, and obtain the following special case of (16.3):

$$d(x + iy_1, x + iy_2) = \left|\log \left(\frac{y_2}{y_1}\right)\right|. \tag{16.4}$$

We take the absolute value in (16.3) and (16.4) so that distance remains non-negative when z_1 and z_2 are interchanged.

Of course, we have not yet given a proof that the quantity defined in (16.3) satisfies the axioms of a metric. In particular, it is not obvious why the triangle inequality holds. One can prove the triangle inequality by first defining the *length* of a curve $\gamma\colon [0, 1] \to \mathbb{H}^2$ using d, and then showing that the lines and circles in $\mathcal{G}$ are precisely the *geodesics*—that is, curves of minimal length between z_1 and z_2. However, we shall not use this fact.

Because fractional linear transformations preserve the cross-ratio and map lines and circles in $\mathcal{G}$ to lines and circles in $\mathcal{G}$, they preserve the hyperbolic distance $d(z_1, z_2)$. In particular, d is invariant under the following maps:

(1) Horizontal translations $z \mapsto z + x$, where $x \in \mathbb{R}$.
(2) Homotheties $z \mapsto \lambda z$, where $\lambda \in (0, \infty)$.
(3) Circle inversions such as $z \mapsto 1/\overline{z}$.

We can use the first two of these to illustrate the similarities and differences between the hyperbolic metric on $\mathbb{H}^2$ and the familiar Euclidean metric. Consider the sequence of points $z_n = n + i$; in the Euclidean metric, the distance between any two successive points is 1. In the hyperbolic metric, we see that $d(z_n, z_{n+1}) = d(z_0, z_1)$ for every n, since horizontal translation does not change the hyperbolic distance. Thus although the hyperbolic distance between two successive points is not the same as the Euclidean distance, the sequence still has the property that the distance between two successive points is constant.

Now consider the sequence of points $z_n = ie^{-n}$. In the Euclidean metric, the points z_n converge to 0, and so the distance between them goes to 0. In the hyperbolic metric, however, we have

$$d(z_n, z_{n+1}) = \left|\log \left(\frac{e^{-n}}{e^{-(n+1)}}\right)\right| = 1$$

for every n, and so successive points are always a distance 1 apart. This illustrates the fact that the hyperbolic metric distorts the usual one by a factor of $1/y$ at the point $z = x + iy$, so the distortion becomes more and more pronounced as we approach the real line, which is the ideal boundary.

c. Isometries of the hyperbolic plane. We have now seen three two-dimensional metric geometries: the elliptic plane, the Euclidean plane, and the hyperbolic plane. In the first of these, every even isometry was a rotation, with a single fixed point (or in the case of rotation by π, a fixed point and a fixed line). In the second, we had two possibilities, contingent on the presence of a fixed point: every even isometry is either a translation or a rotation.

What happens in $\mathbb{H}^2$? Do we have a similar classification? Are there new classes of isometries? To answer this, we use the algebraic description of even isometries of $\mathbb{H}^2$ as fractional linear transformations, which gives us an efficient way of determining whether or not an isometry has fixed points, and where they lie.

Consider an isometry $f(z) = (az + b)/(cz + d)$, where $ad - bc = 1$, and $a, b, c, d \in \mathbb{R}$. A point $z \in \mathbb{C}$ is fixed by f if and only if $z = (az+b)/(cz+d)$—that is, if and only if

$$cz^2 + dz = az + b.$$

Collecting all the terms on one side and using the quadratic formula, we obtain

$$\begin{aligned} z &= \frac{a - d \pm \sqrt{(a-d)^2 + 4bc}}{2c} \\ &= \frac{a - d \pm \sqrt{(a+d)^2 - 4ad + 4bc}}{2c} \\ &= \frac{a - d \pm \sqrt{(a+d)^2 - 4}}{2c}, \end{aligned} \tag{16.5}$$

where the last equality uses the fact that $ad - bc = 1$. Thus we see that the number and type of the fixed points of f is determined by the absolute value of $a + d$, which is the trace of $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$, and hence is invariant under conjugacy. There are three possibilities.

Hyperbolic transformations. If $|a + d| > 2$, then f has two fixed points on the ideal boundary. If $c \neq 0$, these are given by (16.5); if $c = 0$, one fixed point is $b/(a - d)$, and the other is ∞. Denote the fixed points by w_1 and w_2, and let γ be any circle (or line) in $\mathbb{H}^2$ (not necessarily a circle or line in $\mathcal{G}$) with endpoints w_1 and w_2. Every point in $\mathbb{H}^2$ lies on exactly one such curve, and each such curve is preserved by f, which is in some sense a "translation" along γ. Thus hyperbolic transformations are as close counterparts of translations in the Euclidean plane as one can get: such a transformation has an invariant line (connecting two fixed points at infinity), the "axis", and the "translation vector" of a particular length along this line. However, since the notion of parallelism in hyperbolic geometry is quite different from the Euclidean geometry one cannot say that any point moves parallel to the axis. Instead of the family of invarant parallel lines there is a family of *equidistant curves* represented by arcs of circles connecting the fixed points at infinity. Such a curve is indeed a locus of points lying on

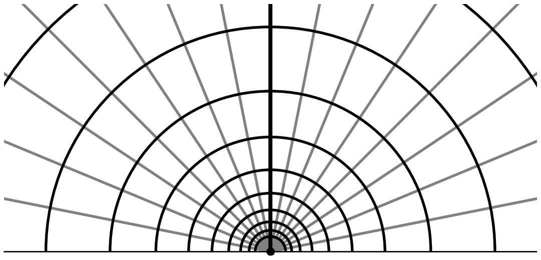


FIGURE 3.6. A homothety acting on circles centred at the origin.

one side of the axis and at a fixed distance from it. but unlike the case of Euclidean geometry it is not a line itself.

The easiest model of a hyperbolic transformation to visualise is the one corresponding to the matrix $\left(\begin{smallmatrix} \lambda & 0 \\ 0 & \lambda^{-1} \end{smallmatrix}\right)$, which acts on $\mathbb{H}^2$ as $z \mapsto \lambda^2 z$ (see Figure 3.6).

Parabolic transformations. If $|a+d| = 2$, then f has exactly one fixed point on the ideal boundary. The easiest model of a parabolic transformation comes when we take this point to be ∞, and consider the map $z \mapsto z+1$.

Elliptic transformations. If $|a+d| < 2$, then f has two fixed points in $\mathbb{C}$, which are conjugate and non-real. One of these lies in the upper half-plane, the other in the lower half-plane. Thus f has exactly one fixed point in $\mathbb{H}^2$, and acts as "rotation" around this point in a certain sense. Using the upper half-plane model of hyperbolic geometry, the image of f as a rotation appears distorted; however, there is another model of hyperbolic geometry, the unit disc, in which f becomes a genuine rotation in the Euclidean sense.

To describe the conjugacy classes in $\mathrm{Isom}^+(\mathbb{H}^2)$, we need to give the conjugacy invariants within each class of isometries. Recall that $\mathrm{Isom}^+(\mathbb{H}^2)$ is isomorphic to $PSL(2,\mathbb{R})$, and that the conjugacy invariants in the latter group are the eigenvalues of $A \in PSL(2,\mathbb{R})$. If the eigenvalues are real (λ and $1/\lambda$), then f_A is a hyperbolic transformation; if they are complex ($e^{i\theta}$ and $e^{-i\theta}$), then f_A is an elliptic transformation. If they are unity ($\lambda_1 = \lambda_2 = 1$), then f_A is either the identity (if $A = \mathrm{Id}$) or a parabolic transformation. Furthermore, all matrices of the form $\left(\begin{smallmatrix} 1 & t \\ 0 & 1 \end{smallmatrix}\right)$ with $t \neq 0$ are conjugate, and so all parabolic transformations are conjugate.

EXERCISE 16.1. Express the hyperbolic length of translation along the axis of a hyperbolic transformation through the eigenvalues of the matrix.

Lecture 17. Wednesday, October 14

a. Matrix groups. A *linear group* is an abstract group which admits a faithful representation as a matrix group—that is, a subgroup of $GL(n, \mathbb{R})$ for some n. We have already seen a number of important examples of linear groups, such as the six groups in Table 1 (Lecture 14). Each of these groups corresponded to a particular sort of two-dimensional geometry; for the time being, we will set immediate geometric applications aside and broaden our horizons by considering other linear groups.

Recall that an $n \times n$ matrix A is *upper-triangular* if $A_{ij} = 0$ whenever $i > j$. The determinant of an upper-triangular matrix is the product of its diagonal entries A_{ii}, which are the eigenvalues of A; in particular, A is invertible if and only if all its diagonal entries are non-zero. Let $UT(n)$ denote the group of all invertible upper-triangular $n \times n$ matrices with real entries.

Given $A \in UT(n)$, let D be the diagonal matrix $D = \text{diag}(A_{11}, \dots, A_{nn})$, and let N be the matrix $N = A - D$. Observe that $N_{ij} = 0$ for all $i \geq j$; it follows from the formula for matrix multiplication that $(N^k)_{ij} = 0$ for all $i > j - k$. More visually, one may say that the main diagonal of N contains only zeros, that the diagonal above the main diagonal of N^2 contains only zeros, and in general, that every diagonal less than k spaces above the main diagonal of N^k contains only zeros. In particular, $N^n = 0$; a matrix N with this property is called *nilpotent*.

The upshot of all this is that any upper-triangular matrix A can be written as

$$A = D + N, \tag{17.1}$$

where D is diagonal and N is nilpotent. Of course, the binary operation in the group $UT(n)$ is multiplication, not addition; by decomposing A as the sum of two matrices with special forms and properties, we are making use of the fact that matrix groups have an additive structure which is in some sense extrinsic to the group structure.

To illustrate why this additive structure is not captured by the group structure, we observe that in the first place, $UT(n)$ is not closed under addition (the sum of two invertible matrices may not be invertible), and in the second place, the nilpotent matrices used in the above decomposition do not actually lie in $UT(n)$, as they are not invertible.

The analogues of the nilpotent matrices within the group $UT(n)$ are the *unipotent* matrices—that is, matrices of the form $I + N$, where N is nilpotent. Given two unipotent matrices $I + N$ and $I + N'$, we observe that

$$(I + N)(I + N') = I + N + N' + NN'$$

is unipotent as well, which follows from the fact that

$$(N + N' + NN')_{ij} = N_{ij} + (N')_{ij} + \sum_{k=1}^{n} N_{ik}N'_{kj} = 0$$

for all $i \geq j$. Furthermore, recalling the geometric series

$$(1+x)^{-1} = 1 - x + x^2 - x^3 + x^4 - \cdots$$

and using the fact that $N^k = 0$ for all $k \geq n$, we see that

$$(I+N)^{-1} = I - N + N^2 - N^3 + \cdots + (-1)^{n-1}N^{n-1} \tag{17.2}$$

is again a unipotent matrix. It follows that the set of unipotent matrices forms a subgroup of $UT(n)$, which we denote $\mathcal{U}_n$. Writing $\mathcal{D}_n$ for the subgroup of all diagonal matrices, we see that (17.1) leads to the group decomposition

$$UT(n) = \mathcal{D}_n\mathcal{U}_n, \tag{17.3}$$

since writing $N' = D^{-1}N$ gives $D + N = D(I + N')$.

b. Algebraic structure of the unipotent group. We now investigate the algebraic properties of the group $UT(n)$; we will do this by stating a general result (for all n), and then doing the calculations in the specific case $n = 3$, which is representative of the general case but easier on the eyes notationally.

First we need a little more notation. Writing $\mathcal{N}_n$ for the collection of all $n \times n$ upper-triangular nilpotent matrices, we consider the following classes of nilpotent matrices for $1 \leq k \leq n$:

$$\mathcal{N}_n^k = \{N \in \mathcal{N}_n \mid N_{ij} = 0 \text{ for all } j < i + k\}. \tag{17.4}$$

That is, $\mathcal{N}_n^k$ is the set of all upper-triangular nilpotent matrices with k empty diagonals (including the main diagonal); equivalently, $\mathcal{N}_n^k = \{N \in UT(n) \mid N^{n+1-k} = 0\}$. We see that

$$\mathcal{N}_n = \mathcal{N}_n^1 \supset \mathcal{N}_n^2 \supset \cdots \supset \mathcal{N}_n^{n-1} \supset \mathcal{N}_n^n = \{0\}.$$

Given $N \in \mathcal{N}_n^k$ and $N' \in \mathcal{N}_n^{k'}$, we have

$$(NN')_{ij} = \sum_{m=1}^{n} N_{im}N'_{mj},$$

and we see that the only non-vanishing terms are those for which $m \geq i + k$ and $j \geq m + k'$. In particular, $(NN')_{ij} = 0$ unless there exists m such that $j \geq m + k' \geq i + k + k'$, and so we have $(NN')_{ij} = 0$ for all $j < i + k + k'$, whence $NN' \in \mathcal{N}_n^{k+k'}$. Thus the sets $\mathcal{N}_n^k$ have the following property:

$$\mathcal{N}_n^k \cdot \mathcal{N}_n^{k'} \subset \mathcal{N}_n^{k+k'}. \tag{17.5}$$

Let $\mathcal{U}_n^k$ be the set of all unipotent matrices of the form $I+N$, where $N \in \mathcal{N}_n^k$. Equivalently

$$\mathcal{U}_n^k = \{A \in UT(n) \mid (A - I)^{n+1-k} = 0\}.$$

It follows from (17.2) and (17.5) that $\mathcal{U}_n^k$ is a subgroup of $UT(n)$, and we have

$$\mathcal{U}_n = \mathcal{U}_n^1 \supset \mathcal{U}_n^2 \supset \cdots \supset \mathcal{U}_n^{n-1} \supset \mathcal{U}_n^n = \{I\}.$$

Visually, $\mathcal{U}_n^k$ is the set of all unipotent matrices with at least $k-1$ blank diagonals above the main diagonal.

Proposition 17.1. *The commutator subgroup of the group of upper-triangular matrices is the subgroup of unipotent matrices:*

$$[UT(n), UT(n)] = \mathcal{U}_n. \tag{17.6}$$

Furthermore, for every $1 \leq k < n$, we have

$$[\mathcal{U}_n, \mathcal{U}_n^k] = \mathcal{U}_n^{k+1}. \tag{17.7}$$

Thus $\mathcal{U}_n$ is a nilpotent group, and $UT(n)$ is a solvable group.

Proof of (17.6). Let $\varphi \colon UT(n) \to \mathcal{D}_n$ be the map which takes $A \in UT(n)$ to the diagonal matrix $\operatorname{diag}(A_{11}, \dots, A_{nn})$, and observe that φ is a homomorphism, since for upper-triangular matrices A and B we have $(AB)_{ii} = A_{ii}B_{ii}$. It follows that $\varphi([A,B]) = \varphi(ABA^{-1}B^{-1}) = I$, and hence $[A,B] \in \ker \varphi = \mathcal{U}_n$ for every $A, B \in UT(n)$, which establishes the inclusion $[UT(n), UT(n)] \subset \mathcal{U}_n$.

The proof that every unipotent matrix can be obtained as the commutator of two upper-triangular matrices is left as an exercise. □

Remark. Thanks to Proposition 17.1, we can construct many nontrivial examples of nilpotent groups. It should be observed that the word "nilpotent" may be used to describe a group or a matrix, and that the meaning in the two cases is somewhat different, although the two are certainly related.

c. The Heisenberg group. We prove the second half of Proposition 17.1 in the case $n = 3$; the same arguments go through in the more general setting. The group $\mathcal{U}_3$ is called the *Heisenberg group*; we have

$$\mathcal{U}_3 = \left\{ \begin{pmatrix} 1 & x & z \\ 0 & 1 & y \\ 0 & 0 & 1 \end{pmatrix} \mid x, y, z \in \mathbb{R} \right\}.$$

The Heisenberg group $\mathcal{U}_3$ joins $\operatorname{Isom}(\mathbb{R}^2)$, $SO(3)$, and $SL(2,\mathbb{R})$ on our list of interesting three-dimensional groups. Unlike its counterpart $\mathcal{U}_2$ (which is isomorphic to the real numbers with addition), it is non-abelian, but the non-commutativity enters in a relatively simple way. To wit, we see that

$$\begin{pmatrix} 1 & x_1 & z_1 \\ 0 & 1 & y_1 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & x_2 & z_2 \\ 0 & 1 & y_2 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & x_1 + x_2 & z_1 + z_2 + x_1 y_2 \\ 0 & 1 & y_1 + y_2 \\ 0 & 0 & 1 \end{pmatrix}, \tag{17.8}$$

and the only term that gets in the way of commutativity is $x_1 y_2$.

To compute the inverse of an element $I + N \in \mathcal{U}_3$, we use (17.2):

$$\begin{aligned}(I+N)^{-1} &= I - N + N^2 \\ &= \begin{pmatrix}1 & -x & -z \\ 0 & 1 & -y \\ 0 & 0 & 1\end{pmatrix} + \begin{pmatrix}0 & 0 & xy \\ 0 & 0 & 0 \\ 0 & 0 & 0\end{pmatrix} \\ &= \begin{pmatrix}1 & -x & -z+xy \\ 0 & 1 & -y \\ 0 & 0 & 1\end{pmatrix}.\end{aligned}$$

Proof of (17.7) for $n = 3$. We could show that $[\mathcal{U}_3, \mathcal{U}_3] = \mathcal{U}_3^2$ by directly computing the entries of $[I+N, I+N']$. Instead, we opt to once again use (17.2), and expand the following expression:

$$[I+N, I+N'] = (I+N)(I+N')(I-N+N^2)(I-N'+N'^2).$$

We get a polynomial in N and N'; using (17.5), we see that every term of cubic or higher order vanishes, and so

$$\begin{aligned}&= (I+N+N'+NN')(I-N+N^2-N'+NN'+N'^2) \\ &= (I+N+N'+NN') - (N+N^2+N'N) + N^2 \\ &\qquad - (N'+NN'+N'^2) + NN' + N'^2 \\ &= I + NN' - N'N.\end{aligned} \tag{17.9}$$

It follows that $[\mathcal{U}_3, \mathcal{U}_3] \subset \mathcal{U}_3^2$. Once again, the reader may verify that we in fact have equality.

To see that $[\mathcal{U}_3, \mathcal{U}_3^2] = \mathcal{U}_3^3 = \{I\}$, it suffices to observe that $\mathcal{U}_3^2$ is the centre of $\mathcal{U}_3$. This follows from either (17.8) or (17.9), by observing that $NN' - N'N$ vanishes for every N' if and only if $N^2 = 0$. □

The general procedure for $n > 3$ is similar: one shows that $[I+N, I+N'] = I + P(N, N')$, where P is a polynomial with no linear terms, which thus outputs nilpotent matrices of lower degree than its inputs.

Remark. The expression $NN' - N'N$ in (17.9) is also called a commutator, this time in the ring theoretic sense. This relates the ring structure of nilpotent matrices to the group structure of unipotent matrices.

Lecture 18. Friday, October 16

a. Groups of unipotent matrices. We claimed in the last lecture (Proposition 17.1) that the group of unipotent matrices $\mathcal{U}_n$ is nilpotent, which gives us our first general class of examples of non-abelian nilpotent groups. Before completing the proof in the general case, we observe that the converse statement is *not* true: there are plenty of nilpotent matrix groups which are not contained in $\mathcal{U}_n$. For example, the diagonal subgroup $\mathcal{D}_n$ is abelian, and hence nilpotent.

Proof of Proposition 17.1. To complete the proof, we continue to use the fact that nilpotent matrices have an additive structure alongside the multiplicative one.

Given $1 \leq i, j \leq n$, let $e_{ij} \in \mathcal{N}_n$ be the matrix which has a 1 in the ith row and jth column, and all other entries are 0. Observe that $\mathcal{N}_n$ is a real vector space of dimension $(n^2 - n)/2$ with basis $\{e_{ij}, 1 \leq i < j \leq n\}$.

A simple calculation shows that $e_{ij}e_{kl} = \delta_{jk}e_{il}$, where δ_{jk} is the Kronecker delta—it is equal to 1 if $j = k$, and 0 otherwise. The geometric meaning of all this is that if we write $\{\mathbf{e}_i\}$ for the standard basis vectors in $\mathbb{R}^n$, then e_{ij} acts on $\mathbb{R}^n$ as a linear operator, taking $\mathbf{e}_j$ to $\mathbf{e}_i$, and all other basis vectors to $\mathbf{0}$. In particular, the only way that e_{ij} and e_{kl} can have a non-zero product is if the range of e_{kl} (which is spanned by $\mathbf{e}_k$) is in the complement of the kernel of e_{ij} (this complement is spanned by $\mathbf{e}_j$).

For the corresponding unipotent matrices, we have the following multiplication rule:

$$(I + se_{ij})(I + te_{kl}) = I + se_{ij} + te_{kl} + st\delta_{jk}e_{il}. \tag{18.1}$$

In particular, since $i < j$, we have

$$(I + e_{ij})(I - e_{ij}) = I + e_{ij} - e_{ij} = I,$$

and so $(I + e_{ij})^{-1} = I - e_{ij}$. Indeed, one has the more general formula

$$(I + e_{ij})^t = I + te_{ij}$$

for every $t \in \mathbb{Z}$.[7] Now assume that at least one of the inequalities holds: $i \neq \ell$ or $j \neq k$. The formula for the inverse lets us write the commutator of two basic unipotent matrices:

$$\begin{aligned}[I + e_{ij}, I + e_{kl}] &= (I + e_{ij})(I + e_{kl})(I - e_{ij})(I - e_{kl}) \\ &= (I + e_{ij} + e_{kl} + \delta_{jk}e_{il})(I - e_{ij} - e_{kl} + \delta_{jk}e_{il}) \\ &= (I + e_{ij} + e_{kl} + \delta_{jk}e_{il}) - (e_{ij} + \delta_{il}e_{kj}) \\ &\quad - (e_{kl} + \delta_{jk}e_{il}) + (\delta_{jk}e_{il}) \\ &= I + \delta_{jk}e_{il} - \delta_{il}e_{jk}.\end{aligned} \tag{18.2}$$

Thus we have three cases:

[7] Actually this holds for all $t \in \mathbb{R}$, but to make sense of it in the more general setting we need to say what is meant by A^t when A is a matrix and $t \notin \mathbb{Z}$.

(1) $j \neq k$ and $i \neq l$. In this case $I + e_{ij}$ and $I + e_{kl}$ commute.
(2) $i \neq \ell, j = k$. In this case $[I + e_{ij}, I + e_{kl}] = I + e_{il}$.
(3) $k \neq j, i = \ell$. In this case $[I + e_{ij}, I + e_{kl}] = I - e_{kj}$.[8]

Now assuming that $i < j$ and $k < \ell$ we see that these three cases cover all possibilities and in every case, the commutator lies in $\mathcal{U}_n^{(j-i)+(l-k)}$. This is the prototype for the result that

$$[\mathcal{U}_n^k, \mathcal{U}_n^{k'}] = \mathcal{U}_n^{k+k'}, \tag{18.3}$$

which is a stronger version of (17.7). To see (18.3), we fix $N \in \mathcal{N}_n^k$ and $N' \in \mathcal{N}_n^{k'}$, and then write for every $j \geq 0$

$$\sigma_j = N^j + N^{j-1}N' + N^{j-2}N'^2 + \cdots + NN'^{j-1} + N'^j.$$

Observe that $(I + N)(I + N') = I + \sigma_1 + NN'$, and that

$$\sigma_1 \sigma_j = \sigma_{j+1} + N'N\sigma_{j-1}. \tag{18.4}$$

Furthermore, we have

$$\begin{aligned}(I + N)^{-1}(I + N')^{-1} &= (I - N + N^2 - N^3 + \cdots)(I - N' + N'^2 - N'^3 + \cdots) \\ &= I - \sigma_1 + \sigma_2 - \sigma_3 + \cdots .\end{aligned}$$

This allows us to compute the commutator by applying (18.4):

$$\begin{aligned} &= (I + \sigma_1 + NN')(I - \sigma_1 + \sigma_2 - \sigma_3 + \cdots) \\ &= \sum_{j \geq 0} (-1)^j (I + \sigma_1 + NN')\sigma_j \\ &= \sum_{j \geq 0} (-1)^j (\sigma_j + \sigma_{j+1} + N'N\sigma_{j-1} + NN'\sigma_j \\ &= I + (NN' - N'N) \sum_{j \geq 0} (-1)^j \sigma_j. \end{aligned} \tag{18.5}$$

It follows from (17.5) that $[I + N, I + N'] \in \mathcal{N}_n^{k+k'}$, which establishes one inclusion in (18.3). The other inclusion follows from the fact that given $t \in \mathbb{R}$, $1 \leq i \leq n$ and $m = i + k, j = m + k'$, we have $I + te_{im} \in \mathcal{U}_n^k$, $I + e_{mj} \in \mathcal{U}_n^{k'}$, and

$$[I + te_{im}, I + e_{mj}] = I + te_{ij}.$$

Since every element of $\mathcal{U}_n^{k+k'}$ can be written as a product of matrices of the form $I + te_{ij}$ for $j = i + k + k'$, this establishes (18.3). □

The groups $\mathcal{U}_n^k$ are natural examples of non-abelian nilpotent groups. There are also other interesting subgroups of $\mathcal{U}_n$, which are automatically nilpotent.

[8] The reader can easily write the formula for $[I + e_{ij}, I + e_{ji}]$ that is more complicated and will not be used directly in these lectures.

Example 18.1. Consider the *generalised Heisenberg group*

$$H_n = \{I + N \mid N \in \mathcal{N}_n \text{ has } N_{ij} = 0 \text{ if } i > 1 \text{ and } j < n\},$$

which is the set of unipotent matrices whose non-zero entries are all in either the first row or the last column. Of particular importance are the matrices $h_k = I + e_{1k}$ and $h'_k = I + e_{kn}$, and also $c = I + e_{1n}$. One has that $[h_k, h'_k] = c$, and furthermore, $c \in Z(H_n)$. It follows that H_n is a nilpotent group of nilpotent length 2.

Why are we interested in nilpotent groups? What is the point of the seemingly arbitrary condition that the sequence G_n defined by $G_{n+1} = [G, G_n]$ eventually terminates in the trivial group? There are (at least) two answers to this question. In the first place, nilpotent groups are the next easiest target after abelian groups: since "groups" as a general class are far too complex to understand completely, we must proceed by studying particular classes of groups. Abelian groups are fairly well understood, but once we move into the non-abelian world, things can become quite difficult. Nilpotent groups are "close enough"' to being abelian that we can still prove powerful results. This is related to the second answer, which is that many groups which are very intricate and complicated can be meaningfully studied by understanding their nilpotent subgroups. This gives us a window into the internal structure of objects which might otherwise be inaccessible.

b. Matrix exponentials and relationship between multiplication and addition. The computations in the previous section indicate a relationship between multiplication of certain matrices (unipotent matrices) and addition of others (nilpotent matrices). This is made explicit in (18.5), which relates the group commutator $ABA^{-1}B^{-1}$ to the ring commutator $AB - BA$.

The most familiar object which relates multiplication and addition is the exponential map, and it turns out that we can apply this to matrices as well as to numbers. In particular, given a matrix A, we define the *matrix exponential* of A by the power series

$$e^A = \sum_{k=0}^{\infty} \frac{1}{k!} A^k. \tag{18.6}$$

Observe that this series is absolutely convergent since $\|A^k\| \leq \|A\|^k$, which grows much slower than $k!$.

Exercise 18.1. Show that the matrix exponential may be equivalently defined by

$$e^A = \lim_{k\to\infty} \left(I + \frac{A}{k} \right)^k. \tag{18.7}$$

The matrix exponential may also be defined using a differential equation, although it is not immediately obvious how this ought to be formulated.

The solution is to look for matrix functions $f(t) = e^{tA}$ and consider differentiation with respect to the parameter t. Using (18.6) one sees from a straightforward calculation that

$$\frac{df}{dt} = Af(t). \tag{18.8}$$

and in particular

$$\frac{df}{dt}_{|t=0} = A.$$

One can show that the equation 18.8 with initial condition $f'(0) = A$ has unique solution e^{tA}.

Observe that e^A may be efficiently computed using normal forms: if $A = TDT^{-1}$, where $T \in GL(n, \mathbb{C})$ and $D = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$, then it follows from (18.6) that

$$e^A = Te^D T^{-1} = T \operatorname{diag}(e^{\lambda_1}, \dots, e^{\lambda_n})T^{-1}.$$

Exercise 18.2. What happens if D is not diagonal, but contains a Jordan block?

If A and B commute, then the same proof as for real numbers shows that

$$e^{A+B} = e^A e^B. \tag{18.9}$$

Considering the scalar multiples tA of A, we have a one-parameter family of matrices

$$\varphi_A(t) = e^{tA},$$

where t is any real number. Since tA and sA commute for all $s, t \in \mathbb{R}$, it follows from (18.9) that

$$\varphi_A(t + s) = \varphi_A(t)\varphi_A(s). \tag{18.10}$$

Thus we see two things: in the first place, $\varphi_A(-t) = \varphi_A(t)^{-1}$, and so φ maps $\mathbb{R}$ into $GL(n, \mathbb{R})$, and in the second place, $\varphi\colon \mathbb{R} \to GL(n, \mathbb{R})$ is a homomorphism. The image $\{e^{tA} \mid t \in \mathbb{R}\}$ is a *one-parameter subgroup* of $GL(n, \mathbb{R})$.

As we mentioned above that $\frac{d}{dt}e^{tA} = Ae^{tA}$, and in particular, $\varphi'_A(0) = A$. This leads to the observation that *every* continuous one-parameter subgroup (that is, every homomorphic image of $\mathbb{R}$) in $GL(n, \mathbb{R})$ is obtained via a matrix exponential: given a homomorphism $\varphi\colon \mathbb{R} \to GL(n, \mathbb{R})$, we let $A = \varphi'(0)$, and observe that $\varphi(t) = e^{tA}$. One needs to prove of course that any continuous homomorphism is differentiable.

What happens if we consider *two* one-parameter subgroups of $GL(n,\mathbb{R})$? Given $A, B \in M_n(\mathbb{R})$, we see that

$$\begin{aligned} e^A e^B &= \left(I + A + \frac{1}{2}A^2 + \cdots\right)\left(I + B + \frac{1}{2}B^2 + \cdots\right) \\ &= (I + A + \frac{1}{2}A^2 + \cdots) + (B + AB + \cdots) + \frac{1}{2}B^2 + \cdots \\ &= I + A + B + \frac{1}{2}(A^2 + 2AB + B^2) + \cdots \\ &= I + (A+B) + \frac{1}{2}(A+B)^2 + \frac{1}{2}(AB - BA) + \cdots , \end{aligned}$$

where all omitted terms are of cubic or higher order. Thus if we consider matrices tA and tB, we have

$$e^{tA}e^{tB} = e^{t(A+B)} + \frac{t^2}{2}(AB - BA) + O(t^3). \tag{18.11}$$

This is the beginning of the *Campbell–Hausdorff* formula, which relates e^{A+B} to e^A and e^B when the matrices A and B do not commute. In terms of multi-variable calculus, this is the statement that moving along two non-commuting vector fields produces an error of quadratic order.

For the time being, we observe that for small values of t (and hence matrices tA and tB which are near the identity), the deviation from what we expect is controlled by the commutator $AB - BA$. In particular, (18.11) yields

$$[e^{tA}, e^{tB}] = I + t^2(AB - BA) + O(t^3),$$

which gives a concrete relationship between the group commutator (in terms of multiplication alone) and the ring commutator (in terms of multiplication and addition). This is the beginning of the theory of Lie groups and Lie algebras; we shall not dive into this theory at the present time, but will merely observe that the matrix exponential provides the relationship between nilpotent and unipotent matrices: if $N = e_{ij}$ is a basic nilpotent matrix, then it is straightforward to see that

$$e^N = I + N$$

is a basic unipotent matrix.

Lecture 19. Monday, October 19

a. Lie algebras. Let us make a few of the notions discussed in the previous lecture more precise. What we have been discussing leads naturally into a brief introduction to Lie groups and Lie algebras, which together are one of the central topics in modern mathematics. We will somewhat deviate from the general principle of these lectures (to prove everything stated, except exercises and statements proved by routine calculations, using only assumed standard background) and formulate a key result (Lemma 19.3) without a proof.

Before giving a general definition, we observe that every matrix group we have discussed so far is a Lie group. These objects can be quite complicated both to describe properly and to work with effectively; one reason for this is that the relevant operation, matrix multiplication, is non-commutative and gives rise to a great deal of internal structure. In dealing with the unipotent matrices $\mathcal{U}_n$, we were able to side-step this issue by using the additive structure of the set $\mathcal{N}_n$ of nilpotent matrices, and the fact that the two classes of matrices are naturally related.

This technique is in fact quite general; in the language we are about to introduce, $\mathcal{U}_n$ is the *Lie group*, and $\mathcal{N}_n$ is its associated *Lie algebra*. The link between a Lie algebra and a Lie group is given by the exponential map introduced in (18.6).

As described in the previous lecture, given an arbitrary $n \times n$ matrix A (which may or may not be invertible), the matrix exponential takes the matrices tA, $t \in \mathbb{R}$, to a one-parameter subgroup of $GL(n, \mathbb{R})$. That is, it takes a one-dimensional subspace of the vector space $M(n, \mathbb{R})$ to a one-parameter subgroup of the group $GL(n, \mathbb{R})$.

Remark. In fact, every *continuous* one-parameter subgroup of $GL(n, \mathbb{R})$ is obtained in this manner. If $\varphi\colon \mathbb{R} \to GL(n, \mathbb{R})$ is a homomorphism such that $\lim_{t\to 0} \varphi(t) = I$, then one may show that $A = \varphi'(0)$ exists, and that $\varphi(t) = e^{tA}$. The requirement of continuity is exactly analogous to the fact that an additive map from $\mathbb{R}$ to itself is linear (hence differentiable) as soon as we require continuity at a single point.

What does the matrix exponential do to a *two*-dimensional subspace of $M(n, \mathbb{R})$? Do we get a two-parameter subgroup of $GL(n, \mathbb{R})$? If the subspace V is spanned by two commuting matrices A and B, then (18.9) shows that we do in fact get a two-parameter subgroup, isomorphic to $\mathbb{R}^2$, since $e^{tA+sB} = e^{tA}e^{sB}$. If A and B do not commute, however, (18.11) shows that we should not expect $e^A e^B$ to lie in the subset $\{e^{tA+sB} \mid t, s \in \mathbb{R}\} \subset GL(n, \mathbb{R})$; consequently, we should not expect the image of V under the exponential map to be a subgroup of $GL(n, \mathbb{R})$.

So some subspaces of $M(n, \mathbb{R})$ exponentiate to subgroups of $GL(n, \mathbb{R})$, and others do not. How do we tell the difference? The presence of the expression $AB - BA$ in (18.11), and in the formula for $[e^{tA}, e^{sB}]$, suggests that this expression should play a role in whatever criterion we examine.

Indeed, this expression has intrinsic importance for the structure of $M(n,\mathbb{R})$, which is not just a vector space, but an *algebra*—that is, a vector space on which we have an associative multiplication that respects addition and scalar multiplication. The expression $AB - BA$ measures the degree to which A and B fail to commute.

Definition 19.1. Given two matrices $A, B \in M(n,\mathbb{R})$, the *Lie bracket* of A and B is the matrix $AB-BA$, and is denoted $[A,B]$. A linear subspace $\mathfrak{g} \subset M(n,\mathbb{R})$ is a *linear Lie algebra* if it is closed under the Lie bracket—that is, $[A,B] \in \mathfrak{g}$ for every $A, B \in \mathfrak{g}$.

Remark. Observe that $\mathfrak{g}$ need *not* be closed under matrix multiplication, and hence is not an associative algebra (the sort described above), since as one may readily verify, the Lie bracket is non-associative.

The following properties of the Lie bracket are established by straightforward calculations.

Proposition 19.2. *The Lie bracket has the following properties:*

(1) Bilinearity*:* $[sA_1 + A_2, B] = s[A_1, B] + [A_2, B]$ *and* $[A, tB_1 + B_2] = t[A, B_1] + [A, B_2]$.
(2) Skew-symmetry*:* $[A,B] = -[B,A]$.
(3) Jacobi identity*:* $[[A,B],C] + [[B,C],A] + [[C,A],B] = 0$.

We claim that closure under the Lie bracket is precisely the property a linear subspace of $M(n,\mathbb{R})$ needs in order for its image under the exponential map to be a subgroup of $GL(n,\mathbb{R})$. To see this, one needs the following lemma (which we do not prove):

Lemma 19.3. *Given $A, B \in M(n,\mathbb{R})$, there exists $C \in M(n,\mathbb{R})$ such that the following hold:*

(1) $e^A e^B = e^C$.
(2) C can be written as an absolutely converging infinite sum of matrices of the form

$$[\cdots[[X_1, X_2], X_3], \cdots, X_n], \tag{19.1}$$

where each X_i is either A or B.

The explicit expression for C is called the *Campbell–Hausdorff formula*; the key consequence of this formula is that we can now prove the subgroup property for the image of a Lie algebra under the exponential map.

Theorem 19.4. *If $\mathfrak{g} \subset M(n,\mathbb{R})$ is a Lie algebra, then $e^{\mathfrak{g}}$ is a subgroup of $GL(n,\mathbb{R})$.*

Proof. Let A, B, C be as in Lemma 19.3. It follows from the second part of the lemma that $C \in \mathfrak{g}$, since every term of the form (19.1) is in $\mathfrak{g}$. Thus $e^A e^B = e^C \in e^{\mathfrak{g}}$, and so $e^{\mathfrak{g}}$ is closed under multiplication. Closure under inverses is immediate; hence $e^{\mathfrak{g}}$ is a subgroup. □

b. Lie groups. Theorem 19.4 gives a natural way of producing subgroups of $GL(n,\mathbb{R})$, and it is very profitable to examine such subgroups in terms of their corresponding Lie algebras. Before saying more about this, we observe that $GL(n,\mathbb{R})$ has more than just an algebraic structure; it has a topological structure as well, in which convergence of a sequence of matrices corresponds to convergence of all the sequences of entries. Thus we may naturally consider subsets which are not only closed in the algebraic sense (that is, subgroups), but in the topological sense as well. This brings us to the central definition.

Definition 19.5. A *linear Lie group* is a (topologically) closed subgroup of $GL(n,\mathbb{R})$.

Under this definition, discrete subgroups qualify, but they are not really what we are interested in (and they introduce a great deal of complexity which we prefer not to deal with at the present time). Thus we will restrict our attention to *connected* linear Lie groups.

Any one-dimensional subspace of $M(n,\mathbb{R})$ is a Lie algebra, and so we may repeat our earlier observation that every matrix $A \in M(n,\mathbb{R})$ generates a one-parameter subgroup of $GL(n,\mathbb{R})$, which comprises all the matrices e^{tA}. However, this subgroup is *not* automatically a Lie group.

Example 19.6. Fix $\alpha \in \mathbb{R}$, and consider the matrix $A = \left(\begin{smallmatrix} 0 & \alpha \\ -\alpha & 0 \end{smallmatrix}\right) \in M(2,\mathbb{R})$. Observe that

$$\begin{aligned} e^A &= I + A + \frac{1}{2}A^2 + \frac{1}{3!}A^3 + \frac{1}{4!}A^4 + \cdots \\ &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} 0 & \alpha \\ -\alpha & 0 \end{pmatrix} + \frac{1}{2}\begin{pmatrix} -\alpha^2 & 0 \\ 0 & -\alpha^2 \end{pmatrix} + \frac{1}{3!}\begin{pmatrix} 0 & -\alpha^3 \\ \alpha^3 & 0 \end{pmatrix} + \cdots \\ &= \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix}. \end{aligned}$$

Now consider the matrix

$$B = \begin{pmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & \alpha \\ 0 & 0 & -\alpha & 0 \end{pmatrix};$$

a similar computation shows that

$$e^{tB} = \begin{pmatrix} \cos t & -\sin t & 0 & 0 \\ \sin t & \cos t & 0 & 0 \\ 0 & 0 & \cos t\alpha & -\sin t\alpha \\ 0 & 0 & \sin t\alpha & \cos t\alpha \end{pmatrix}.$$

It is not difficult to show that the subgroup $\{e^{tB} \mid t \in \mathbb{R}\}$ is dense in the set of all matrices of the form

$$e^{tB} = \begin{pmatrix} \cos t & -\sin t & 0 & 0 \\ \sin t & \cos t & 0 & 0 \\ 0 & 0 & \cos s & -\sin s \\ 0 & 0 & \sin s & \cos s \end{pmatrix},$$

where $s, t \in \mathbb{R}$ are arbitrary; consequently, this subgroup is not closed, and hence is not a Lie group.

One may ask if every linear Lie group is of the form $e^{\mathfrak{g}}$ for some Lie algebra $\mathfrak{g}$. It turns out that this is indeed the case. Letting $G \subset GL(n, \mathbb{R})$ be an arbitrary linear Lie group, we may consider the continuous one-parameter subgroups of G; these correspond to continuous homomorphisms $\varphi \colon \mathbb{R} \to G$, and as before, they are generated by the exponentials of the derivatives $\varphi'(0)$. Considering all matrices in $M(n, \mathbb{R})$ obtained as $\varphi'(0)$ for some one-parameter subgroup, one may show that we have a Lie algebra $\mathfrak{g}$ such that $G = e^{\mathfrak{g}}$. However, we do not give the details of the proof here.

Remark. For the benefit of the reader who has some familiarity with differentiable manifolds, we observe that a Lie group is a differentiable manifold, and continuous one-parameter subgroups correspond to certain curves through I; in particular, the Lie algebra just described is nothing but the tangent space to the manifold at I.

Remark. We have given the definitions of Lie groups and Lie algebras in a very concrete way, in terms of matrices. One can also give an abstract definition of both these objects, which does not mention matrices at all; however, if the elements of the Lie algebra are merely abstract entities rather than matrices, it is not *a priori* obvious how to define the exponential map. In fact, there are certain topological issues that get in the way of a completely clean correspondence.

Before turning our attention to specific examples, we remark that the technique of studying a Lie group by examining its associated Lie algebra is actually quite reminiscent of what we do in basic calculus, where we study finite objects (functions, curves) using infinitesimal tools (their derivatives). The finite objects are non-linear, while their infinitesimal counterparts are linear, and hence easier to study. In the present case, the finite objects are Lie groups (which are "non-linear" in an appropriate sense) and their infinitesimal counterparts are Lie algebras (which are linear spaces).

c. Examples. All the matrix groups that we have studied are given by nice equations, and so we quickly see that they are closed subgroups of $GL(n, \mathbb{R})$, hence Lie groups. For example, a limit of upper triangular matrices is upper triangular; hence $UT(n)$ is closed. Similarly for the unipotent group $\mathcal{U}_n$, the Heisenberg group H_n, the special linear group $SL(n, \mathbb{R})$, the special orthogonal group $SO(n)$, etc. For $UT(n)$, $\mathcal{U}_n$, and H_n, the defining equations are very simple: each one takes the form $A_{ij} = 0$, $A_{ij} = 1$, or $A_{ij} \neq 0$, depending on the context and the values of i, j. For $SO(n)$, there are $n(n+1)/2$ equations, each of which has only terms of quadratic order. For $SL(n, \mathbb{R})$, there is a single equation, $\det A = 1$, which comprises terms of order n when written in terms of the entries of A.

Remark. As an aside, we observe that $SL(2, \mathbb{R})$ is a hyperboloid of sorts in the four-dimensional vector space $M(n, \mathbb{R})$. Indeed, the condition for a matrix $\left(\begin{smallmatrix} x_1 & x_2 \\ x_3 & x_4 \end{smallmatrix}\right)$ to lie in $SL(2, \mathbb{R})$ is that $x_1x_4 - x_2x_3 = 1$, which may be rewritten

$$(x_1 + x_4)^2 - (x_1 - x_4)^2 - (x_2 + x_3)^2 + (x_2 - x_3)^2 = 4.$$

Given a Lie group G, we write $\mathcal{L}(G)$ for its associated Lie algebra. We will compute $\mathcal{L}(G)$ for the above examples by examining the possible values of $\varphi'(0)$, where $\varphi\colon \mathbb{R} \to G$ is a differentiable map with $\varphi(0) = I$.

Example 19.7. If $\varphi(t) \in \mathcal{D}_n$ for each t, then $\varphi'(t)$ is diagonal as well; thus $\mathcal{L}(\mathcal{D}_n)$ is contained in the set of all diagonal $n \times n$ matrices. Conversely, if $A = \operatorname{diag}(\lambda_1, \dots, \lambda_n)$, then $e^{tA} = \operatorname{diag}(e^{t\lambda_1, \dots, t\lambda_n}) \in \mathcal{D}_n$, and so $\mathcal{L}(\mathcal{D}_n)$ is equal to the set of all diagonal $n \times n$ matrices. Since all these matrices commute, we see that the Lie bracket on this Lie algebra is trivial.

Example 19.8. Let $G = \mathcal{U}_n$ be the Lie group of unipotent matrices. Given a differentiable map $\varphi\colon \mathbb{R} \to G$ with $\varphi(0) = I$, let $\varphi(t)_{ij}$ denote the i,jth entry of $\varphi(t)$. Since $\varphi(t)$ is unipotent for all t, we have $\varphi(t)_{ij} = 0$ for all $i > j$ and $\varphi(t)_{ii} = 1$ for all t. In particular, $\varphi'(0)_{ij} = 0$ for all $i \geq j$, hence $\varphi'(t)$ is nilpotent and upper-triangular, and we have $\mathcal{L}(\mathcal{U}_n) \subset \mathcal{N}_n$.

Conversely, if N is any nilpotent matrix, then so is N^k for all k, and it follows from (18.6) that $e^N \in \mathcal{U}_n$. Thus $\mathcal{L}(\mathcal{U}_n) = \mathcal{N}_n$.

Example 19.9. Let $G = SL(n, \mathbb{R})$ be the Lie group of matrices with unit determinant. Given a differentiable map $\varphi\colon \mathbb{R} \to G$ with $\varphi(0) = I$, we observe that

$$\begin{aligned} 0 &= \frac{d}{dt}(\det \varphi(t))|_{t=0} \\ &= \frac{d}{dt}\left(\sum_{\sigma \in S_n} \operatorname{sgn} \sigma \prod_{i=1}^{n} \varphi(t)_{i,\sigma(i)}\right)\Big|_{t=0} \\ &= \sum_{\sigma \in S_n} \operatorname{sgn} \sigma \sum_{j=1}^{n} \varphi'(0) \prod_{i \neq j} \varphi(0) \\ &= \sum_{j=1}^{n} \varphi'(0) = \operatorname{Tr} \varphi'(0), \end{aligned} \tag{19.2}$$

where we write $\operatorname{sgn} \sigma = +1$ for an even permutation $\sigma \in S_n$ and $\operatorname{sgn} \sigma = -1$ for an odd permutation. The last equality uses the fact that $\varphi(0)_{ij} = \delta_{ij}$ since $\varphi(0) = I$, which implies that the only non-vanishing term comes when

σ is the identity permutation. We see from (19.2) that $\mathcal{L}(G)$ is contained in the space of matrices with zero trace, which we denote by $\mathfrak{sl}(n, \mathbb{R})$.

Furthermore, given any traceless matrix $A \in \mathfrak{sl}(n, \mathbb{R})$, we observe that $\lambda \in \mathbb{C}$ is an eigenvalue of A if and only if e^λ is an eigenvalue of e^A (and with the same multiplicity). Since the trace of a matrix is the sum of its eigenvalues, and the determinant is their product, we immediately see that $\det e^A = e^{\operatorname{Tr} A}$, and in particular, $\det e^{tA} = 1$ for all $t \in \mathbb{R}$. This shows that $\mathfrak{sl}(n, \mathbb{R}) = \mathcal{L}(SL(n, \mathbb{R}))$.

Example 19.10. Let $G = SO(n, \mathbb{R})$ be the Lie group of orthogonal matrices with unit determinant. Given a differentiable map $\varphi \colon \mathbb{R} \to G$ with $\varphi(0) = I$, we observe that $\varphi(t)\varphi(t)^T = I$ for every t, and that entry-wise, this may be written $\sum_{k=1}^n \varphi(t)_{ik}\varphi(t)_{jk} = \delta_{ij}$ for every i, j. Differentiating, we obtain

$$\sum_{k=1}^{n} \varphi'(0)_{ik}\varphi(0)_{jk} + \varphi(0)_{ik}\varphi(0)_{jk} = 0, \tag{19.3}$$

and once more using the fact that $\varphi(0)_{jk} = \delta_{jk}$, we see that there are only two non-vanishing terms here, which yield

$$\varphi'(0)_{ij} + \varphi'(0)_{ji} = 0.$$

That is, $\varphi'(0)$ lies in the space of skew-symmetric $n \times n$ matrices, which we denote $\mathfrak{so}(n, \mathbb{R})$. To see that $\mathcal{L}(SO(n, \mathbb{R})) = \mathfrak{so}(n, \mathbb{R})$, we observe that given an arbitrary skew-symmetric matrix $A \in \mathfrak{so}(n, \mathbb{R})$, we have $A + A^T = 0$, and since in general $(e^A)^T = e^{A^T}$, we get

$$(e^A)(e^A)^T = e^{A+A^T} = e^0 = I,$$

using the fact that A and $A^T = -A$ commute. Thus $e^A \in SO(n, \mathbb{R})$ (the fact that $\det e^A = 1$ follows from the fact that $\operatorname{Tr} A = 0$), and we are done.

Remark. Once again, we remark that the theory can be developed in an abstract setting, without immediately restricting our attention to matrix groups. A Lie group can be defined as a differentiable manifold with a group structure whose operations are differentiable, and a Lie algebra can be defined as a vector space endowed with a Lie bracket satisfying Proposition 19.2.

The question is, is the abstract setting really any more general? It turns out that the answer is yes... but just barely. To see what happens, observe that in going from $\mathfrak{sl}(2, \mathbb{R})$ to $SL(2, \mathbb{R})$, we have the relationship $\exp\left(\begin{smallmatrix} 0 & 2\pi \\ 2\pi & 0 \end{smallmatrix}\right) = I$. One can construct an abstract Lie group whose Lie algebra is $\mathfrak{sl}(2, \mathbb{R})$, but which is not $SL(2, \mathbb{R})$, because the matrix given has a non-trivial image under the exponential map. This has to do with the topology of $SL(2, \mathbb{R})$, in particular with the so-called *fundamental group* that happens to be our next subject in these lectures.

CHAPTER 4

Fundamental group: A different kind of group associated to geometric objects

Lecture 20. Wednesday, October 21

a. Isometries and homeomorphisms. By and large, we have been considering groups that arise from geometric objects as collections of symmetries; now we turn our attention to a different class of groups, which opens the door on the world of algebraic topology.

We begin by highlighting the distinction between geometry and topology in the context of metric spaces. As with so many distinctions between various closely related fields of mathematics, the distinction hinges on the conditions under which we consider two metric spaces to be "the same" or "equivalent".

The natural equivalence relation in metric geometry is isometry. Recall that two metric spaces (X, d) and (X', d') are *isometric* if there exists an isometric bijection between them—that is, a bijection $f\colon X \to X'$ such that $d'(f(x_1), f(x_2)) = d(x_1, x_2)$ for all $x_1, x_2 \in X$. For example, any two circles in $\mathbb{R}^2$ with the same radius are isometric, regardless of their centre, while the circles $x^2 + y^2 = 1$ and $x^2 + y^2 = 4$ are *not* isometric to each other, nor to the square with vertices at $(\pm 1, \pm 1)$, nor to the line $x = 1$.

Nevertheless, we feel that the two circles are in some sense more akin to each other than they are to either the square or the line, and that the circles and the square are somehow more akin than the circles and the line, or the square and the line. One may make the first feeling precise by observing that there is a similarity transformation $f\colon \mathbb{R}^2 \to \mathbb{R}^2$ that takes the circle of radius 1 to the circle of radius 2; indeed, *any* two circles are equivalent up to a similarity transformation. Thus passing from metric geometry to similarity geometry is a matter of weakening the conditions under which two objects may be considered equivalent.

Weakening these conditions still further, we may consider allow even more general maps f. Writing X for the square with vertices at $(\pm 1, \pm 1)$ and S^1 for the unit circle, we may define a bijection $f\colon X \to S^1$ by $f(\mathbf{x}) = \mathbf{x}/\|\mathbf{x}\|$, where $\|(x, y)\| = \sqrt{x^2 + y^2}$. One may easily verify that given a sequence of points $\mathbf{x}_n \in X$, we have $\mathbf{x}_n \to \mathbf{x}_0$ on the square if and only if $f(\mathbf{x}_n) \to f(\mathbf{x}_0)$ on the circle—that is, both f and f^{-1} are continuous. Such a map f is called a *homeomorphism*, and is the natural equivalence relation between topological spaces.

Remark. In fact, a more general definition of continuity uses not convergence of sequences (a local property), but open sets (a global property): a map $f\colon X \to Y$ is continuous if $f^{-1}(U)$ is open in X whenever U is open Y. If X and Y are metric spaces, this is equivalent to the definition in terms of sequences. Using this language, a homeomorphism is a bijection such that $f(U)$ is open if and only if U is open.

A fundamental observation at this point is that given a metric space (X, d), there are in general many different metrics one can place on X that induce the same topology as d does—that is, there are many metrics d' on X such that $d(x_n, x_0) \to 0$ if and only if $d'(x_n, x_0) \to 0$. Thus when we are interested in topological matters, the primary importance of a distance function is not the geometric structure it induces (which is unique to that particular metric), but rather the topological structure (which is held in common by many equivalent metrics).

Exercise 20.1. Let $X = \mathbb{R}^n$, and for every $p \geq 1$ consider the function

$$d_p(\mathbf{x}, \mathbf{y}) = ((x_1 - y_1)^p + \cdots + (x_n - y_n)^p)^{\frac{1}{p}}. \tag{20.1}$$

Shown that d_p is a metric for every $p \geq 1$ and that each of these metrics defines the same topology on $\mathbb{R}^n$ (the standard one).

Exercise 20.2. Consider the following distance function in $\mathbb{R}^2$:

$$d_L((x_1, x_2), (y_1, y_2)) = |x_1 - y_1| + |x_2 - y_2| + 1 - \delta_{x_2, y_2}.$$

Prove that d_L defines a metric that is not equivalent to the standard one.

b. Tori and $\mathbb{Z}^2$. Now we consider the various sorts of tori we have encountered so far. Visually, of course, the most natural way to view a torus is as the surface of a bagel or doughnut embedded in $\mathbb{R}^3$. However, we also refer to the quotient group $\mathbb{R}^2/\mathbb{Z}^2$ as a torus. Indeed, given any two linearly independent vectors $\mathbf{v}$ and $\mathbf{w}$ in $\mathbb{R}^2$, we may consider the lattice $L = \{a\mathbf{v} + b\mathbf{w} \mid a, b \in \mathbb{Z}^2\}$, which is a normal subgroup of $\mathbb{R}^2$, and take the quotient group $\mathbb{R}^2/L$; this is again a torus.

Another example is direct product of two circles (or two closed curves) in the plane that can be viewed as embedded into $\mathbb{R}^4$. One can also generalize the doughnut example by rotating around the z axis not a circle but another closed curve that may have no internal symmetries of putting a warp on the surface.

Geometrically, these tori are quite different from each other.

- Writing X for the surface of the bagel, or more precisely, the set of points $(x, y, z) \in \mathbb{R}^3$ such that

$$(\sqrt{x^2 + y^2} - 2)^2 + z^2 = 1, \tag{20.2}$$

observe that the only isometries of X are rotations around the z-axis by an arbitrary angle, rotations around any axis through the origin in the xy-plane by an angle of exactly π, and reflections in either the xy-plane or any plane that contains the z-axis. In particular, $\mathrm{Isom}(X)$ only has one continuous degree of freedom.

- Tori obtained by rotating a non-symmetric closed curve have fewer isometries in $\mathrm{Isom}\,\mathbb{R}^3$ that the bagel; symmetry in the xy-plane is lost. It turns out however that it is recovered when one considers intrinsic metric on the torus.
- By putting a warp on the bagel's surface one may destroy all isometries.
- Since we have seen already that $\mathrm{Isom}(\mathbb{R}^2/\mathbb{Z}^2)$ (or $\mathrm{Isom}(\mathbb{R}^2/L)$) contains all translations, and hence has two continuous degrees of freedom, we conclude that the geometries of the two tori are quite different from each other; in particular, the torus $\mathbb{R}^2/\mathbb{Z}^2$ is in some sense more symmetric than the embedded torus X.
- There are also differences in the isometry groups $\mathrm{Isom}(\mathbb{R}^2/L)$ for different lattices. Generically translations form a subgroup of index two in the isometries of $\mathrm{Isom}(\mathbb{R}^2/L)$, the rest being rotations by π; there are no orientation reserving isometries and fewer isometries overall that for $\mathbb{R}^2/\mathbb{Z}^2$. But for the hexagonal lattice there are in a sense more isometries that for the rectangular one, i.e. the index of the translation subgroup in all isometries is higher: twelve rather than eight.
- Interestingly the product of the two circles is isometric to $\mathbb{R}^2/\mathbb{Z}^2$ while the product of two other curves has only discrete groups of isometries in $\mathrm{Isom}\,\mathbb{R}^4$ *intrinsically* though it is isometric to $\mathbb{R}^2/\mathbb{Z}^2$.

All statements about isometry groups in this list may be considered as useful exercises.

Despite the differences in their geometry, all these tori are homeomorphic: in the case of the bagel X the map $f\colon \mathbb{R}^2/\mathbb{Z}^2 \to X$ given by

$$f((s,t)+\mathbb{Z}^2) = ((2+\cos 2\pi s)\cos 2\pi t, (2+\cos 2\pi s)\sin 2\pi t, \sin 2\pi s) \tag{20.3}$$

can be easily verified to be a homeomorphism.

Exercise 20.3. Given a lattice L generated by vectors $\mathbf{v}$ and $\mathbf{w}$, show that L and $\mathbb{Z}^2$ are isomorphic subgroups of $\mathbb{R}^2$—that is, there exists a group isomorphism (invertible additive map) $f\colon \mathbb{R}^2 \to \mathbb{R}^2$ such that $f(\mathbb{Z}^2) = L$. Conclude that the two tori $\mathbb{R}^2/\mathbb{Z}^2$ and $\mathbb{R}^2/L$ are homeomorphic by using the map f to exhibit a homeomorphism between them.

Exercise 20.3 shows that all the tori $\mathbb{R}^2/L$ are homeomorphic; the key tool in the exercise is the fact that the lattices L are all isomorphic to $\mathbb{Z}^2$. This suggests that the group $\mathbb{Z}^2$ somehow plays an important role in the topology of the torus—but how? We will spend the remainder of this lecture and the next one making this notion clear.

Thus it looks as the group $\mathbb{Z}^2$ is somehow important to the structure of the torus. This is obvious if we consider the torus $\mathbb{T}^2 = \mathbb{R}^2/\mathbb{Z}^2$ as a factor group and direct our attention to the original group $\mathbb{R}^2$, within which $\mathbb{Z}^2$ sits as the integer lattice $\{(a,b) \mid a,b \in \mathbb{Z}\}$. However, this does not tell us how $\mathbb{Z}^2$ is related to the *intrinsic* structure of the torus $\mathbb{T}^2$—after all, every point in the integer lattice in $\mathbb{R}^2$ corresponds to the *same* point on the torus, and a single point does not have terribly much internal structure!

Another way of stating the problem is to observe that if the algebraic structure of $\mathbb{Z}^2$ characterises some aspect of the topological structure of the torus, then we should be able to describe $\mathbb{Z}^2$ in terms of *any* torus homeomorphic to $\mathbb{R}^2/\mathbb{Z}^2$. In particular, we want to produce $\mathbb{Z}^2$ in terms of objects on the embedded torus X given in (20.2). But how do we do this?

c. Paths and loops. Thinking once more in terms of the factor space $\mathbb{R}^2/\mathbb{Z}^2$, what we want is a description of the lattice points $\mathbb{Z}^2 \subset \mathbb{R}^2$ that is able to distinguish between different points on the lattice even after we pass to the quotient space $\mathbb{R}^2/\mathbb{Z}^2$. To this end, we consider not just lattice points, but *paths* between lattice points, as in Figure 4.1.

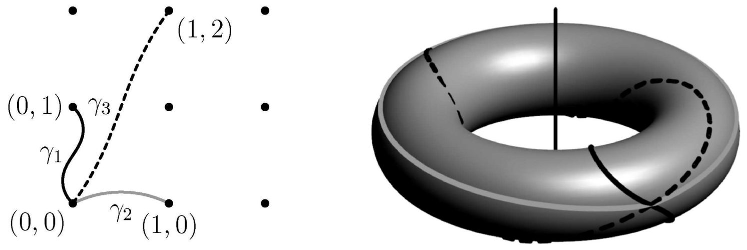


FIGURE 4.1. Paths in the plane and loops on the torus.

Recall that a path in $\mathbb{R}^2$ is given by a continuous function $\gamma\colon [0,1] \to \mathbb{R}^2$; such a path also defines a path on the factor torus by $\tilde{\gamma}\colon t \mapsto \gamma(t) + \mathbb{Z}^2$, and on the embedded torus by $\tilde{\gamma}(t) = f(\gamma(t))$, where f is given by (20.3). Let $\mathbf{p}$ be the point on the embedded torus that corresponds to the lattice points in $\mathbb{Z}^2$ under the map f, and observe that if γ is a path between lattice points, then $\tilde{\gamma}$ is a *loop* on the torus based at $\mathbf{p}$—that is, it has the property that $\tilde{\gamma}(0) = \tilde{\gamma}(1) = \mathbf{p}$. Figure 4.1 shows three such paths, both as paths in $\mathbb{R}^2$ and loops on the torus.

Of course, there are many paths in $\mathbb{R}^2$ that connect a particular pair of lattice points. For example, γ_1 is only one possible path from $\mathbf{0}$ to $\mathbf{x} = (0,1)$; a more natural choice would be $\gamma_0(t) = (0,t)$, which goes between the points along a straight line with uniform speed. These two paths are equivalent in the sense that one can be continuously deformed into the other while keeping the endpoints fixed—this visually obvious property is made precise as follows. Define a map $\Gamma\colon [0,1] \to [0,1]$ by

$$\Gamma(s,t) = (1-s)\gamma_0(t) + s\gamma_1(t). \tag{20.4}$$

The map Γ has several important properties:

(1) Γ depends continuously on both s and t.

(2) $\Gamma(0,t) = \gamma_0(t)$ and $\Gamma(1,t) = \gamma_1(t)$ for all $t \in [0,1]$.
(3) $\Gamma(s,0) = \mathbf{0}$ and $\Gamma(s,1) = \mathbf{x}$ for all $s \in [0,1]$.

The cross-sections $\Gamma(s,\cdot)$ each define a path by $\gamma_s(t) = \Gamma(s,t)$. The first property above states that the paths γ_s are each continuous and that they vary continuously with s. The second property states that the family of paths γ_s connects γ_0 and γ_1—that is, it continuously deforms one into the other. Finally, the third property states every path γ_s runs from $\mathbf{0}$ to $\mathbf{x}$—that is, the endpoints are held fixed even as the rest of the curve moves. We say that γ_0 and γ_1 are *homotopic relative to* $\{\mathbf{0}, \mathbf{x}\}$.

The condition that the endpoints be held fixed is essential. Indeed, if we remove this condition, then any two paths in $\mathbb{R}^2$ can be related by a linear homotopy as in (20.4). but this homotopy does not project to the torus as a family of closed paths. One may of course consider an intermediate condition: a homotopy between loops on the torus that does not fix a point. While this condition (called *free homotopy*) makes perfect sense geometrically, classes of free homotopic paths are not amenable to algebraic manipulations, unlike the classes of paths homotopic relative to a point.

Given $\mathbf{x} \in \mathbb{Z}^2$ and a path γ in $\mathbb{R}^2$ with $\gamma(0) = \mathbf{0}$ and $\gamma(1) = \mathbf{x}$, let $[\gamma]$ denote the set of all paths in $\mathbb{R}^2$ that are homotopic to γ relative to $\{\mathbf{0}, \mathbf{x}\}$—that is, the set of all paths that can be continuously deformed into γ without moving their endpoints. Observe that $[\gamma]$ comprises all paths that start at $\mathbf{0}$ and end at $\mathbf{x}$, and that this gives a one-to-one correspondence between lattice points $\mathbb{Z}^2$ and equivalence classes of paths starting at $\mathbf{0}$.

Thus we have associated the elements of the group $\mathbb{Z}^2$ to equivalence classes of paths in $\mathbb{R}^2$; we will now see that these equivalence classes are still distinguishable when we pass to the torus.

As remarked above, paths γ in $\mathbb{R}^2$ with endpoints in $\mathbb{Z}^2$ correspond to loops $\tilde{\gamma}$ on the torus—paths with $\tilde{\gamma}(0) = \tilde{\gamma}(1) = \mathbf{p}$. We can define equivalence classes just as before: two loops $\tilde{\gamma}_0$ and $\tilde{\gamma}_1$ based at $\mathbf{p}$ are homotopic relative to $\mathbf{p}$ if they can be deformed into each other via a continuous family of continuous paths, each of which is also a loop based at $\mathbf{p}$.

In $\mathbb{R}^2$, we were able to characterise $[\gamma]$ as the set of all paths from $\mathbf{0}$ with the same endpoint as γ; this no longer holds on the torus, since all lattice points are identified with the point $\mathbf{p}$. However, it is not the case that all loops on the torus are homotopic—for example, $\tilde{\gamma}_1$ and $\tilde{\gamma}_2$ in Figure 4.1 cannot be continuously deformed into each other. So what characterises the different homotopy classes?

Heuristically, the answer is as follows (for the torus at least). Let Z denote the z-axis, and let C denote the circle in the xy-plane of radius 2 centred at the origin (the embedded torus in $\mathbb{R}^3$ is the set of all points whose distance from C is exactly 1). Observe that $\tilde{\gamma}_1$, which corresponds to the path γ_1 from $\mathbf{0}$ to $(0,1)$, wraps around C exactly once, and Z not at all; similarly, $\tilde{\gamma}_2$, which corresponds to the path γ_2 from $\mathbf{0}$ to $(1,0)$, wraps around Z exactly once, and C not at all. A slightly more careful look at

Figure 4.1 shows that $\tilde{\gamma}_3$, which corresponds to the path γ_3 from $\mathbf{0}$ to $(1, 2)$, wraps around Z exactly once, and around C twice.

In general, if γ is a path from $\mathbf{0}$ to (a, b), then the corresponding curve $\tilde{\gamma}$ on the embedded torus wraps a times around Z and b times around C. Thus we may think of $(1, 0)$ and $(0, 1)$, the generators of $\mathbb{Z}^2$, as representing the two "holes" in the torus: if we think of the embedded torus as a hollowed-out doughnut, then one hole (corresponding to $(1, 0)$ and Z) is the "doughnut hole" through the centre, and the other hole (corresponding to $(0, 1)$ and C) is the hollowed-out part (where the jelly would go, perhaps).

One thing is not yet clear. We wanted to give an intrinsic description of the group $\mathbb{Z}^2$ in terms of the embedded torus; so far we have described the *elements* of the group as loops on the torus (or rather, as equivalence classes of loops), but have not specified a binary operation. There is a fairly natural candidate, though, using which we can complete the construction, and we do this in the next section.

d. The fundamental group. Consider now an arbitrary metric space X, and fix a point $p \in X$ (this will be our base point). Given any two paths $\gamma_1, \gamma_2 \colon [0, 1] \to X$ with $\gamma_1(1) = \gamma_2(0)$, we can define a concatenated path $\gamma_1 \star \gamma_2$ by

$$(\gamma_1 \star \gamma_2)(t) = \begin{cases} \gamma_1(2t) & 0 \leq t \leq 1/2, \\ \gamma_2(2t-1) & 1/2 \leq t \leq 1. \end{cases} \tag{20.5}$$

That is, $\gamma_1 \star \gamma_2$ is the path that follows first γ_1 and then γ_2, moving with twice the speed of the original parametrisations so as to parametrise the entire path by the interval $[0, 1]$. In particular, if γ_1 and γ_2 are *loops* from p, then $\gamma_1 \star \gamma_2$ is a loop from p as well.

We saw in the previous section that the key objects are not loops *per se*, but equivalence classes of loops. Thus we formalise the discussion there as follows.

Definition 20.1. Let $\gamma_0, \gamma_1 \colon [0, 1] \to X$ be continuous paths with $\gamma_0(0) = \gamma_1(0) = \gamma_0(1) = \gamma_1(1) = p$. We say that γ_0 and γ_1 are *homotopic relative to* p if there exists a continuous function $\Gamma \colon [0, 1] \times [0, 1] \to X$ such that

(1) $\Gamma(0, t) = \gamma_0(t)$ and $\Gamma(1, t) = \gamma_1(t)$ for all $0 \leq t \leq 1$.
(2) $\Gamma(s, 0) = \Gamma(s, 1) = p$ for all $0 \leq s \leq 1$.

In this case we write $\gamma_0 \sim \gamma_1$. The set of all loops from p that are homotopic to γ relative to p is called the *homotopy class* of γ, and is denoted $[\gamma]$.

The binary operation of concatenation works not just on loops, but on homotopy classes of loops: given loops γ and η, we define $[\gamma] \star [\eta]$ to be the homotopy class $[\gamma \star \eta]$. We must check that this is well-defined, but once we do so, we will finally have in our hands the fundamental object of algebraic topology.

Definition 20.2. Given a metric space X and a point $p \in X$, the *fundamental group* of X with base point p is the collection of homotopy classes of loops based at p together with the binary operation of concatenation. We denote this group by $\pi_1(X, p)$.

Of course, this terminology puts the cart before the horse. Right now all we have is a set together with a binary operation (which may not even be well-defined, for all we know). Why is this a group?

PROPOSITION 20.3. *The binary operation $\star$ is well-defined on $\pi_1(X, p)$ and makes it into a group.*

PROOF. We first show that $\star$ is well-defined—that is, that $\gamma_1\star\eta_1 \sim \gamma_2\star\eta_2$ whenever $\gamma_1 \sim \gamma_2$ and $\eta_1 \sim \eta_2$. An equivalent way of stating this condition is that the equivalence class $[\gamma\star\eta]$ is the same no matter which representatives of $[\gamma]$ and $[\eta]$ we work with.

The proof of this is straightforward: if Γ and H are homotopies demonstrating $\gamma_1 \sim \gamma_2$ and $\eta_1 \sim \eta_2$, respectively, we can concatenate them to obtain a homotopy between $\gamma_1 \star \eta_1$ and $\gamma_2 \star \eta_2$. To wit, define a continuous function $G\colon [0, 1] \times [0, 1] \to X$ as follows:

$$G(s,t) = \begin{cases} \Gamma(s, 2t) & 0 \le t \le 1/2, \\ H(s, 2t-1) & 1/2 \le t \le 1. \end{cases}$$

One may easily verify that G is the required homotopy. A visual representation of this procedure is shown in Figure 4.2(a), where the vertical lines are level sets of the function G—that is, values of s and t which G sends to the same point in X.

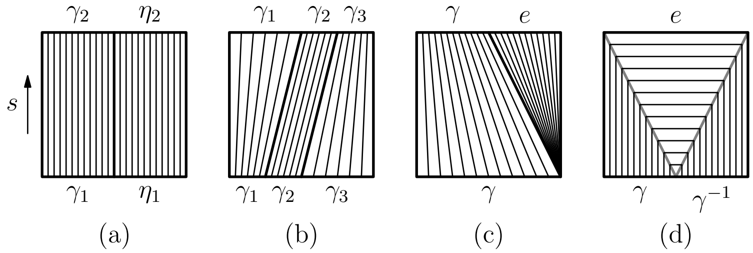


FIGURE 4.2. Homotopy equivalences that make $\pi_1(X)$ a group.

A similar representation is used in Figure 4.2(b)–(d), which essentially contains all the remaining elements of the proof. Let us explain this claim.

Now that we know $\star$ is well-defined, we must show that it is associative. As shown along the bottom edge of part (b) of the figure, $(\gamma_1 \star \gamma_2) \star \gamma_3$ is the curve which traverses γ_1 from $t = 0$ to $t = 1/4$, then γ_2 from $t = 1/4$ to $t = 1/2$, and finally γ_3 from $t = 1/2$ to $t = 1$. The top edge represents $\gamma_1 \star (\gamma_2 \star \gamma_3)$, for which the points traversed are the same, but the parametrisation is different. Using the piecewise linear homotopy

$$G(s,t) = \begin{cases} \gamma_1((s+1)t) & 0 \leq t \leq \frac{s+1}{4}, \\ \gamma_2(t-(s+1)/4) & \frac{s+1}{4} \leq t \leq \frac{s+2}{4}, \\ \gamma_3(1-(s+1)t) & \frac{s+2}{4} \leq t \leq 1, \end{cases}$$

we see that $[(\gamma_1 \star \gamma_2) \star \gamma_3] = [\gamma_1 \star (\gamma_2 \star \gamma_3)]$, and hence $\star$ is associative. Once again, the lines in Figure 4.2(b) correspond to values of s and t which G sends to the same place in X.

Observe that G does not change the geometry of the above paths at all—indeed, it is nothing more than a reparametrisation! This is an important special case of homotopy equivalence, and is also what we need in order to satisfy the next group axiom, the existence of an identity element. The natural candidate for the identity element in the fundamental group $\pi_1(X,p)$ is the trivial loop $e\colon [0,1] \to X$, for which $e(t) = p$ for all $0 \leq t \leq 1$. Concatenating any loop γ with e does not change its geometry, and the simple piecewise linear reparametrisation shown in Figure 4.2(c) suffices to show that $[\gamma] \star [e] = [\gamma \star e] = [\gamma]$ for all loops γ, and similarly $[e] \star [\gamma] = [\gamma]$.

Reparametrisation is *not* enough to get us the final group axiom, the existence of inverse elements. Indeed, as soon as a loop γ is non-trivial and goes to points other than p, it cannot be a reparametrisation of the trivial loop. Rather, a genuine homotopy is required; the key is that we consider loops not just as geometric objects (the image $\gamma([0,1])$), but also record the "history" of movement along the path. Thus the inverse γ^{-1} ought to be the loop which "undoes" γ, so we write $\gamma^{-1}(t) = \gamma(1-t)$ to obtain a loop that traverses the same curve as γ, but does so in the reverse direction.

To show that $\gamma \star \gamma^{-1} \sim e$, we use the homotopy shown in Figure 4.2(d), which may be given the following explicit form:

$$G(s,t) = \begin{cases} \gamma(t) & 0 \leq t \leq \frac{1-s}{2}, \\ \gamma(\frac{1-s}{2}) = \gamma^{-1}(\frac{1+s}{2}) & \frac{1-s}{2} \leq t \leq \frac{1+s}{2}, \\ \gamma^{-1}(t) & \frac{1+s}{2} \leq t \leq 1. \end{cases}$$

The path $G(s,\cdot)$ follows γ as far as $\gamma((1-s)/2)$, then stops and thinks about things for a while, and finally retraces its steps to end where it began, at p. As s goes from 0 to 1, the amount of γ that $G(s,\cdot)$ traverses gets smaller and smaller, until finally $G(1,\cdot)$ is just the trivial loop e. This homotopy establishes that $[\gamma] \star [\gamma^{-1}] = [e]$, and hence $\pi_1(X,p)$ is indeed a group. □

We have successfully produced a group from the intrinsic topological data of X. However, several questions remain. The definition involves an arbitrarily chosen point p; what happens if we choose a different point p as our base point? Do we get a different group? What does this group look like for familiar examples, such as the circle, the sphere, the plane, the torus, etc.? Part of our motivation was to recover the group $\mathbb{Z}^2$ from the intrinsic properties of the torus—did it work? Or is $\pi_1(\mathbb{T}^2, \mathbf{p})$ something else?

We will defer specific examples until the next lecture; for now we address the first question, and consider the groups $\pi_1(X, p)$ and $\pi_1(X, q)$ for points $p \neq q \in X$.

Definition 20.4. A metric space X is *path-connected* if for every $p, q \in X$ there exists a continuous path $\gamma\colon [0,1] \to X$ such that $\gamma(0) = p$ and $\gamma(1) = q$.

Proposition 20.5. *If X is a path-connected metric space, then $\pi_1(X, p)$ and $\pi_1(X, q)$ are isomorphic for any $p, q \in X$.*

Proof. Given $p, q \in X$, let $\alpha\colon [0,1] \to X$ be a continuous path such that $\alpha(0) = p$ and $\alpha(1) = q$. Define a map $\varphi\colon \pi_1(X,p) \to \pi_1(X,q)$ by $\varphi([\gamma]) = [\alpha^{-1} \star \gamma \star \alpha]$. The proof that φ is well-defined exactly mirrors the proof for $\star$ in Proposition 20.3. Furthermore, φ is a homomorphism, since

$$\begin{aligned}\varphi([\gamma] \star [\eta]) &= [\alpha^{-1} \star \gamma \star \eta \star \alpha] \\ &= [\alpha^{-1} \star \gamma \star \alpha \star \alpha^{-1} \star \eta \star \alpha] \\ &= \varphi([\gamma]) \star \varphi([\eta]),\end{aligned}$$

where the second equality uses the fact that $\alpha \star \alpha^{-1} \sim e_p$, and that $\alpha^{-1} \star \gamma \star e_p \star \eta \star \alpha$ is a reparametrisation of $\alpha^{-1} \star \gamma \star \eta \star \alpha$.

Now we observe that φ is onto, since φ^{-1} can be defined by $[\zeta] \mapsto [\alpha \star \zeta \star \alpha^{-1}]$ for every $[\zeta] \in \pi_1(X, q)$. Furthermore, $\varphi([\gamma]) = [e_q]$ implies $\gamma \sim \alpha \star e_q \star \alpha^{-1} \sim e_p$, and so φ is one-to-one. It follows that φ is an isomorphism. □

As a consequence of Proposition 20.5, we can (and will) speak of the fundamental group of X, and write $\pi_1(X)$, without explicitly mentioning the base point, since changing the base point merely yields an isomorphic group.

e. Algebraic topology. In algebraic topology one associates to various kinds of topological spaces algebraic objects, usually groups, moduli or rings. The fundamental group we just described is a premiere and arguably most geometrically transparent example of such an association. Two leading principle of algebraic topology are *invariance* and *functoriality.* The former requires that equivalent spaces are associated with isomorphic objects and that the association is independent of auxiliary elements involved in the construction of an algebraic object. We already have an example in the case of fundamental group of a path connected space: construction does not depend on the base point used and homeomorphic spaces have isomorphic fundamental groups. Functoriality in its simplest form requires that continuos maps between spaces "naturally" induce homomorphisms between associated algebraic object; the direction of this homomorphism may be the same as for the map (*covariant* constructions) or the opposite (*contravariant* constructions). The fundamental groups is an example of the former.

Furthermore this homomorphisms should behave properly under the composition of maps.

Proposition 20.6. *Let $f : X \to Y$ be a continuous map and $p \in X$. Then for $[\gamma] \in \pi_1(X, p)$ the path $\gamma \circ f : [0, 1] \to Y$ defines an element $f_*([\gamma]) \in \pi_1(Y, f(p))$ and $f_: * \pi_1(X, p) \to \pi_1(Y, f(p))$ is a group homomorphism.*
If $g : Y \to Z$ then $(gf)_* = g_* f_*$.

Proof. Since composition of a path homotopy in X with a continuous map is a path homotopy in Y the map f_* is correctly defined. Concatenation of paths goes into concatenation of their images, hence the map f_* is a homomorphism. The last statement is obvious since it is true already at the level of paths. □

Lecture 21. Friday, October 23

a. Homotopic maps, contractible spaces and homotopy equivalence. The notion of homotopy from the previous lecture can be applied not just to paths, but to *any* continuous maps. Given two metric spaces X and Y, we say that two continuous maps $f, g\colon X \to Y$ are *homotopic* if there exists a continuous function $\Gamma\colon [0,1] \times X \to Y$ such that $\Gamma(0,x) = f(x)$ and $\Gamma(1,x) = g(x)$ for all $x \in X$. Heuristically, this means that the functions $\Gamma(s,\cdot)\colon X \to Y$ are a continuous one-parameter family of continuous maps that deform f into g.

This is the notion of *absolute* homotopy; observe that we place no restrictions on the functions $\Gamma(s,\cdot)$, in contrast to the previous lecture, where we required each of the paths $\Gamma(s,\cdot)$ to have endpoints at a fixed base point. Even though we later showed that the isomorphism class of the fundamental group is independent of the choice of base point, this base point still plays a prominent role in the definitions. This is emblematic of many topological constructions: in order to define a very general object, one must use definitions which in and of themselves depend on an arbitrary choice, but in the end the objects so defined are independent of which particular choice is made.

For the fundamental group, the "particular choice" is a choice of base point, which appears in the definitions via the notion of *relative* homotopy. Given two continuous maps $f, g\colon X \to Y$ and a subset $A \subset X$, we say that f and g are *homotopic relative to* A if there exists a continuous homotopy $\Gamma\colon [0,1] \times X \to Y$ with the properties above, along with the additional property that $\Gamma(s,x) = f(x) = g(x)$ for all $s \in [0,1]$ and $x \in A$. Thus relative homotopy is a matter of continuously deforming the map f into the map g, while keeping the action of the map $\Gamma(s,\cdot)$ on the set A fixed; in the previous lecture, we used homotopy relative to the set of endpoints $A = \{0,1\}$.

Once we have a definition of homotopy for maps, it is natural to ask what the possible homotopy classes of maps from X to Y are. For example, if $X = Y = S^1$, then it is intuitively clear that the homotopy class of $f\colon S^1 \to S^1$ is the set of all maps that "wind around the circle the same number of times as f does". We will make this precise shortly.

In the meantime, we note that given *any* metric space X, there are two natural maps from X to itself. One is the identity map, $\mathrm{Id}\colon x \to x$, and the other is the trivial (or constant) map $e_p\colon x \to p$, where p is some arbitrarily chosen point in X. Thus Id fixes every point in X, while e_p collapses all of X to a single point. We say that X is *contractible to the point p* if these two maps are homotopic—that is, if there exists a continuous map $\Gamma\colon [0,1] \times X \to X$ such that $\Gamma(0,x) = x$ and $\Gamma(1,x) = p$ for all $x \in X$.

Proposition 21.1. *Given any two points $p, q \in X$, X is contractible to p if and only if X is contractible to q.*

Proof. First notice that if X is contractible to a point p it is path connected, since for any $q \in X$ the homotopy $\Gamma(t, q)$ is a path connecting q with p. Combining the contraction to p with this path in the opposite direction, i.e. $\Gamma(1-t, q)$, gives a contraction of X to q. □

Thanks to Proposition 21.1, we may simply refer to X as being *contractible* without mentioning which point it is contractible to, since if it contractible to one point it is contractible to any point. This is another example of a general property that must be defined with reference to an arbitrarily chosen object, whose precise choice turns out not to matter.

Example 21.2. $\mathbb{R}^n$ is contractible: consider the homotopy $\Gamma(s, \mathbf{x}) = (1-s)\mathbf{x}$. We have $\Gamma(0, \cdot) = \mathrm{Id}$ and $\Gamma(1, \cdot) = e_{\mathbf{0}}$. Similarly, any open or closed ball in $\mathbb{R}^n$ is contractible: given $\mathbf{p} \in \mathbb{R}^n$ and $r > 0$, the identity map on the closed ball $X = \{\mathbf{x} \in \mathbb{R}^n \mid d(\mathbf{p}, \mathbf{x}) \leq r\}$ can be homotoped to the trivial map by

$$\Gamma(s, \mathbf{p} + \mathbf{x}) = (1-s)\mathbf{x} + \mathbf{p}. \tag{21.1}$$

In fact, this gives a broad class of contractible spaces: we say that $X \subset \mathbb{R}^n$ is *convex from a point* (or *star-shaped*) if there exists $\mathbf{p} \in X$ such that the line segment from $\mathbf{p}$ to $\mathbf{x}$ is contained in X for every $\mathbf{x} \in X$. If X is convex from the point $\mathbf{p}$, then (21.1) gives a homotopy between Id_X and $e_{\mathbf{p}}$.

Remark. The fact that open balls are contractible emphasises the fact that for nice spaces that look locally as Euclidean spaces (such spaces are called *manifolds*) at a local level, everything is homotopic, and that homotopy is really a *global* theory, which captures large-scale properties of spaces and maps.

Definition 21.3. Recall that a *graph* is a finite or countable collection of vertices (which we may think of as lying in $\mathbb{R}^n$) together with a collection of edges joining certain pairs of vertices. A *cycle* is a collection of edges that forms a closed loop, and a graph without cycles is a *tree*.

Proposition 21.4. *Every finite tree is contractible.*

Proof. Let us use induction in the number of edges. The tree with zero edges is a point and hence contractible. Now let $\mathcal{T}$ be a tree with n edges. Removing an edge e makes the rest of the tree disconnected since otherwise the endpoints of e could be connected in $\mathcal{T} \setminus e$ and adding e one would get a cycle. Thus $\mathcal{T}$ with the interior of e removed is the union of two disjoint trees, each having fewer that n edges. By inductive hypothesis each of the two parts can be contracted to the corresponding endpoint of the removed edge. Combining these contractions with contraction of the edge to, say, its midpoint, completes the argument. □

As it turns out, countable trees are also contractible. This will be proven later in the course of our studies of certain graphs as geometric objects related to certain groups.

EXAMPLE 21.5. Let $X = \mathbb{R}^2 \setminus \{\mathbf{0}\}$; then as we will shortly see, X is *not* contractible.

We have observed that homeomorphic spaces have the same fundamental group; however, there is a weaker condition under which two spaces must have the same fundamental group. The condition that X and Y be homeomorphic may be stated as the existence of maps $f\colon X \to Y$ and $g\colon Y \to X$ such that $f \cdot g = \mathrm{Id}_Y$ and $g \cdot f = \mathrm{Id}_X$. Since the fundamental group is stated not in terms of paths but rather of homotopy classes of paths, it makes sense to weaken these equalities to homotopic equivalences.

DEFINITION 21.6. Two metric spaces X and Y are *homotopic* (or *homotopy equivalent*) if there exist maps $f\colon X \to Y$ and $g\colon Y \to X$ such that $f \cdot g \sim \mathrm{Id}_Y$ and $g \cdot f \sim \mathrm{Id}_X$.

EXAMPLE 21.7. Any contractible space is homotopic to a point; to see this, let X be contractible, fix a point $p \in X$, and let $Y = \{p\}$. Then defining $f\colon X \to \{p\}$ by $e_p\colon x \to p$ and $g\colon \{p\} \to X$ as the inclusion map $g(p) = p$, we see that $f\colon g = \mathrm{Id}_Y$ and $g\colon f = e_p \sim \mathrm{Id}_X$, where the last statement follows from the definition of contractibility.

EXAMPLE 21.8. Writing S^1 for the circle $\{\mathbf{x} \in \mathbb{R}^2 \mid \|\mathbf{x}\| = 1\}$, we see that the punctured plane $\mathbb{R}^2 \setminus \{\mathbf{0}\}$ is homotopic to S^1. Indeed, we may let $f\colon \mathbb{R}^2 \setminus \{\mathbf{0}\} \to S^1$ be the radial projection $f(\mathbf{x}) = \mathbf{x}/\|\mathbf{x}\|$ and $g\colon S^1 \to \mathbb{R}^2 \setminus \{\mathbf{0}\}$ be the inclusion map $g(\mathbf{x}) = \mathbf{x}$: then $f \circ g = \mathrm{Id}_{S^1}$, and $g \circ f$ is homotopic to $\mathrm{Id}_{\mathbb{R}^2 \setminus \{\mathbf{0}\}}$ via the linear homotopy

$$\Gamma(s, \mathbf{x}) = s\mathbf{x} + (1-s)\frac{\mathbf{x}}{\|\mathbf{x}\|}.$$

Similarly, one may show that both the Möbius strip and the cylinder are homotopic to the circle, and hence to each other (since homotopy is an equivalence relation), although they have the same dimension but not homeomorphic

Fundamental group that is an invariant of homeomorphism between spaces by construction in fact possesses a stronger invariance property.

PROPOSITION 21.9. *Show that if X and Y are homotopically equivalent, then $\pi_1(X)$ and $\pi_1(Y)$ are isomorphic.*

PROOF. By Proposition 20.6 one simply needs to check that $(g \cdot f)_*$ is an isomorphism between $\pi_1(X, p)$ and $\pi_1(X, g(f(p))$. This is very similar to the argument in the proof of Proposition 20.5. Let Γ be the homotopy between Id and $g \cdot f$ and $\alpha(t) = \Gamma(t, p)$. Let us associates to a path γ at p the path $\alpha^{-1} \star \gamma \star \alpha$ at $g(f(p))$. Γ establishes homotopy between these two paths. Reversing the direction of Γ one gets homotopy in the opposite direction. □

b. The fundamental group of the circle. We now describe the homotopy classes of maps from S^1 to itself, which also lets us compute its fundamental group $\pi_1(S^1)$. We need to formalise the notion of a map $f\colon S^1 \to S^1$ as "wrapping the circle around itself". To do this, we recall that the circle S^1 can also be obtained (algebraically and topologically, if not geometrically) as the factor space $\mathbb{R}/\mathbb{Z}$. Thus any continuous map $F\colon \mathbb{R} \to \mathbb{R}$ projects a map $f\colon \mathbb{R}/\mathbb{Z} \to \mathbb{R}/\mathbb{Z}$ provided it is well-defined—that is, provided $F(x) - F(y) \in \mathbb{Z}$ whenever $x - y \in \mathbb{Z}$. Furthermore, for any such map, the quantity $F(x+1) - F(x)$ varies continuously in x and takes integer values, and hence is independent of x; it is called the *degree* of the map f and denoted by $\deg f$. We may think of the degree as the number of times f wraps the circle around itself.

Does it go in the other direction? Do we get *every* map of the circle this way? That is, given a continuous map $f\colon \mathbb{R}/\mathbb{Z} \to \mathbb{R}/\mathbb{Z}$, can we produce a continuous map $F\colon \mathbb{R} \to \mathbb{R}$ such that the following diagram commutes?

$$\begin{tikzcd}
\mathbb{R} \arrow[r, "F"] \arrow[d, "\pi"] & \mathbb{R} \arrow[d, "\pi"] \\
\mathbb{R}/\mathbb{Z} \arrow[r, "f"] & \mathbb{R}/\mathbb{Z}
\end{tikzcd} \tag{21.2}$$

Here $\pi\colon \mathbb{R} \to \mathbb{R}/\mathbb{Z}$ is the natural projection $\pi(x) = x + \mathbb{Z}$.

It turns out that such a map F does indeed exist; we call this the *lift* of f. To produce F, we begin by specifying $F(0)$ as any element of $f(0+\mathbb{Z})$. Once this is done, the requirement that F be continuous specifies it uniquely; fixing a small $\varepsilon > 0$ and considering any $y \in (-\varepsilon, \varepsilon)$, we must choose $F(y)$ to be the element of $f(y+\mathbb{Z})$ that lies nearest to $F(0)$. Continuing in this manner, we can define F on $(-2\varepsilon, 2\varepsilon)$, $(-3\varepsilon, 3\varepsilon)$, and so on.

Given a lift $F\colon \mathbb{R} \to \mathbb{R}$, we see that $F(1) - F(0)$ is the *degree* of f defined above. Observe that since $x \mapsto F(x+1) - F(x)$ is continuous and integer-valued, we must have $F(x+1) - F(x) = F(1) - F(0)$ for all $x \in \mathbb{R}$, and thus the choice of 0 to determine the degree is (once again) irrelevant, although *some* choice was necessary.

Proposition 21.10. *If $f\colon S^1 \to S^1$ and $g\colon S^1 \to S^1$ are homotopic, then* $\deg f = \deg g$.

Proof. If f and g are homotopic, then their lifts F and G are homotopic as well. Let Γ be this homotopy, and observe that $\Gamma(s, \cdot)\colon S^1 \to S^1$ varies continuously in s, so $\Gamma(s,1) - \Gamma(s,0)$ varies continuously in s as well. Since it takes integer values, it must be constant. □

We can easily define a circle map with any given degree: for any $n \in \mathbb{Z}$, let $E_n\colon S^1 \to S^1$ be the linear map $E_n(x+\mathbb{Z}) = nx + \mathbb{Z}$—that is, E_n is the projection of the map $x \mapsto nx$ from the real line onto the circle. In fact, from the point of view of homotopy, these maps are all there is.

Proposition 21.11. *Every circle map of degree n is homotopic to E_n.*

Proof. Let $f\colon S^1 \to S^1$ have degree n, and let $F\colon \mathbb{R} \to \mathbb{R}$ be its lift to the real line. Consider the linear homotopy

$$\Gamma(s, x) = (1 - s)F(x) + snx, \tag{21.3}$$

and observe that $\Gamma(0, x) = F(x)$ and $\Gamma(1, x) = nx$. Furthermore, we have

$$\begin{aligned}\Gamma(s, x + 1) &= (1 - s)F(x + 1) + sn(x + 1)\\ &= (1 - s)(F(x) + n) + snx + sn\\ &= \Gamma(s, x) + n,\end{aligned}$$

and so $\Gamma(s, \cdot)\colon \mathbb{R} \to \mathbb{R}$ projects to a well-defined continuous map $\gamma(s, \cdot)\colon \mathbb{R}/\mathbb{Z} \to \mathbb{R}/\mathbb{Z}$. Since $\gamma(0, x + \mathbb{Z}) = f(x + \mathbb{Z})$ and $\gamma(1, x + \mathbb{Z}) = nx + \mathbb{Z}$, we see that γ is the desired homotopy. □

Corollary 21.12. *The fundamental group of the circle is* $\pi_1(S^1) = \{[E_n]\} \cong \mathbb{Z}$*, where the group operation is* $[E_n] \star [E_m] = [E_{n+m}]$.

Proof. A loop in X with base point p can be written as a continuous map $S^1 = \mathbb{R}/\mathbb{Z} \to X$ which maps $0 + \mathbb{Z}$ to p. Taking $p = 0 + \mathbb{Z} \in S^1$, we see that E_n has this property as well, and so any loop in S^1 of degree n with base point $0 + \mathbb{Z}$ is homotopic to E_n via the homotopy coming from (21.3). □

Remark. In the end, this result is purely topological, and applies to any space homotopic to the circle—a punctured plane, a Möbius strip, a cylinder, etc. However, in order to prove it, we found it beneficial to consider a very particular representative from this homotopy class—namely, the factor circle $\mathbb{R}/\mathbb{Z}$, which carries an extra algebraic structure that was essential in the proof.

c. Direct products, spheres, and bouquets of circles. Upon observing that the torus $\mathbb{R}^2/\mathbb{Z}^2$ is the direct product of two copies of S^1, we can finally complete our description of the fundamental group of the torus, using the following result.

Theorem 21.13. *Let X and Y be path-connected metric spaces. Then* $\pi_1(X \times Y) \cong \pi_1(X) \times \pi_1(Y)$.

Proof. Fix base points $x_0 \in X$ and $y_0 \in Y$, and let $P_X\colon (x, y) \mapsto x$ and $P_Y\colon (x, y) \mapsto y$ be the natural projections from $X \times Y$ to X and Y, respectively.

Now if γ_X and γ_Y are loops in X and Y with base points x_0 and y_0, then they determine a unique loop in $X \times Y$ with base point (x_0, y_0) by

$$\gamma(t) = (\gamma_X(t), \gamma_Y(t)). \tag{21.4}$$

Conversely, every loop γ in $X \times Y$ based at (x_0, y_0) determines loops in X and Y based at x_0 and y_0 by the projections $P_X(\gamma)$ and $P_Y(\gamma)$. This map also works for homotopies, so it defines a map $\pi_1(X \times Y) \to \pi_1(X) \times \pi_(Y)$; similarly, the map (21.4) defines a map $\pi_1(X) \times \pi_1(Y) \to \pi_1(X \times Y)$. Writing down definitions in a straightforward way one sees that these maps are homomorphisms and are inverses of each other, which proves the result. □

Corollary 21.14. *The n-dimensional torus $\mathbb{R}^n/\mathbb{Z}^n$ has fundamental group $\mathbb{Z}^n$.*

Corollary 21.14 is a concrete example of an important general scheme: many interesting spaces are obtained as X/G, where X is a topological space and G is a group acting on X. We will see in the next lecture that for "simple" enough X, and a "nice" action of G one can obtain $\pi_1(X/G) = G$.

Definition 21.15. If $\pi_1(X)$ is trivial, we say that X is *simply connected*.

Obviously, every contractible space is simply connected. The converse fails, however: being contractible is a stronger property than being simply connected. A simplest but fundamental example is provided by the spheres. To see this, consider the sphere S^2. The sphere is not contractible (this looks clear intuitively but requires a proof that will be given later), but has trivial fundamental group.

Proposition 21.16. *The sphere S^2 is simply connected.*

Proof. We observe that if x is any point on the sphere, then $S^2 \setminus \{x\}$ is homeomorphic to $\mathbb{R}^2$ via stereographic projection. Since $\mathbb{R}^2$ is contractible, any curve on $\mathbb{R}^2$ is homotopic to a point, and in particular, any loop $\gamma\colon S^1 \to S^2$ which misses a point (that is, $\gamma(S^1) \neq S^2$) can be homotoped to a point by using stereographic projection from a point $x \in S^2 \setminus \gamma(S^1)$.

However, one must deal with the fact that there are continuous and surjective functions $\gamma\colon S^1 \to S^2$—these so-called *Peano curves* cannot be immediately dealt with in the above fashion. They turn out not to cause too much trouble, as any curve γ is homotopic to a piecewise smooth approximation. In the plane, this can be seen by letting $\gamma\colon [0,1] \to \mathbb{R}^2$ be any curve and considering the piecewise linear approximations γ_n that are defined by the property $\gamma_n(k/n) = \gamma(k/n)$ for integers $0 \leq k \leq n$, and are linear in between these points. We have $\gamma_n \sim \gamma$, and a similar construction works on the sphere, replacing line segments with arcs of great circles. Since the curves γ_n *cannot* cover the entire sphere (being piecewise smooth), this suffices to show that $\pi_1(S^2) = \{e\}$. □

This argument extends straightforwardly to higher dimensions: for $n \geq 2$

$$\pi_1(S^n) = \{e\}.$$

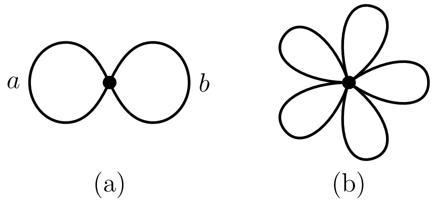


Figure 4.3. Bouquets of circles.

A more complicated example is given by the "bouquets" of n circles shown in Figure 4.3. When $n = 2$, we get the figure-eight shape shown in Figure 4.3(a); part (b) of the diagram shows the case $n = 5$. What is the fundamental group of these spaces?

The case $n = 1$ is just the circle S^1, where the key to deciphering the fundamental group was to classify curves in terms of how often they looped around the circle. Thus we expect a similar classification to be important here, and indeed, given a loop γ, we may profitably ask what the degree of γ is on each "leaf" of the bouquet (to mix our metaphors a little). However, we soon find that this is not quite sufficient for a complete understanding. Labeling the leaves of the figure-eight as shown in Figure 4.3(a), we write a for the loop that goes around the left-hand leaf in the clockwise direction, and a^{-1} for the loop that goes around it counter-clockwise, and similarly for b and b^{-1}. Then the loop $\gamma = a \star b \star a \star b^{-1}$ has degree 0 around both leaves, but is *not* homotopic to the identity.

This indicates that the fundamental group in this case is more complicated; in particular, it is non-abelian. We will return to this example in the next lecture, where we will once again be able to "lift" loops from the figure-eight to a certain *covering space*, just as we lifted loops from the circle to the real line. However, it is not immediately obvious what this covering space of the figure-eight should be, and so we will first need to see what happens when we "unfold" this picture.

Lecture 22. Monday, October 26

a. Fundamental groups of bouquets of circles. Let $B_n(S^1)$ denote the bouquet of n circles shown in Figure 4.3; we will focus our attention on the case $n = 2$, as the situation for larger values of n is analogous.

What is the fundamental group of the "figure eight" shape $B_2(S^1)$? So far, we have used three different techniques to compute the fundamental group of a space X:

(1) Show that $\pi_1(X)$ is trivial by showing that any loop can be contracted to a point.
(2) In the case $X = S^1$, use the fact that we can lift loops to paths in $\mathbb{R}$ to define the *degree* of a loop, and show that this defines an isomorphism between $\pi_1(S^1)$ and $\mathbb{Z}$.
(3) Show that X is homotopic to a space whose fundamental group is known, or obtain X as the direct product of such spaces.

For the figure eight $X = B_2(S^1)$, the first and the third methods are useless, and so we must look more closely at the second. As we did for the circle, we want to exhibit a standard family of loops in $B_2(S^1)$ that carries a clear group structure, and which is universal in the sense that every loop in $B_2(S^1)$ is homotopic to something from this family.

The first step in obtaining this family for the circle was to use the fact that the circle is a factor space $\mathbb{R}/\mathbb{Z}$, and that loops on the circle can be lifted to paths in $\mathbb{R}$. The standard projection $\pi\colon \mathbb{R} \to S^1 = \mathbb{R}/\mathbb{Z}$ can be written in a number of different forms.

(1) If we think of the circle as the interval $[0, 1)$, where the missing endpoint 1 is identified with 0, then $\pi(x) = x \pmod 1$.
(2) If we think of the circle as the factor space $\mathbb{R}/\mathbb{Z}$, so that points on the circle are equivalence classes in $\mathbb{R}$, then $\pi(x) = x + \mathbb{Z}$.
(3) If we think of the circle as the unit circle in $\mathbb{C}$, then $\pi(x) = e^{2\pi i x}$.

Whichever model of the circle we use, the key property of the projection π is that it is a local homeomorphism—that is, for every $x \in \mathbb{R}$ there exists a neighbourhood $U \ni x$ such that $\pi\colon U \to \pi(U) \subset S^1$ is a homeomorphism. In particular, if $V \subset S^1$ is any sufficiently small neighbourhood, the preimage $\pi^{-1}(V) \subset \mathbb{R}$ is a disjoint union of open sets, each of which is homeomorphic to V via the action of π.

In the language of topology, we say that π is a *covering map*, and $\mathbb{R}$ is a *covering space*; in fact, it is what is called the *universal covering space*. We will postpone general definitions of these concepts, and focus instead on the techniques involved.

Example 22.1. The natural projection $S^2 \to \mathbb{R}P(2)$ that takes $\mathbf{x}$ to $\{\mathbf{x}, -\mathbf{x}\}$ is also a local homeomorphism (and indeed, a covering map). In this case, however, each point has only two preimages, rather than countably many, as is the case for the circle.

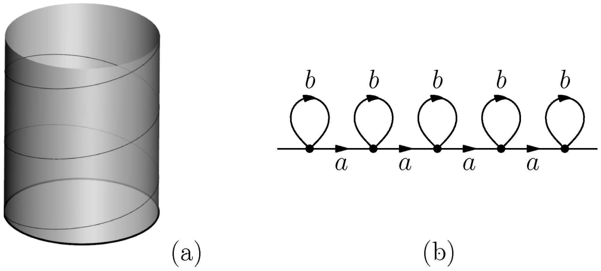


Figure 4.4. Unwinding the circle and the figure-eight.

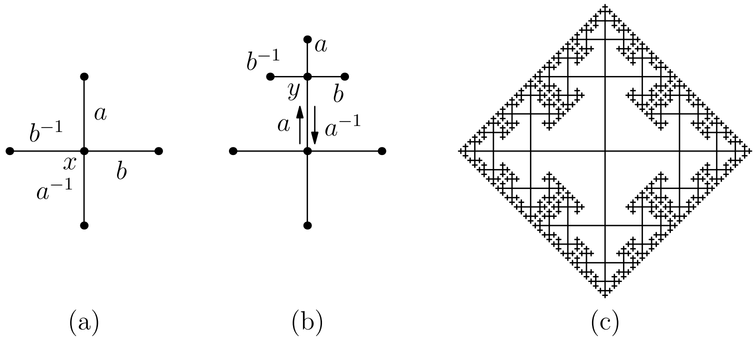
One visualisation of the covering map π is shown in Figure 4.4(a). Topologically, the helix $\{(\cos t, \sin t, t) \mid t \in \mathbb{R}\}$ is equivalent to the real line, and the projection $(x, y, z) \mapsto (x, y)$ is a local homeomorphism from the helix to the circle.

If we unwind just one of the (topological) circles in $B_2(S^1)$, say a, then we obtain the space X shown in Figure 4.4(b); the circle labeled by a unwinds into a copy of $\mathbb{R}$, just as S^1 did, but now the resulting line has a circle corresponding to b attached to every integer value. There is a natural projection from X back down to $B_2(S^1)$; however, in the end X is not quite the space we were after. Recall that one of the key tools in our analysis of $\pi_1(S^1)$ was the fact that the homotopy type of a loop in S^1 only depended on the endpoints of its lift to a path in $\mathbb{R}$. In particular, this required every loop in $\mathbb{R}$ to be homotopic to the trivial loop; in other words, it was essential that $\mathbb{R}$ be simply connected. The space X is *not* simply connected, and so we will run into difficulties if we try to study $\pi_1(B_2(S^1))$ using X.

Thus to obtain the proper covering space for $B_2(S^1)$, we must unwind X still further until we have something simply connected. The space we get is to be locally homeomorphic to $B_2(S^1)$—that is, every point must either have a neighbourhood that is a segment of a path or be a vertex from which four paths emanate. This means that the space we are looking for is a graph in which every vertex has degree 4. Furthermore, in order to be simply connected, it cannot have any loops, and hence must be a tree.

This is enough to describe the space completely—see Figure 4.5. Let p be the point at which that two circles in $B_2(S^1)$ intersect. We construct the universal covering space, which we call Γ_4, and the covering map $\pi\colon \Gamma_4 \to B_2(S^1)$ by beginning with a single preimage x of the point p, which is the centre of the cross in Figure 4.5(a). There are four edges emanating from x, which correspond to the paths a, b, a^{-1}, b^{-1}; at the other end of each of these edges is another preimage of p, distinct from the first one.

Consider the point $y \in \Gamma_4$ shown in Figure 4.5(b); this point is the preimage of p lying at the other end of the edge labeled a. The loop a in $B_2(S^1)$ corresponds to following this edge from x to y; following the edge in the reverse direction, from y to x, corresponds to the loop a^{-1}. There



(a) (b) (c)

FIGURE 4.5. An infinite tree of homogeneous degree 4.

must be three other edges emanating from y, and they are labeled a, b, b^{-1}, as shown.

Similarly, each of the other three vertices we constructed at the first step must be the source of three more edges; once these are all drawn, we have five vertices of degree 4 and twelve vertices of degree 1. Each of these twelve vertices must likewise be the source of three further edges, so that it has one edge corresponding to each of the four labels a, b, a^{-1}, b^{-1}; this process continues *ad infinitum*. Thus Γ_4 is an infinite tree of the sort shown in Figure 4.5(c); observe that at every step, the vertices we add are disjoint from those that came before and from each other, since otherwise we would produce a loop.

REMARK. The lengths of the edges of Γ_4 are irrelevant for topological questions, which is what we are interested in. However, the geometric nature of Figure 4.5 is worth noting. Edges further away from x are drawn to be shorter; in particular, if we let $n(z)$ denote the minimum number of edges we must move along to reach x from a vertex z, then the edges emanating from z are drawn with length $2^{-(n(z)+1)}$. This lets us embed Γ_4 in the unit disc, and results in a fractal-like pattern near the edges of the disc that is reminiscent of some of M.C. Escher's artwork. Recalling that these drawings are based on the unit disc model of the hyperbolic plane, we may suspect that there is some connection between Γ_4 and hyperbolic geometry. This is indeed the case, although we shall not pursue the connections at the present time.

The projection map $\pi\colon \Gamma_4 \to B_2(S^1)$ is defined in the obvious way: every vertex of Γ_4 is mapped to p, and every edge is mapped to the loop corresponding to its label. In particular, the restriction of π to any small neighbourhood $U \subset \Gamma_4$ is a homeomorphism between U and its image. This is the key to the following result, which says that we can lift paths from $B_2(S^1)$ to Γ_4.

PROPOSITION 22.2. *Let $\gamma\colon [0,1] \to B_2(S^1)$ be a continuous path such that $\gamma(0) = \gamma(1) = p$. Then there exists a unique continuous path $\tilde{\gamma}\colon [0,1] \to \Gamma_4$ such that $\tilde{\gamma}(0) = x$ and $\pi \circ \tilde{\gamma} = \gamma$.*

PROOF. The idea is this: because π is a local homeomorphism, there exists a neighbourhood $V \ni p = \gamma(0)$ such that if we write U for the connected component of $\gamma^{-1}(V)$ containing x, then $\pi\colon U \to V$ is a homeomorphism. Write $\varphi\colon V \to U$ for the inverse of this homeomorphism; then taking $\varepsilon > 0$ such that $\gamma(t) \in V$ for all $t \in (0, \varepsilon)$, the unique path $\tilde{\gamma}$ satisfying $\pi \circ \tilde{\gamma} = \gamma$ on $(0, \varepsilon)$ is $\tilde{\gamma} = \varphi \circ \gamma$. Repeating this construction and using compactness of the unit interval, one obtains the result.

To make this a little more precise, let $r > 0$ be such that every ball of radius r in $B_2(S^1)$ is simply connected. For example, if the two (topological) circles in Figure 4.3 each have diameter 1, then any $r < 1/2$ will suffice, as a ball of radius r in $B_2(S^1)$ is not big enough to contain a complete loop. Now since $\gamma\colon [0,1] \to B_2(S^1)$ is continuous and $[0,1]$ is compact, γ is uniformly continuous, so there exists $\varepsilon > 0$ such that $d(\gamma(s), \gamma(t)) < r$ whenever $|s - t| < \varepsilon$.

Now for every $t \in [0,1]$, we may write $B(\gamma(t), r)$ for the ball of radius r centred at $\gamma(t)$ in $B_2(S^1)$, and we observe that if U is a connected component of $\pi^{-1}(B(\gamma(t), r))$, then $\pi|_U$ is a homeomorphism from U to $B(\gamma(t), r)$. Given ε as above, we see that once $\tilde{\gamma}(t)$ is chosen, the connected component is fixed, and so there exists a unique lift of γ to $\tilde{\gamma}$ on $(t - \varepsilon, t + \varepsilon)$.

Thus we start with $\tilde{\gamma}(0) = x$, and observe that this determines a unique $\tilde{\gamma}$ on $[0, \varepsilon)$. Applying the above argument to $t = \varepsilon/2$, we get a unique $\tilde{\gamma}$ on $[0, 3\varepsilon/2)$; applying it to $t = \varepsilon$, we get $[0, 2\varepsilon)$, and so on. Within a finite number of steps, we have determined $\tilde{\gamma}$ uniquely on $[0,1]$. □

In fact, the above argument lets us lift more than just paths. We can also lift homotopies, which gives a direct link between $\pi_1(B_2(S^1))$ and $\pi_1(\Gamma_4)$.

PROPOSITION 22.3 (Principle of covering homotopy). *If $\gamma_0, \gamma_1\colon [0,1] \to B_2(S^1)$ are continuous loops based at p and $\Gamma\colon [0,1] \times [0,1] \to B_2(S^1)$ is a homotopy from γ_0 to γ_1, then there exists a unique lift of Γ to a homotopy from $\tilde{\gamma}_0$ to $\tilde{\gamma}_1$, the lifts guaranteed by Proposition 22.2. Furthermore, the lifted homotopy is a homotopy relative to endpoints.*

PROOF. Apply Proposition 22.2 to $\Gamma(s, \cdot)$ for each $0 \leq s \leq 1$. We can (indeed, must) hold the endpoints fixed because the set of preimages of the base point is discrete. □

In order to describe the homotopy classes of loops in $B_2(S^1)$, we need to give a list of standard representatives, along with a complete homotopy invariant that identifies which element from the list corresponds to a given loop. For the circle S^1, the homotopy invariant was the degree of a loop, which tracked how many times the loop went around the circle; upon being lifted to $\mathbb{R}$, this became the total displacement of the lifted path.

For the torus $\mathbb{T}^2$, the homotopy invariant was a pair of integers specifying the degrees of the projections of the lifted path; this integer pair corresponded to the second endpoint of the lifted path on the integer lattice, which was the lift of the base point.

For the figure eight $B_2(S^1)$, we may likewise expect that the homotopy invariant will be the second endpoint of the lifted path on the preimage of the base point under the projection map π. This preimage is the set of vertices of Γ_4, and every such vertex may be specified by the sequence of edges we follow to reach it from the "centre" of Γ_4. To see this, we first consider a finite sequence of symbols from the set $\{a, a^{-1}, b, b^{-1}\}$—such a sequence is called a *word*. If a word w has the property that the symbols a and a^{-1} never appear next to each other, and similarly for b and b^{-1}, then w is called a *reduced word*. Any word can be transformed into a reduced word by repeatedly cancelling all adjacent pairs of inverses. For brevity of notation, we abbreviate aa as a^2, aaa as a^3, and so on; thus $aaab^{-1}aba^{-1}a^{-1}bbb$ may be written $a^3b^{-1}aba^{-2}b^3$.

Now labeling the edges of Γ_4 with the symbols a, a^{-1}, b, b^{-1} (see Figure 4.5), we associate to each reduced word w the following path in Γ_4. Beginning at the centre x, follow the edge corresponding to the first symbol in w; once the second vertex of this edge is reached, follow the edge corresponding to the second symbol in w, and so on. Observe that because w never contains a symbol followed by its inverse, we will never backtrack. Parametrising this path with uniform speed, one associates to each reduced word a standard path in Γ_4.

This exhibits a one-to-one correspondence between reduced words and standard paths; there is also a one-to-one correspondence between standard paths and vertices in Γ_4. By Proposition 22.3, any two homotopic loops in $B_2(S^1)$ lift to paths in Γ_4 that are homotopic relative to endpoints. In particular, they correspond to the same reduced word.

Write F_2 for the set of all reduced words in the symbols a, b, a^{-1}, b^{-1}. We have now shown that the process of lifting loops in $B_2(S^1)$ to paths in Γ_4 gives a map $\psi \colon \pi_1(B_2(S^1)) \to F_2$. The previous paragraph shows that ψ is well-defined, and it is obvious that ψ is surjective.

Furthermore, ψ is one-to-one. To see this, we must show that any two loops in $B_2(S^1)$ that lift to paths with the same endpoint in Γ_4 are actually homotopic.

LEMMA 22.4. *Every loop based at p in $B_2(S^1)$ is homotopic to one of the standard loops described above.*

PROOF. As in the proof of Proposition 22.2, let $r > 0$ be such that every ball of radius r in $B_2(S^1)$ is contractible. Given a loop γ based at p, let $\varepsilon > 0$ be such that $\gamma((t-\varepsilon, t+\varepsilon))$ is contained in such a ball for every $\varepsilon > 0$. (This uses uniform continuity of γ.)

Now consider the set $E = \{t \in [0,1] \mid \gamma(t) = p\}$, which contains all parameter values t at which γ returns to the basepoint. Because γ is continuous, $E = \gamma^{-1}(p)$ is closed, and so $[0,1] \setminus E$ is open—in particular, this complement is a countable union of open intervals. Denote these intervals by (s_n, t_n), and observe that if $|t_n - s_n| < \varepsilon$, then $\gamma|_{[s_n,t_n]}$ is homotopic to the constant map $t \mapsto p$.

This shows that γ is homotopic to a loop γ_1 with the property that $[0,1] \setminus \gamma_1^{-1}(p)$ is a *finite* union of open intervals (since there are at most $1/\varepsilon$ values of n such that $t_n - s_n \geq \varepsilon$). Again, denote these by (s_n, t_n), and observe that each $\gamma|_{[s_n,t_n]}$ is a loop on a circle (which corresponds to either a or b), and hence is homotopic to one of the standard representatives from $\pi_1(S^1)$.

We have shown that γ is homotopic to a concatenation of standard loops on circles; a straightforward reparametrisation shows that such a concatenation is homotopic to one of the standard loops described above. □

Lemma 22.4 shows that ψ is a bijection between $\pi_1(B_2(S^1))$ and F_2. In order to complete our description of the fundamental group, it remains to put a group structure on F_2 and show that ψ is in fact an isomorphism.

As with paths, words can be multiplied by concatenation; in order to obtain a reduced word, we must then cancel adjacent inverse symbols. Thus, for example,

$$(aba^2b) \star (b^{-1}a^{-1}b) = aba^2bb^{-1}a^{-1}b = aba^2a^{-1}b = abab.$$

This gives a group structure on F_2, which we call the *free group* with two generators. It is immediate that ψ is a homomorphism, since the operation in both groups is concatenation; upon observing that everything we have done generalises immediately to $B_n(S^1)$ for $n > 2$, we have the following result.

Theorem 22.5. *The fundamental group of the bouquet of n circles is isomorphic to the free group with n generators: $\pi_1(B_n(S^1)) \cong F_n$.*

Remark. The free group is in some sense the most non-abelian group possible, in that it has the fewest relations—none. One can show that $[F_2, F_2]$ is the set of words in which a and a^{-1} occur with equal frequency, and also for b and b^{-1}. Furthermore, one finds that $F_2/[F_2, F_2] \cong \mathbb{Z}^2$, so $\mathbb{Z}^2$ is the *abelianisation* of F_2.

CHAPTER 5

From groups to geometric objects and back

Lecture 23. Wednesday, October 28

a. Cayley graphs. We constructed the graph Γ_4 as a covering space of $B_2(S^1)$, and thought of it as somehow an "unfurling" of that space to eliminate all loops. We then found that apart from this topological meaning, it also had algebraic significance by leading us to the free group F_2. In fact, we can also go in the other direction and construct Γ_4 from the purely algebraic properties of F_2: this is an example of a very general construction, which we now describe.

Let G be a finitely generated group—that is, a group for which there exists a finite set $\mathcal{B} = \{g_1, \dots, g_n\} \subset G$ such that the only subgroup of G containing $\mathcal{B}$ is G itself. The elements g_i are *generators* of G; we say that the set of generators is *symmetric* if $g^{-1} \in \mathcal{B}$ whenever $g \in \mathcal{B}$. If G is finitely generated, then it has a symmetric finite set of generators; every element of G can be written as a finite product of elements from this set. (We can modify the following construction to avoid making this symmetry assumption, but this way simplifies certain aspects of the notation.)

Given a group G and a symmetric set of generators $\{g_1, \dots, g_n\}$, we construct the *Cayley graph* as follows. Begin with a single vertex corresponding to the identity element e, and draw n edges emanating from e, corresponding to the n generators. Label each of the vertices at the other ends of these edges with the corresponding generator.

This is the first step of an iterative process; so far we have vertices corresponding to the group elements $e, g_1, \dots, g_n$. In the second step, we add vertices corresponding to each of the elements $g_i g_j$, and draw edges from g_i to $g_i g_j$.

The iterative process continues as described: after n steps, we have a collection of vertices that correspond to group elements represented by reduced words of length $\leq n$ in the set of generators. The same group element may be represented by different words: for example, if G is abelian then $g_1 g_2$ and $g_2 g_1$ give the same vertex of the Cayley graph. Two vertices corresponding to the elements g and g' are connected by an edge if and only if there exists a generator $g_i \in \mathcal{B}$ such that $g g_i = g'$.

By appending all possible generators to all possible reduced words of length exactly n, we obtain the next level of vertices. Once again, we stress that if G is not free in the generators $\mathcal{B}$, then there may be some non-trivial relationships—that is, different reduced words that correspond to the same vertex. In particular, since a reduced word describes a path along the graph, two different reduced words corresponding to the same vertex describe two different paths with the same endpoints, and thus give a loop in the Cayley graph.

EXAMPLE 23.1. For $G = F_2$ and $\mathcal{B} = \{a, b, a^{-1}, b^{-1}\}$, we start with a single vertex (the identity), add four vertices at the first step, and add 12 more at the second step. For $G = \mathbb{Z}^2$ and $\mathcal{B} = \{(1, 0), (0, 1), (-1, 0), (0, -1)\}$, we start with a single vertex, add four vertices at the first step, and then add only four more at the second step. This is because there are different combinations $g_i g_j$ which yield that same element of the group: for example, $(1, 0) + (0, 1) = (0, 1) + (1, 0)$. In this representation the Cayley graph of $\mathbb{Z}^2$ is the set of vertices and edges of the square lattice in the plane.

EXERCISE 23.1. (1) Describe the Cayley graph of $\mathbb{Z}^2$ for a different choice of two generators and their inverses.
(2) Show how to obtain the vertices and edges of a the triangular lattice as the Cayley graph of $\mathbb{Z}^2$.
(3) Show how to obtain the hexagonal lattice as the Cayley graph of $\mathbb{Z}^2$ by choosing a non-symmetric system of generators.

EXERCISE 23.2. Show that all Cayley graphs that appear in Exercise 23.1 are homotopy equivalent to the bouquet of countably many circles and calculate their fundamental group.

We now have a short dictionary between algebraic objects in the group G and graph-theoretic objects in the Cayley graph: group elements correspond to vertices, generators correspond to edges, words correspond to paths, relations correspond to loops, and freeness of the group is equivalent to the statement that the Cayley graph is a tree.

There is one aesthetically unpleasant part of all this. The Cayley graph depends not just on the group G, but also on the choice of generators. If we choose a different set of generators, we will obtain a different graph; is there anything intrinsic about all this, or do the generators really play a fundamental role? Exercise 23.2 indicate that there may be some similarities in the properties of those graphs.

Consider first the free group F_2 and its associated Cayley graph Γ_4. We can pass to the *boundary* of the Cayley graph by considering *infinite* words in the generators $\{a, b, a^{-1}, b^{-1}\}$, which correspond to infinite paths from the centre of the graph out to its edge. We will not go through the details here, but merely mention that one can introduce a natural topology on the space of infinite words under which it becomes a Cantor set, and has certain intrinsic properties. A similar construction can be carried out for more general groups, although we must be more careful because there may be infinite words whose corresponding paths do not escape to the edge of the graph.

b. Homotopy types of graphs. We now use the notion of homotopy equivalence to further our understanding of the topological structure of arbitrary graphs. Understanding the topological structure of graphs up to homeomorphism is quite a difficult proposition, as homeomorphisms preserve a great deal of combinatorial information about the graph, such as the degree of every vertex (although vertices of degree 2 can be absorbed into their adjoining edges). However, homotopy equivalence is a more flexible matter, and it turns out that we can say a great deal without exerting an unreasonable amount of effort. We begin by showing that the absence of loops implies contractibility (trivial homotopy type).

PROPOSITION 23.2. *Any tree (finite or infinite) is contractible.*

PROOF. Let $\mathcal{T}$ be a tree, and fix a vertex $v \in \mathcal{T}$. Put a metric on $\mathcal{T}$ under which every edge has length 1, and observe that for every point $x \in \mathcal{T}$ (whether at a vertex or not) there exists a path $p_x\colon [0,1] \to \mathcal{T}$ that runs from v to x with constant speed (with respect to the metric just imposed)—that is, $p_x(0) = v$ and $p_x(1) = x$. Because $\mathcal{T}$ does not contain any loops, p_x is unique; in particular, $p_x(t)$ varies continuously with x.

Now define a homotopy $\Gamma\colon [0,1] \times X \to X$ by $\Gamma(t,x) = p_x(1-t)$, and observe that $\Gamma(0,\cdot) = \mathrm{Id}_{\mathcal{T}}$ and $\Gamma(1,\cdot) = e_v$. It follows that $\mathrm{Id}_{\mathcal{T}} \sim e_v$, so $\mathcal{T}$ is contractible. □

In order to deal with graphs that contain loops, we want to find trees in these graphs to which we can apply Proposition 23.2.

DEFINITION 23.3. Given a connected graph $\mathcal{G}$, a *maximal tree* in $\mathcal{G}$ is a subgraph $\mathcal{T} \subset \mathcal{G}$ such that

(1) $\mathcal{T}$ is a tree;
(2) If $\mathcal{T} \underset{\neq}{\subsetneq} \mathcal{T}' \subset \mathcal{G}$, then $\mathcal{T}'$ is not a tree.

Equivalently, $\mathcal{T}$ contains every vertex of $\mathcal{G}$, and no edge of $\mathcal{G}$ can be added to $\mathcal{T}$ without forming a loop.

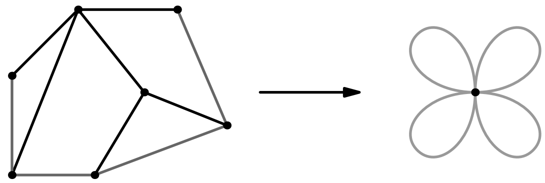


FIGURE 5.1. Collapsing a maximal tree.

Figure 5.1 shows an example of a maximal tree, as well as illustrating what we want to do with such trees: collapse them to a single point using Proposition 23.2, and then show that $\mathcal{G}$ is homotopic to the resulting bouquet of circles.

First, though, we need to verify that maximal trees exist. For finite trees, this is a straightforward induction; for infinite trees, things are a little more subtle, and we need the following abstract tools.

Definition 23.4. A *total ordering* on a set S is a binary relation $\preceq$ such that

(1) $x \preceq x$ for all $x \in S$;
(2) if $x \preceq y$ and $y \preceq x$, then $x = y$;
(3) if $x \preceq y$ and $y \preceq z$, then $x \preceq z$;
(4) for every $x, y \in S$, either $x \preceq y$ or $y \preceq x$.

If the first three of these hold, but we do not require the fourth, then $\preceq$ is called a *partial ordering*, and $(S, \preceq)$ is called a *partially ordered set*. If $x \preceq y$ or $y \preceq x$, we say that x and y are *comparable*; any two elements of a totally ordered set are comparable, but this is not necessarily true in a partially ordered set.

Example 23.5. The set of real numbers with the usual ordering is a totally ordered set.

Example 23.6. The set of natural numbers with divisibility ($a \preceq b$ if and only if a divides b) is a partially ordered set.

Example 23.7. Given any set X, the *power set* of X, denoted $\mathcal{P}(X)$, is the set of all subsets of X. Inclusion gives a natural partial ordering on $\mathcal{P}(X)$: $A \preceq B$ if and only if $A \subset B$.

Given a partially ordered set S, an element $x \in S$ is *maximal* if there does not exist $y \in S$, $y \neq x$, such that $x \preceq y$. (Note that this does *not* imply that $y \preceq x$ for all $y \in S$, since not every pair of elements is comparable.) We are often interested in finding maximal elements of certain partially ordered sets—for example, we often want to find a subset of X that is maximal among all subsets with a certain property. The standard tool for doing this is the following statement, which is equivalent to the Axiom of Choice.

Lemma 23.8 (Zorn's lemma). *Let S be a partially ordered set and suppose that every totally ordered subset of S has an upper bound. Then S itself has a maximal element.*

Remark. If we take S to be the collection of all linearly independent subsets of an arbitrary vector space V (ordered by inclusion), then Zorn's lemma lets us prove that every vector space has a basis. This statement, which is straightforward for finite-dimensional vector spaces, becomes more mysterious when we consider examples such as $\mathbb{R}$ as a vector space over $\mathbb{Q}$ (which has uncountable dimension).

Applying Zorn's lemma to trees in a graph $\mathcal{G}$, we can obtain a maximal tree.

Proposition 23.9. *Any graph (finite or infinite) contains a maximal tree.*

Proof. Let S be the collection of subgraphs of $\mathcal{G}$ that are also trees. Observe that if $C \subset S$ is totally ordered, then $\bigcup_{\mathcal{T}\in C} \mathcal{T} \in S$ is an upper bound for C, and so Zorn's lemma applies. □

Combining Propositions 23.2 and 23.9, we can classify graphs up to homotopy.

Theorem 23.10. *Every graph is homotopy equivalent to a bouquet of circles: in particular, if $\mathcal{T} \subset \mathcal{G}$ is a maximal tree and n is the number of edges of $\mathcal{G}$ that are* not *contained in $\mathcal{T}$, then $\mathcal{G} \sim B_n(S^1)$.*

Proof. Let $\mathcal{T} \subset \mathcal{G}$ be a maximal tree, whose existence is guaranteed by Proposition 23.9. (Note that $\mathcal{T}$ is not unique.) Recall that the quotient space $\mathcal{G}/\mathcal{T}$ is the result of identifying all points in $\mathcal{T}$ to a single point—that is,

$$\mathcal{G}/\mathcal{T} = \left(\bigcup_{x\in\mathcal{G}\setminus\mathcal{T}} \{x\} \right) \cup \{\mathcal{T}\},$$

and a sequence $x_n \in \mathcal{G}/\mathcal{T}$ converges to $x \in \mathcal{G}/\mathcal{T}$ if and only if $x_n \to x$ in $\mathcal{G}$ (if $x \in \mathcal{G} \setminus \mathcal{T}$) or if $\inf_{y\in\mathcal{T}} d(x_n, y) \to 0$ (if $x = \{\mathcal{T}\}$). Observe that $\mathcal{G}/\mathcal{T}$ is homeomorphic to $B_n(S^1)$, where n is the number of edges of $\mathcal{G}$ not contained in $\mathcal{T}$, and so it suffices to show that $\mathcal{G} \sim \mathcal{G}/\mathcal{T}$.

To show that two spaces are homotopic, we produce maps f and g in opposite directions such that $f \circ g$ and $g \circ f$ are both homotopic to the identity. In this case, we let $f\colon \mathcal{G} \to \mathcal{G}/\mathcal{T}$ be the canonical projection

$$f(x) = \begin{cases} \mathcal{T} & x \in \mathcal{T}, \\ x & x \notin \mathcal{T}. \end{cases}$$

To define $g\colon \mathcal{G}/\mathcal{T} \to \mathcal{G}$, we first fix a point $v \in \mathcal{T}$ and set $g(\mathcal{T}) = v$. Now we observe that every point $x \in \mathcal{G} \setminus \mathcal{T}$ lies on an edge γ of $\mathcal{G}$ that is not contained in $\mathcal{T}$; writing x, y for the endpoints of γ, we let p_x and p_y denote the (unique) paths in $\mathcal{T}$ from v to x and y, respectively. Let $g_\gamma\colon \gamma \to \mathcal{G}$ be the map that uniformly stretches the path γ to cover the path $p_x \star \gamma \star p_y^{-1}$, and define $g(x) = g_\gamma(x)$ for all $x \in \gamma$.

Now we see that $f \circ g\colon \mathcal{G}/\mathcal{T} \to \mathcal{G}/\mathcal{T}$ simply reparametrises each loop of the bouquet, and that $g \circ f\colon \mathcal{G} \to \mathcal{G}$ similarly stretches each edge in $\mathcal{G}/\mathcal{T}$ in a way that can be continuously homotoped to the identity, which completes the proof. □

If we consider the trivial group a free group with zero generator and remember that $\mathbb{Z}$ is the free group with one generator we obtain.

Corollary 23.11. *The fundamental group of any graph is a free group.*

c. Covering maps and spaces. We now extend some of the ideas used in the previous lectures; namely, the procedure of covering the torus by the plane, and the figure-eight by the Cayley graph of the free group. These are specific cases of a more general theory, which we now introduce.

Let C and X be metric spaces, and suppose that $\rho\colon C \to X$ is a continuous map such that for every $x \in X$, there exists a neighbourhood $U \ni x$ such that $\rho^{-1}(U)$ is a collection of disjoint neighbourhoods in C on which ρ is homeomorphic—that is, $\rho^{-1}(U) = \bigcup_i V_i$, where the V_i are disjoint open sets in X, and $\rho\colon V_i \to U$ is a homeomorphism for all i. (The union may be finite or countable.) Then ρ is called a *covering map*, and C is a *covering space* of X.

Remark. The number of connected components of the preimage $\rho^{-1}(U)$ takes discrete values and varies continuously in x. Thus it is locally constant, and hence constant everywhere if X is path connected, the only case to be considered forthwith.

We have already seen a number of examples of covering spaces and covering maps. For example, the map $f(t) = e^{2\pi i t}$ is a covering map from $\mathbb{R}$ to $S^1 \subset \mathbb{C}$. We also saw covering maps from $\mathbb{R}^2$ to $\mathbb{T}^2$, and from Γ_4 to the bouquet $B_2(S^1)$.

As we have shown in Proposition 20.6 given *any* metric spaces X, Y and a base point $x_0 \in X$, *any* continuous map $f\colon X \to Y$ induces a homomorphism $f_*\colon \pi_1(X, x_0) \to \pi_1(Y, f(x_0))$ that takes $[\gamma] \in \pi_1(X, x_0)$ to $[f \circ \gamma] \in \pi_1(Y, f(x_0))$. In particular, a covering map $\rho\colon C \to X$ induces a homomorphism $\rho_*\colon \pi_1(C, x_0) \to \pi_1(X, \rho(x_0))$.

Proposition 23.12. *If $\rho\colon C \to X$ is a covering map, then ρ_* is injective.*

Proof. The key is the *principle of covering homotopy* that generalizes Propositions 22.3

Lemma 23.13. *If $\rho\colon C \to X$ is a covering map, then:*

(1) Every path $\gamma\colon [0,1] \to X$ has a lift $\tilde{\gamma}\colon [0,1] \to C$—that is, $\rho \circ \tilde{\gamma} = \gamma$. Furthermore, $\tilde{\gamma}$ is unique up to the choice of $\tilde{\gamma}(0)$, which can be any point in $\rho^{-1}(\gamma(0))$.

(2) If $\Gamma\colon [0,1] \times [0,1] \to X$ is a homotopy such that $\Gamma(s,0) = \Gamma(s,1) = x_0$ for all $0 \le s \le 1$, then there exists a continuous homotopy $\tilde{\Gamma}\colon [0,1] \times [0,1] \to C$ such that $\rho \circ \tilde{\Gamma} = \Gamma$. Again, $\tilde{\Gamma}$ is unique up to the choice of $\tilde{\Gamma}(0,0)$.

Proof. It suffices to prove the second statement. Observe that given $x \in X$, there exists a neighbourhood U_x of x in X such that $\rho^{-1}(U_x)$ is a disjoint union of neighbourhoods in C on which ρ is a homeomorphism. In particular, there exists a family of maps $L_i\colon U_x \to C$ such that L_i is a homeomorphism onto its image (the L_i are the inverse branches of ρ).

Now given $(s,t) \in [0,1] \times [0,1]$, let $x = \Gamma(s,t) \in X$, and let $\varepsilon = \varepsilon(s,t) > 0$ be such that $\Gamma(s',t') \in U_x$ for all $|s - s'| < \varepsilon$ and $|t - t'| < \varepsilon$. Denote this range of (s',t') by $B(s,t,\varepsilon)$, and observe that $\rho(\tilde{\Gamma}(s',t')) = \Gamma(s',t')$ on $B(s,t,\varepsilon)$ if and only if $\tilde{\Gamma}(s',t') = L_i(\Gamma(s',t'))$ for some i.

We can now construct the lift of the homotopy. The open sets $B(s,t,\varepsilon(s,t))$ cover the compact unit square $[0,1]\times[0,1]$, and so by compactness there exists a finite set $\{(s_1,t_1),\dots,(s_n,t_n)\}$ such that the open sets $B_i = B(s_i,t_i,\varepsilon(s_i,t_i))$ cover the unit square.

Fix $\tilde{\Gamma}(0,0)\in\rho^{-1}(\Gamma(0,0))$; we claim that this determines $\tilde{\Gamma}(s,t)\in\rho^{-1}(\Gamma(s,t))$ uniquely for every $(s,t)\in[0,1]\times[0,1]$. Indeed, for any such (s,t), let $\eta(r) = (rs,rt)$ for $0\le r\le 1$, and for $1\le i\le n$ with $\eta([0,1])\cap B_i\neq\emptyset$, let

$$r_i^- = \inf\{r\in[0,1] \mid \eta(r)\in B_i,$$
$$r_i^+ = \sup\{r\in[0,1] \mid \eta(r)\in B_i.$$

Observe that there exist $i_1,\dots,i_m$ such that

$$0 = r_{i_1}^- < r_{i_2}^- < r_{i_1}^+ < r_{i_3}^- < r_{i_2}^+ < \cdots < r_{i_m}^- < r_{i_{m-1}}^+ < r_{i_m}^+ = 1.$$

There exists a unique inverse branch L_1 of ρ on $\Gamma(\eta([0,r_{i_1}^+)))$ such that $L_1(\Gamma(0,0)) = \tilde{\Gamma}(0,0)$. Similarly, there exists a unique inverse branch L_2 of ρ on $\Gamma(\eta((r_{i_2}^-,r_{i_2}^+)))$ such that $L_2(\Gamma(\eta(r_{i_1}^+))) = L_1(\Gamma(\eta(r_{i_1}^+)))$, and so on for $L_3,\dots,L_m$. Define $\tilde{\Gamma}(s,t) = L_m(\Gamma(s,t))$.

Now observe that $\tilde{\Gamma}(s,t)$ was uniquely determined by $\tilde{\Gamma}(0,0)$, and that the choice of L_m is stable under small perturbations of (s,t), so $\tilde{\Gamma}$ is continuous in s and t. □

Using Lemma 23.13, we can prove that the map $\rho_*\colon \pi_1(X,x_0)\to\pi_1(C,\rho(x_0))$ is injective. Indeed, if $\rho\circ\gamma$ is a contractible loop in C, then Lemma 23.13 allows us to lift the homotopy between $\rho\circ\gamma$ and $e_{\rho(x_0)}$ to a homotopy between γ and e_{x_0}; hence $[\gamma] = [e_{x_0}]$ whenever $\rho_*[\gamma] = [e_{f(x_0)}]$, so ρ_* is an injective homomorphism. □

The key consequence of Proposition 23.12 is that if C is a covering space for X, then $\pi_1(C)$ is isomorphic to a subgroup of $\pi_1(X)$. If C is simply connected, then this subgroup is trivial and we refer to C as the *universal covering space.*

Remark. Definite article here need justification; indeed, universal covering space is unique up to a homeomorphism that commutes with the covering maps. The proof of this is not too difficult and provides a useful exercise. However we will not use uniqueness in our considerations.

There are many examples, though, for which C is not simply connected and we obtain a non-trivial subgroup. For example, let X be the space shown in Figure 4.4(b); X is a covering space for the figure-eight $B_2(S^1)$, but is not simply connected and the tree Γ_4 is in turn a covering space for X. Notice that X is homotopic to the bouquet if infinitely many circles and hence its fundamental group is F_∞, the free group with infinitely many generators. By Proposition 23.12 it is mapped injectively to F_2 for the homomorphism induced by the covering map and hence is isomorphic to a subgroup of it. Indeed the generators of this subgroup are a^nba^{-n}, $n = 1, 2, \dots$. Since F_∞ contains F_n for any natural number n as a subgroup and F_2 is isomorphic to to a subgroup of $_n$ for any $n > 2$ we obtain the following interesting property of free groups that superficially looks somewhat paradoxical:

Proposition 23.14. *Let $m, n \in \{2, 3, \dots ; \infty\}$. Then F_m contains a subgroup isomorphic to F_n.*

d. Deck transformations and group actions. Returning to our original example of the torus $\mathbb{T}^2$, we recall that $\pi_1(\mathbb{T}^2) = \mathbb{Z}^2$, and that $\mathbb{R}^2$ is a covering space for the torus via the canonical projection. Indeed, $\mathbb{R}^2$ is simply connected (even contractible), so it is the universal covering space. We know that in this example the fundamental group $\mathbb{Z}^2$ does not merely sit passively by, but acts on the universal covering space $\mathbb{R}^2$ by translations. In fact, this is once again a specific manifestation of a general phenomenon.

Let $\rho \colon C \to X$ be a covering map. Then given a loop γ in X based at x_0, we can define a map $f_\gamma \colon C \to C$ as follows.

(1) Given $x \in C$, let $\eta \colon [0, 1] \to X$ be a path in X from x_0 to $\rho(x)$.
(2) Consider the loop $\gamma_x = \eta \star \gamma \star \eta^{-1}$, which is a loop in X based at $\rho(x)$.
(3) Using Lemma 23.13, lift γ_x to a path $\tilde{\gamma}_x \colon [0, 1] \to C$ with $\tilde{\gamma}_x(0) = x$.
(4) Define $f_\gamma(x)$ as the other endpoint of $\tilde{\gamma}_x$: $f_\gamma(x) = \tilde{\gamma}_x(1)$.

Once again using Lemma 23.13, one sees that $f_{\gamma_1} = f_{\gamma_2}$ whenever $\gamma_1 \sim \gamma_2$, and so we may write $f_{[\gamma]}$ for any element $[\gamma] \in \pi_1(X)$. To summarise, each element of the fundamental group of X induces a map on the covering space C; this map is called a *deck transformation*.

Example 23.15. If $C = \mathbb{R}^2$ and $X = \mathbb{T}^2$, then $(a, b) \in \mathbb{Z}^2 = \pi_1(X)$ acts on $C = \mathbb{R}^2$ as translation by the vector (a, b).

This action allows us to obtain X as a factor space: $X = C/\pi_1(X)$.

Observe that if C is the universal covering space, then this action is *free*: $f_{[\gamma]}(x) \neq x$ for every $x \in C$ and non-trivial $[\gamma] \in \pi_1(X)$.

Conversely, if G is a group acting freely and discretely on a space X, then the natural projection $X \to X/G$ is a covering map (action by isometries is the simplest case). The case of a simply connected space X will be central for our purposes so let us summarize it:

Theorem 23.16. *Let X be a complete path-connected simply connected metric space and G be a finite of countable group that acts on X by isometries in such a way that there exists $r > 0$ such that for any $g \in G$, $g \neq \mathrm{Id}$ and any $x \in X$ the distance between x and its image $g(x)$ is greater than r. Then $\pi_1(X/G) = G$.*

The condition of freeness is important: for example, the plane $\mathbb{R}^2$ modulo rotation by $2\pi/3$ does not give a covering map, since the origin behaves badly (we get a cone). Thus it is not so easy to get a space with fundamental group $\mathbb{Z}/3\mathbb{Z}$.

Finally, if G is a topological group and Γ is a discrete subgroup of G, then Γ acts freely and discretely on G by left translations. Thus $G \mapsto G/\Gamma$ is a covering map, and $\pi_1(G/\Gamma) = \Gamma$ (this is another interpretation of the torus). The really interesting examples start to turn up when we let G be the group of isometries of $\mathbb{H}^2$ (fractional linear transformations). We will consider such examples in a little while.

Lecture 24. Friday, October 30

a. Subgroups of free groups are free. The theory of covering spaces introduced in the previous lecture can be used to prove a remarkable purely algebraic result, which we will state in a moment. First consider the free group F_n: what are its subgroups?

As we have seen in the previous lecture (Proposition 23.14) among subgroups of $F - N$ are free groups with any number of generators, finite or countable. It is quite remarkable that there is nothing else. Recall that an arbitrary group G is *free* if there exist generators $a_1, \dots, a_n$ for G such that no non-trivial reduced word in the elements $a_i^{\pm 1}$ is equal to the identity element.

Theorem 24.1. *Every subgroup of a free group is free.*

Proof. We use two topological facts proved in the previous lecture: first, that any group acting freely and discretely on a simply connected space then appears as the fundamental group of the factor space (Theorem 23.16), and second, that the fundamental group of any graph is free (Corollary 23.11).

Let F_n be a free group on n generators, and recall that $F_n = \pi_1(B_n(S^1))$. The universal cover of $B_n(S^1)$ is the infinite tree $\mathcal{T}_n$ whose vertices all have degree $2n$. As described in the previous lecture, F_n acts on $\mathcal{T}_n$ by deck transformations; thus any subgroup $G \subset F_n$ also acts on $\mathcal{T}_n$ by deck transformations. This action is free and discrete, so we get

$$G = \pi_1(\mathcal{T}_n/G).$$

The result follows upon observing that $\mathcal{T}_n/G$ is a graph, and hence is homotopic to a bouquet of circles. □

Theorem 24.1 is a purely algebraic result; however, direct algebraic proofs of it are considerably more involved that the elegant geometric argument we presented. Thus the use of topological methods provides a surprisingly powerful tool to address an ostensibly purely algebraic matter.[1]

b. Abelian fundamental groups. As we have seen, fundamental groups can have a very complicated algebraic structure. However, there is one instance worth noting in which this structure simplifies significantly, and the fundamental group $\pi_1(X)$ turns out to be abelian. This occurs then the space X is not just a topological space, but carries a group structure as well.

Theorem 24.2. *Let G be a metrisable path-connected topological group. Then $\pi_1(G)$ is abelian.*

[1] An even more dramatic phenomenon occurs regarding the so-called Fundamental Theorem of Algebra, which has no known purely algebraic proof.

Proof. We take the identity element e as our base point, and consider two loops $\alpha, \beta\colon [0,1] \to G$ with $\alpha(0) = \beta(0) = \alpha(1) = \beta(1) = e$. We must show that $\alpha \star \beta \sim \beta \star \alpha$ by using group multiplication in G to produce a homotopy.

Using the fact that α and β take the value e at the endpoints of the interval, we observe that

$$\alpha \star \beta(t) = \begin{cases} \alpha(2t)\beta(0) & 0 \leq t \leq 1/2, \\ \alpha(0)\beta(2t-1) & 1/2 \leq t \leq 1, \end{cases}$$

and a similar formula holds for $\beta \star \alpha$. Observe that if $\Gamma(s,t)$ is the desired homotopy—that is, $\Gamma(0,t) = \alpha \star \beta(t)$ and $\Gamma(1,t) = \beta \star \alpha(t)$—then for $0 \leq t \leq 1/2$, we must have

$$\Gamma(s,t) = \begin{cases} \alpha(2t)\beta(0) & s = 0, \\ \alpha(0)\beta(2t) & s = 1, \end{cases} \tag{24.1}$$

and for $1/2 \leq t \leq 1$,

$$\Gamma(s,t) = \begin{cases} \alpha(1)\beta(2t-1) & s = 0, \\ \alpha(2t-1)\beta(1) & s = 1. \end{cases} \tag{24.2}$$

It is now easy to see that the following homotopy works:

$$\Gamma(s,t) = \begin{cases} \alpha((1-s)(2t))\beta(s(2t)) & 0 \leq t \leq 1/2, \\ \alpha(s(2t-1)+1-s)\beta((1-s)(2t-1)+s) & 1/2 \leq t \leq 1. \end{cases}$$

One needs only observe that Γ satisfies (24.1) and (24.2), is continuous, and has $\Gamma(s,0) = \Gamma(s,1) = e$ for all $0 \leq s \leq 1$. □

c. Finitely presented groups. Recall that a group G is *finitely generated* if there exists a finite set $\{a_1, \dots, a_n\}$ such that every element $g \in G$ can be written in the form $g = a_{i_1}^{k_1} \cdots a_{i_m}^{k_m}$, where $i_j \in \{1, \dots, n\}$ and $k_j \in \mathbb{Z}$.

If G is the free group on the n generators $a_1, \dots, a_n$, then this representation is unique provided we choose a reduced word—that is, one in which $k_j \neq 0$ and $i_j \neq i_{j+1}$ for all j. For other groups G, however, such representations are not unique. Two prominent examples of this are as follows:

(1) If a generator a_i has order p, then we can always replace a_i^k with $a_i^{k \pm p}$.
(2) If two generators a_i and a_j commute, then we can always reverse their order whenever they occur in a representation of $g \in G$.

The situation, then, is this. There is a one-to-one correspondence between reduced words in the generators and their inverses on the one hand, and elements of the free group on the other; a similar correspondence exists for a group that is not free, but it is no longer one-to-one, and elements of the group correspond not to single reduced words, but to *equivalence classes* of reduced words. For example, in the cyclic group $\langle g \mid g^n = e \rangle$, all the words in the set $\{\dots, g^{1-2n}, g^{1-n}, g, g^1+n, g^1+2n, \dots\}$ are equivalent, and each of them corresponds to the generator of the group.

To get a feel for this process, let us see what happens as we go from F_2 to $\mathbb{Z}^2$ to $(\mathbb{Z}/2\mathbb{Z}) \times (\mathbb{Z}/4\mathbb{Z})$. Each of these groups is generated by two generators, which we write a and b, but the equivalence classes of words corresponding to individual group elements will grow as we go from the first group to the second, and from the second to the third.

To begin with, every element of F_2 corresponds to a unique reduced word; the product of the two generators is ab, and no other reduced word in a and b corresponds to this element. Once we pass to $\mathbb{Z}^2$, we are dealing with an abelian group, and so as elements of $\mathbb{Z}^2$, ab and ba are equal, despite being two non-equivalent reduced words in F_2. Indeed, there are many more words corresponding to this element: for example, $ab = ba = a^{-1}ba^2 = ba^{-1}ba^2b^{-1}$, and so on. Every reduced word w in a and b can be written in the form $w = a^{j_1}b^{k_1}a^{j_2}b^{k_2}\cdots a^{j_m}b^{k_m}$, and one may show without much difficulty that two words correspond to the same element of $\mathbb{Z}^2$ if and only if $\sum_i j_i = \sum_i j_i'$ and $\sum_i k_i = \sum_i k_i'$. Thus the equivalence classes of words are infinite.

Every element of $\mathbb{Z}^2$ corresponds to a word of the form $a^m b^n$, where $m, n \in \mathbb{Z}$. As elements of $\mathbb{Z}^2$, these words are all distinct, but this is not true when we pass to $(\mathbb{Z}/2\mathbb{Z}) \times (\mathbb{Z}/4\mathbb{Z})$. In this (finite) group, a and b have orders 2 and 4, respectively, so now two words $a^m b^n$ and $a^{m'} b^{n'}$ are equivalent if $m - m'$ is a multiple of 2 and $n - n'$ is a multiple of 4. We see that with the addition of more relations and restrictions on the generators, the equivalence classes have grown once again…

Surely there must be some nice way to formalise all this. The second transition above can be given a relatively nice form: the group $(\mathbb{Z}/2\mathbb{Z}) \times (\mathbb{Z}/4\mathbb{Z})$ can be written as the factor group $\mathbb{Z}^2/\langle a^2, b^4\rangle$, and we see that the two restrictions we added are precisely the relationships $a^2 = b^4 = e$. Can we do a similar thing with the transition from F_2 to $\mathbb{Z}^2$?

First we need to observe that every relationship between the generators can be restated by giving a word that is to be considered trivial: for example, the relationship $ab = ba$ is equivalent to the relationship $aba^{-}b^{-1} = e$. Thus a naïve generalisation of the technique just given would say that we ought to produce $\mathbb{Z}^2$ as $F_2/\langle aba^{-1}b^{-1}\rangle$. The first signs are promising: we see that $ab = aba^{-1}b^{-1}ba$, and so ab and ba lie in the same right coset of $\langle aba^{-1}b^{-1}\rangle$. However, things sort of fizzle out at this point, since $\langle aba^{-1}b^{-1}\rangle$ is not a normal subgroup of F_2, and so we do not actually obtain a factor group.[2]

This suggests an obvious remedy: let H be the smallest normal subgroup of F_2 that contains $aba^{-1}b^{-1}$. Then F_2/H is indeed a factor group; of course, it may be trivial, if F_2 does not contain any normal subgroups between $\langle aba^{-1}b^{-1}\rangle$ and the whole group.

Exercise 24.1. Show that $[F_2, F_2]$ is the smallest normal subgroup of F_2 that contains $aba^{-1}b^{-1}$. Furthermore, show that $\mathbb{Z}^2 = F_2/[F_2, F_2]$.

[2] Instead, we get something called a *homogeneous space*.

It is now relatively clear what the general procedure ought to be. Beginning with the free group F_n on n generators $\{a_1, \dots, a_n\}$, we consider a finite list of words $W = \{w_1, \dots, w_m\}$ in the generators a_i. These words are the combinations of the generators that are to be considered trivial, and encode all the relationships between the generators. Let S_W be the smallest normal subgroup of F_n that contains W; we say that the factor group F_n/S_W is the group with generators $a_1, \dots a_n$ defined by the relations $w_1, \dots, w_m$, and write this group as

$$\langle a_1, \dots a_n \mid w_1, \dots w_m \rangle \tag{24.3}$$

The expression (24.3) is a *presentation* of the group. If an arbitrary group G is isomorphic to a group of this form, we say that G is *finitely presented.* One could also consider infinitely presented groups namely, groups that are presented by a finite number of generators but an infinite number of relations. We will see later that interesting properties can be observed in finitely generated infinitely presented groups. Finitely presented groups are the most natural class of countable groups from many viewpoints including geometric applications. It is already a broad enough class that a complete classification is out of the question. However many natural questions about possible properties of finitely generated of groups do not have known answers in the class of finitely presented groups, while infinitely presented groups with such properties may be constructed. We will encounter such examples later in this course.

Example 24.3. The discrete groups we have played with so far are all finitely presented. For example:

(1) $F_2 = \langle a, b \rangle$.
(2) $\mathbb{Z}^2 = \langle a, b \mid [a, b] \rangle$.
(3) $\mathbb{Z}/n\mathbb{Z} = \langle a \mid a^n \rangle$.
(4) $D_n = \langle a, b \mid a^n, b^2, abab \rangle$.

Remark. In everyday usage, it is common to write relationships in the form of equalities: for example, the presentation of the dihedral group is often given as $\langle a, b \mid a^n = b^2 = e, ab = ba^{-1} \rangle$.

Remark. If we are given two finitely presented groups G and H, it is in general a highly non-trivial problem to determine whether or not they are isomorphic. Thus it is not always easy (or indeed possible) to tell whether or not two finite presentations are actually talking about the same group.

One fully tractable class of finitely presented groups is the class of finitely generated abelian groups. One can show (though we do not do so here), that every finitely generated abelian group is isomorphic to a group of the form

$$\mathbb{Z}^n \oplus \left(\bigoplus_{p \text{ prime}} \bigoplus_{m \geq 1} (\mathbb{Z}/p^m\mathbb{Z})^{k(p,m)} \right). \tag{24.4}$$

Furthermore, the integers n and $k(p,m)$ form a complete system of invariants; that is, two finitely generated abelian groups are isomorphic if and only if $n = n'$ and $k(p,m) = k'(p,m)$ for every p, m.

Remark. Arbitrary countable abelian groups have considerably more complicated albeit still tractable structure. The group of rational numbers by addition provides an instructive example.

Remark. The use of the direct sum $\oplus$ rather than the direct product $\times$ in (24.4) is important. For collections of finitely many abelian groups, these are the same: $G \oplus H = G \times H$. However, the story is different for infinite collections. Each element of an infinite direct *product* $G_1 \times G_2 \times \cdots$ corresponds to a sequence $(g_1, g_2, \dots)$, where $g_i \in G_i$ may be arbitrary.This is an uncountable abelian group. In a direct *sum*, however, there is a cofiniteness requirement: the sequences that occur in an infinite direct sum are precisely those which have only finitely many non-identity elements. This group is countable although not finitely generated.

For example, if we take the direct product of countably many copies of $\mathbb{Z}/2\mathbb{Z} = \{0,1\}$, we are dealing with infinite sequences of 0s and 1s, added coordinate-wise (modulo 2). If, however, we take the direct *sum*, then we are only dealing with *finite* sequences (albeit of arbitrarily length), since every permissible sequence has a "tail" which is eventually all 0s.

d. Free products. Certain groups, such as $SL(2,\mathbb{Z})$, can be best described not by group presentations, but by a different construction, which we will soon describe. Observe that $SL(2,\mathbb{Z})$ is not a free group, as it contains elements of finite order, such as $\left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$, which has order 4, and $\left(\begin{smallmatrix} 0 & 1 \\ -1 & 1 \end{smallmatrix}\right)$, which has order 6. However, it *does* have a finite index free subgroup.

The construction of the free group can be generalised to the *free product* of two groups. Given groups G and H, their free product $G * H$ is the set of all words $g_1 h_1 g_2 h_2 \cdots g_m h_m$, where $g_i \in G$ and $h_i \in H$, with the obvious cancellations, and with binary operation given by concatenation.

Example 24.4. $F_2 = \mathbb{Z} * \mathbb{Z}$, and more generally, F_n is the free product of n copies of $\mathbb{Z}$.

Exercise 24.2. What is $(\mathbb{Z}/n\mathbb{Z}) * (\mathbb{Z}/m\mathbb{Z})$? *Hint: Construct its Cayley graph.*

We will show later that $PSL(2,\mathbb{Z})$, the factor of $SL(2,\mathbb{Z})$ by its center $\pm \mathrm{Id}$ is isomorphic to $(\mathbb{Z}/2\mathbb{Z}) * (\mathbb{Z}/3\mathbb{Z})$.

Lecture 25. Monday, November 2

a. Planar models. The torus can be thought of in many ways. It can be embedded into $\mathbb{R}^3$; it can be a factor space $\mathbb{R}^2/\mathbb{Z}^2$; and it can also be thought of as the unit square with opposite edges identified. If we label the vertical edges a and the horizontal edges b, then upon "rolling" the square into the embedded torus, a and b correspond to loops on $\mathbb{T}^2$ that generate the fundamental group $\mathbb{Z}^2$. The union of the loops a and b gives an embedding of the figure-eight $B_2(S^1)$ into $\mathbb{R}^3$, which has fundamental group F_2. To obtain the torus from this skeleton, we may think of stretching a "film" of some sort so that it is anchored to the loops a and b, and composes the rest of the surface of the torus. This film corresponds to the interior of the unit square, and we see that the boundary of the square is $aba^{-1}b^{-1}$, which is contractible through that interior.

Thus adding the film corresponds to imposing the relationship $aba^{-1}b^{-1} = e$ on F_2. As we saw earlier, this leads to $F_2/[F_2, F_2] = \mathbb{Z}^2$, which explains the relationships between the fundamental groups.

Here is another example of a similar construction. Consider the hexagon with opposite edges identified, as shown in Figure 5.2. Label the three pairs of opposite edges a, b, c: let $\mathbf{v}_1$ be the vector from x_1 to x_3, so that translation by $\mathbf{v}_1$ maps one edge labeled a onto the other, and let $\mathbf{v}_2$ and $\mathbf{v}_3$ be similarly defined, as shown in the picture.

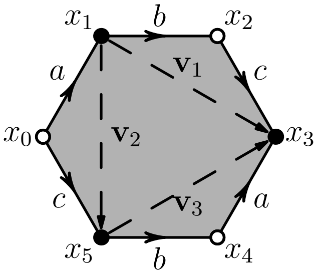


FIGURE 5.2. A planar model on a hexagon.

This is an example of a general sort of construction: we begin with a polygon P (in this case, a hexagon), and then introduce an equivalence relation $\sim$ on the boundary of P, under which certain pairs of edges are identified. The pair $(P, \sim)$ is called a *planar model* of the topological space obtained as the quotient $X = P/\sim$.

As we have seen, the square with opposite edges identified is a planar model of the torus: Figure 5.3 shows how horizontal and vertical lines in the planar model correspond to parallels and meridians on the torus. The fact that all the horizontal lines in the planar model are the same length, while the parallels on the torus (circles around the z-axis) are of varying lengths, illustrates the fact that we are capturing the *topology* of the situation, and not its geometry.

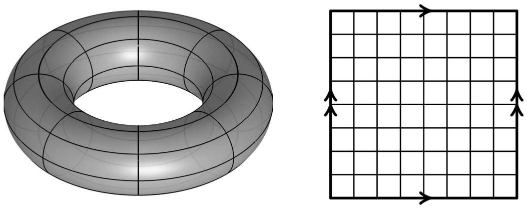


FIGURE 5.3. A planar model of the torus.

What is the hexagon with opposite edges identified a planar model of? Observe that the images of the hexagon under all translations generated by $T_{\mathbf{v}_1}$, $T_{\mathbf{v}_2}$, and $T_{\mathbf{v}_3}$ tiles the plane, and thus the subgroup $H \subset \mathrm{Isom}(\mathbb{R}^2)$ generated by these three translations acts freely and discretely on $\mathbb{R}^2$. The quotient $P/\sim$ is just the factor space $\mathbb{R}^2/H$; furthermore, because $\mathbf{v}_1 = \mathbf{v}_2 + \mathbf{v}_3$, we see that in fact $H = \langle \mathbf{v}_1, \mathbf{v}_2 \rangle$ is a lattice, and this factor space is a torus.

Observe that when we identify opposite edges of the square, all vertices are identified, and hence each edge becomes a loop. This does not happen on the hexagon; instead we get two different equivalence classes of vertices ($\{x_0, x_2, x_4\}$ and $\{x_1, x_3, x_5\}$), and so a, b, c are not themselves loops. Writing ab for $a \star b$, and so on, we see that ab, ca^{-1}, and cb *are* loops. Each of these loops is non-contractible; however, the product $abca^{-1}b^{-1}c^{-1}$ corresponds to the perimeter of the hexagon traversed clockwise beginning at x_0, and *is* contractible by the obvious homotopy through the interior of the hexagon.

So we got a torus from both the square and the hexagon. What if we move up to the octagon, and again consider the planar model with opposite edges identified? In this case we immediately see that the tiling procedure from before fails: $\mathbb{R}^2$ can be tiled with squares and with hexagons, but not with octagons. Indeed, all eight vertices of the octagon are to be identified, but the sum of the internal angles is 6π, not 2π.

It follows that the translations matching opposite edges of the regular octagon do not generate a discrete subgroup of $\mathbb{R}^2$, which means that this planar model does not admit a nice geometric interpretation as $\mathbb{R}^2$ modulo some discrete group of isometries.

We have two options, then: we can forget about the geometry of the situation and adopt a combinatorial and topological approach, or we can use a different geometry. We will follow the first of these for now, and return to the second later on.

EXERCISE 25.1. Using cutting and pasting, show that the planar model of the octagon with opposite sides identified is equivalent to the planar model shown in Figure 5.4. (Note that rather than placing arrows along the edges to indicate the direction of identification, we write a for the edge a directed counterclockwise, and a^{-1} to indicate a clockwise direction.)

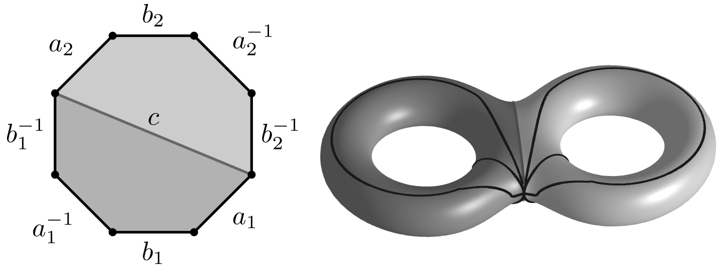


Figure 5.4. From an octagon to a surface of genus two.

The planar model in Figure 5.4 has two halves, each of which is a pentagon with two pairs of sides identified and the fifth side left free. Figure 5.5 shows that this pentagon may be obtained by cutting a hole in the usual planar model of a torus (to obtain a so-called "handle"), and so the octagon with opposite edges identified is equivalent to the surface obtained by cutting holes in two tori and gluing them together along those holes. This gives the "pretzel" surface shown in Figure 5.4.

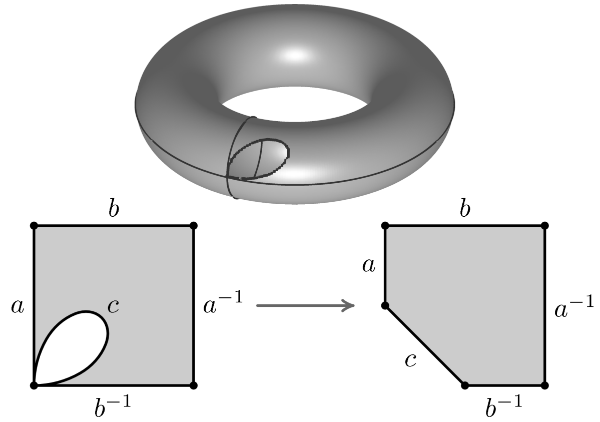


Figure 5.5. Cutting a hole in a torus.

All eight vertices of the octagon are identified, and so each of the curves a_1, a_2, b_1, b_2 becomes a loop in the quotient space, as shown in Figure 5.4. Intuitively, we expect that these loops generate the fundamental group of this surface; furthermore, since the curve $a_1 b_1 a_1^{-1} b_1^{-1} a_2 b_2 a_2^{-1} b_2^{-1}$ is the perimeter of the octagon, it is contractible (through the octagon), and we expect that this is the only relationship between the generators.

This construction can be generalised to any $4n$-gon to produce a surface with more "holes"—or, to use the technical term, a surface of higher genus. (The genus is the number of holes.) Based on the above considerations, we expect that the fundamental group of the surface of genus n is the *surface* *group* with presentation

$$\langle a_1, b_1, \dots, a_n, b_n \mid [a_1, b_1] \cdots [a_n, b_n] \rangle.$$

This will appear as a particular case of a more general construction.

b. The fundamental group of a polygonal complex. One of our key results so far has been to exhibit every free group on finitely or countably many generators as the fundamental group of a graph, by associating the generators of the group to loops in the graph. What we have *not* been able to do so far is to take a group G in which those generators satisfy certain relationships and exhibit a topological space that has G as its fundamental group.

In the previous section, we saw that planar models place a relationship on a collection of loops (the edges of the polygon) by allowing homotopy through the interior of the polygon. Using the suitable generalisation of this idea, we will be able to produce a topological space with arbitrary finitely presented fundamental group.

The idea, then, is to generalise the notion of a graph (which is a collection of zero-dimensional vertices and one-dimensional edges with some combinatorial relationships) to a two-dimensional object by adding some faces.

For our purposes, a graph can be defined as a metric space that is the union of a collection V of vertices (points) together with a collection E of edges (homeomorphic images of the open interval $(0, 1)$) such that

(1) no vertex in V lies on an edge $e \in E$;
(2) the endpoints of every edge $e \in E$ are vertices in V; and
(3) every pair of distinct edges $e \neq e' \in E$ is in fact disjoint.

The generalisation of this definition to two-dimensional objects is straightforward. A *polygonal complex* is a metric space obtained as the union of a collection V of vertices, a collection E of edges, and a collection F of faces (homoeomorphic images of the open unit disc $\{(x, y) \mid x^2 + y^2 < 1\}$, such that

(1) no vertex in V lies on an edge $e \in E$ or a face $f \in F$, and no edge $e \in E$ has non-trivial intersection with a face $f \in F$;
(2) the boundary of every face $f \in F$ is a union of edges and vertices, and the endpoints of every edge are vertices in V;
(3) every pair of distinct edges is disjoint, as is every pair of distinct faces.

Remark. This definition can also be generalized to dimensions higher than two. While the straightforward one is not used very often there are two versions, one more specialized (simplicial complexes) and one more general (CW-complexes), that are central objects of algebraic topology.

Observe that the planar models described in the previous section are all examples of polygonal complexes. For example, the square with opposite sides identified is a polygonal complex with one vertex, two edges, and one face. Similarly, the hexagon with opposite sides identified has two vertices, three edges, and one face.

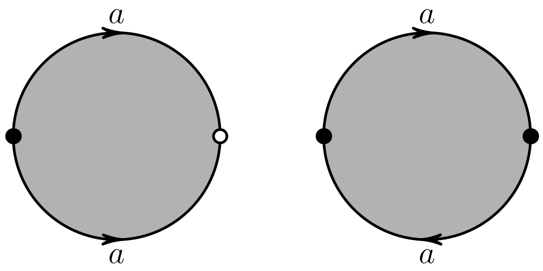


FIGURE 5.6. Polygonal complexes for the sphere and the projective plane.

There is no requirement that a face have enough edges to make it a polygon in the Euclidean sense; indeed, Figure 5.6 shows two polygonal complexes whose only face is a 2-gon. The first of these has one face, one edge, and two vertices; if we "fold" the disc up to identify the two edges labeled a (sort of like folding up a taco), we see that this polygonal complex is nothing but the sphere S^2.

What about the second complex in Figure 5.6, which has exactly one face, one edge, and one vertex? We see that this space is the unit disc with antipodal boundary points identified; this is homeomorphic to the upper hemisphere of S^2 with antipodal points on the equator identified. This is in turn homeomorphic to the entire sphere with *all* pairs of antipodal points identified, which we know to be the projective plane $\mathbb{R}P(2)$.

Recall that since $\mathbb{R}P(2)$ is the factor of S^2 by the free and discrete action of $\mathbb{Z}/2\mathbb{Z}$, and furthermore $\pi_1(S^2)$ is trivial, we have $\pi_1(\mathbb{R}P(2)) = \mathbb{Z}/2\mathbb{Z}$. As with the planar models we saw earlier, we observe that the fundamental group has presentation $\langle a \mid a^2\rangle$, where the generator is the single edge and the relation corresponds to the perimeter of the single face.

In fact, this gives a general procedure for finding the fundamental group of a (connected) polygonal complex.

Let $\mathcal{C} = (V, E, F)$ be a connected polygonal complex, and write $E = \{e_1, \dots, e_n\}$. Fix an orientation on each edge e_i. Every face $f \in F$ determines a word in the symbols $e_1, \dots, e_n$ by recording the symbols corresponding to the edges traversed in one counterclockwise circuit of the perimeter of f. Write $w_1, \dots, w_m$ for the collection of such words.

We now have the tools for a group presentation: to the connected polygonal complex $\mathcal{C}$ we associate the group

$$G(\mathcal{C}) = \langle e_1, \dots, e_n \mid w_1, \dots, w_m\rangle. \tag{25.1}$$

Based on our previous experience, we expect to find that $G(\mathcal{C}) = \pi_1(\mathcal{C})$. Is this true? In the first place, the edges e_i are not necessarily loops if V contains more than a single vertex. Furthermore, we need to show that every loop in $\mathcal{C}$ is in fact homotopic to a loop corresponding to a concatenation of edges.

THEOREM 25.1. *If $\mathcal{C}$ is a connected polygonal complex, then $\pi_1(\mathcal{C}) = G(\mathcal{C})$.*

Right now we will only prove one half of this theorem, namely we will prove that $\pi_1(\mathcal{C})$ is a factor-group of $G(\mathcal{C})$. We define a map $\varphi\colon G(\mathcal{C}) \to \pi_1(\mathcal{C})$ using the idea of a maximal tree. Observe that the collection of all vertices and edges in $\mathcal{C}$ is a graph $\mathcal{G}$, which has a maximal tree $\mathcal{T}$. Fix a vertex $v_0 \in \mathcal{T}$, and for every vertex $v \in V$, let γ_v be the unique path in $\mathcal{T}$ that moves from v_0 to v with unit speed.

Now given an edge $e \in E$ that runs from v to w, define a loop in $\mathcal{C}$ by $\gamma_v \star e \star \gamma_w^{-1}$. Thus we may define φ on each generator $e = (v, w)$ by

$$\varphi((v, w)) = [\gamma_v \star e \star \gamma_w^{-1}] \in \pi_1(\mathcal{C}, v_0).$$

To obtain a homomorphism from $G(\mathcal{C})$ to $\pi_1(\mathcal{C})$, we extend φ in the natural way. One needs to check that φ is well-defined, and indeed, if w is a word in the generators e_i that lies in the normal subgroup of F_n generated by $w_1, \dots, w_m$, then the corresponding loop in $\mathcal{C}$ can be contracted through the faces of $\mathcal{C}$, and so $\varphi(w)$ is the identity element.

It remains to show that φ is a bijection. To show that φ is onto, it suffices to observe that if γ is any loop in $\mathcal{C}$, then the section of γ lying inside any given face can be homotoped to lie on the edges adjacent to that face, and once a loop γ lies on the graph $\mathcal{G}$, it is homotopic to a loop generated by the loops $\gamma_v \star (v, w) \star \gamma_w^{-1}$. Thus we proved that $\pi_1(\mathcal{C})$ is a factor-group of $G(\mathcal{C})$.

As a consequence of Theorem 25.1, we can obtain any finitely presented group as the fundamental group of a compact metric space.

c. Factor spaces of the hyperbolic plane. We return now to the surface of genus two, the "pretzel" surface obtained by identifying opposite sides of the octagon. As we saw earlier, this surface cannot be obtained as $\mathbb{R}^2/G$ for any subgroup $G \subset \mathrm{Isom}(\mathbb{R}^2)$; we will see now that if we replace the Euclidean plane $\mathbb{R}^2$ with the hyperbolic plane $\mathbb{H}^2$, things are quite different.

First observe that if T_1 and T_2 are *any* Euclidean translations, then they obey the relationship $T_1 \circ T_2 = T_2 \circ T_1$. Things are quite different in the hyperbolic plane, where the analogues of translations are hyperbolic transformations, which have two fixed points on the ideal boundary, and move points in $\mathbb{H}^2$ along circles connected these two ideal fixed points. Such a transformation in the upper half-plane was illustrated in Figure 3.6; we will work now in the unit disc model, where the geometry of the transformation is as shown in Figure 5.7(a).

A hyperbolic transformation with fixed points w_1 and w_2 on the ideal boundary moves points along the geodesic connecting w_1 and w_2 (the horizontal line). The transversal curves such as γ_1, which are circles intersection

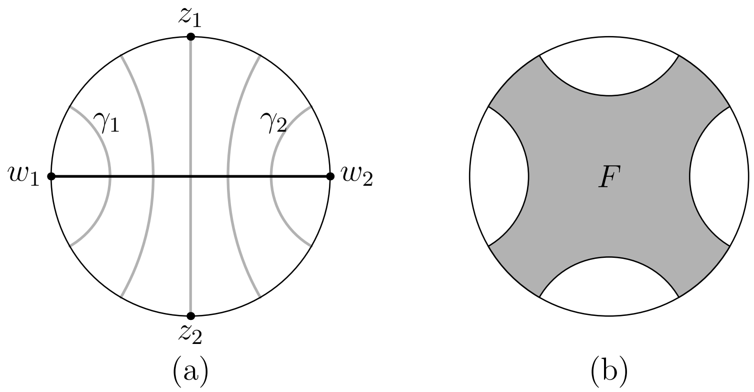


FIGURE 5.7. A hyperbolic transformation.

this geodesic and the ideal boundary orthogonally, are moved into other such curves.

Let f be a hyperbolic transformation that fixes w_1 and w_2 and takes γ_1 to γ_2, and let g be a hyperbolic transformation that acts similarly on the geodesic from z_1 to z_2. Let F be the subset of $\mathbb{H}^2$ shown in Figure 5.7(b), and let $G = \mathbb{H}^2 \setminus F$. Then $f^k(F) \subset G$ and $g^k(F) \subset G$ for any integer $k \neq 0$.

EXERCISE 25.2. Using the action of f and g on the region F, show that f and g generate a free subgroup of $\mathrm{Isom}(\mathbb{H}^2)$.

Exercise 25.2 is a particular case of the "ping-pong lemma", which gives a general method of showing that a subgroup of a group acting on a set is actually free.

The existence of a free subgroup of $\mathrm{Isom}(\mathbb{H}^2)$ stands in stark contrast to the situation in $\mathrm{Isom}(\mathbb{R}^2)$. We will show in the next lecture that the latter does not contain a free subgroup with two generators, discrete or not. To accentuate complexity of the situation let us add that $SO(3)$ and hence $\mathrm{Isom}(\mathbb{R}^3)$ contains a free subgroup but not a discrete free subgroup.

To return to the octagon, we observe that we may replace F in Figure 5.7(b) with a figure bounded by symmetrically located eight geodesics, rather than four. If we increase the Euclidean length of these geodesics by moving them inwards toward the centre of the disc, they will eventually intersect, so that they bound an octagon. At the first moment of intersection, they will be tangent, and so the sum of the angles of the octagon is 0. As they move closer and closer to the centre, the angles of intersection grow, approaching the Euclidean limit of $3\pi/4$, and so the sum of the angles approaches 6π. By continuity, we can find geodesics that bound an octagon whose angles sum to exactly 2π. It turns out that the hyperbolic translations matching opposite sides of this octagon generate a discrete subgroup of $\mathrm{Isom}(\mathbb{H}^2)$, and that the images of the octagon under this subgroup tile the hyperbolic plane.

Lecture 26. Wednesday, November 4

a. Hyperbolic translations and fundamental domains. Let us now spend a little more time with the embedding of the free group F_2 into $PSL(2,\mathbb{R})$, the group of isometries of the hyperbolic plane $\mathbb{H}^2$. First recall that $GL(2,\mathbb{C})$ acts on the Riemann sphere $\mathbb{C}\cup\infty$ by fractional linear transformations: to a matrix $A = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in GL(2,\mathbb{C})$ we associate the transformation

$$\varphi_A(z) = \frac{az+b}{cz+d}.$$

The kernel of this action is the set of scalar multiples of the identity matrix: that is, $\varphi_A = \varphi_{A'}$ if and only if $A = \lambda A'$ for some $\lambda \in \mathbb{C}$.

Exercise 26.1. Show that $\varphi_A \circ \varphi_B = \varphi_{AB}$.

If A has real entries, then φ_A preserves the extended real line and maps the upper half-plane $H = \{z \in \mathbb{C} \mid \operatorname{Im} z > 0\}$ to itself, which lets us consider it as a transformation of the hyperbolic plane $\mathbb{H}^2$.

We can also consider the unit disc model for $\mathbb{H}^2$, given by $D = \{z \in \mathbb{C} \mid |z| < 1\}$. The Möbius transformation $\varphi_B(z) = \frac{-z+i}{z+i}$ associated to the matrix $B = \left(\begin{smallmatrix} -1 & i \\ 1 & i \end{smallmatrix}\right)$ maps H bijectively onto D, taking the ideal boundary $\mathbb{R}\cup\infty$ to the unit circle S^1 (in particular, 0 to 1 and ∞ to -1), and taking i to 0, the centre of the disc. Taking the inverse of B and rescaling the matrix, we see that $\varphi_B^{-1}(z) = \frac{-z+1}{-iz-i}$ maps D back onto H.

Treating ϕ_B as a change of coordinates, we define the map $\psi_A \colon D \to D$ such that the following diagram commutes:

$$\begin{tikzcd}
H \arrow[r, "\varphi_A"] \arrow[d, "\varphi_B"] & H \arrow[d, "\varphi_B"] \\
C \arrow[r, "\psi_A"] & C
\end{tikzcd}$$

That is, $\psi_A = \varphi_B \circ \varphi_A \circ \varphi_B^{-1}$, and so A acts on D as $\varphi_{A'}$, where

$$\begin{aligned} A' &= \begin{pmatrix} -1 & i \\ 1 & i \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} -1 & 1 \\ -i & -i \end{pmatrix} \\ &= \begin{pmatrix} (a+d)+(b-c)i & (a-d)+(b+c)i \\ (a-d)-(b+c)i & (a+d)-(b-c)i \end{pmatrix} = \begin{pmatrix} u & w \\ \overline{w} & \overline{u} \end{pmatrix}; \end{aligned}$$

here $u = (a+d)+(b-c)i$ and $w = (a-d)+(b+c)i$, and we may take $\det A' = |u|^2 - |w|^2 = 1$.

Remark. Matrices of the form given above compose a group known as $SU(1,1)$; it is reminiscent of the special unitary group $SU(2)$, except the Hermitian product (12.1) is replaced by $z_1\overline{w_1} - z_2\overline{w_2}$. This group acts on the unit disc in the same way $SL(2,\mathbb{R})$ acts on the upper half-plane, preserving the ideal boundary (in this case, the unit circle), and mapping the disc to itself.

Recall that a fractional linear transformation $\varphi_A\colon H \to H$ is *hyperbolic* if $A \in SL(2,\mathbb{R})$ has trace $\operatorname{Tr} A > 4$, in which case φ_A fixes two points on the ideal boundary $\mathbb{R} \cup \infty$. In this case, the images of these points under the change of coordinates φ_B are fixed points of ψ_A on the ideal boundary S^1 of the unit disc D. The situation when these two points are antipodal was illustrated in Figure 5.7; of course, they may also lie in some other configuration, and Figure 5.8 shows the general setup.

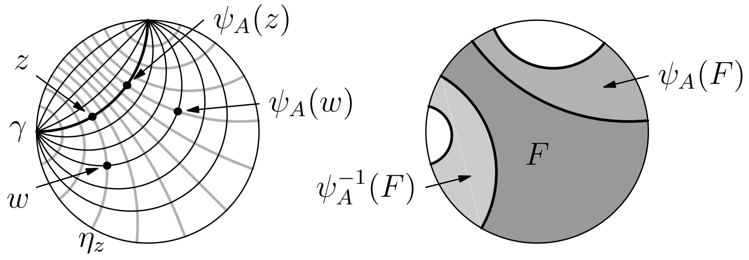


FIGURE 5.8. A hyperbolic transformation.

Geodesics in $\mathbb{H}^2$ are lines and circles that are orthogonal to the ideal boundary. There is a unique geodesic γ connecting the ideal fixed points of ψ_A, and ψ_A preserves γ. It also preserves the other circles through the ideal fixed points shown in Figure 5.8, although these are not geodesics. (They are the so-called *equidistant curves*, which each run a fixed (hyperbolic) distance from the geodesic γ.)

Another important family of curves for ψ_A is the set of geodesics orthogonal to γ. These partition $\mathbb{H}^2$, and furthermore, if we write η_z for the geodesic through $z \in \gamma$ orthogonal to γ, then $\psi_A(\eta_z) = \eta_{\psi_A(z)}$. Thus ψ_A acts on this family of curves.

Fix $z \in \gamma$ and let $(z, \psi_A(z))$ denote the set of points on γ that lie between z and its image $\psi_A(z)$. Then consider the region

$$F = \bigcup_{w \in (z,\psi_A(z))} \eta_w \tag{26.1}$$

that comprises all geodesics orthogonal to γ through points in the interval $(z, \psi_A(z))$. F is significant for understanding the action of ψ_A on D; we start by recalling the following definition.

DEFINITION 26.1. Let G be a group acting on a set X. A subset $F \subset X$ is a *fundamental domain* for this group action if

(1) $g_1 F \cap g_2 F = \emptyset$ for all $g_1 \neq g_2 \in G$, and
(2) $\bigcup_{g \in G} gF = X$.

That is, a fundamental domain is a subset whose images under all the elements of G tile the set X.

Remark. In order for a fundamental domain to exist, G must act freely on X, otherwise there exists $g \in G$ and $x \in X$ such that $g(x) = x$, and hence if $x \in hF$, we also have $x \in ghF$.

If X is not just a set, but carries a topology as well, then we usually want to avoid fundamental domains that are topologically unpleasant. For example, the set $[0, 1/2) \cup (3/2, 2)$ is a fundamental domain for the action of $\mathbb{Z}$ on $\mathbb{R}$, but not one we ever really want to use...

Thus in this case we restrict ourselves to *connected* fundamental domains, such as $[0, 1)$. Furthermore, it is somewhat unwieldy to have to include some of the boundary, but not all of it; this motivates a slight modification of the above definition.

Definition 26.2. Let G be a group acting on a topological space X. A subset $F \subset X$ is a *topological fundamental domain* for this group action if

(1) F is open,
(2) $g_1F \cap g_2F = \emptyset$ for all $g_1 \neq g_2 \in G$, and
(3) $\bigcup_{g \in G} g\overline{F} = X$.

Example 26.3. Every topological fundamental domain for the action of $\mathbb{Z}$ on $\mathbb{R}$ by addition ($n\colon x \mapsto x + n$) is an open interval $(a, a + 1)$ for some $a \in \mathbb{R}$.

Example 26.4. Given two linearly independent vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^2$, we have an action of $\mathbb{Z}^2$ on $\mathbb{R}^2$ by $(m, n)\colon \mathbf{x} \mapsto \mathbf{x} + m\mathbf{u} + n\mathbf{v}$. The obvious topological fundamental domain for this action is the parallelogram

$$\{a\mathbf{u} + b\mathbf{v} \mid 0 < a < 1, 0 < b < 1\}, \tag{26.2}$$

or any translation of this parallelogram. However, there are many other fundamental domains as well.

In order for a topological fundamental domain to exist, the group G must act discretely on X. However, the action need not necessarily be free; for example, $\mathbb{Z}/3\mathbb{Z}$ acts on $\mathbb{R}^2$ by rotations by multiples of $2\pi/3$, and the sector

$$\{(r\cos\theta, r\sin\theta) \mid 0 < r < \infty, 0 < \theta < 2\pi/3\}$$

is a topological fundamental domain.

A useful general construction of fundamental domain when G acts by isometries is the *Dirichet domain* D_x: pick a point $x \in X$ and let D_x be the set of points that are closer to x than to any other point on the orbit of x. Conditions (1) and (2) of Definition 26.2 are obviously satisfied. In order for (3) to hold the set of points equidistant from two point must be nowhere dense since the boundary of D_x consists of parts of several such sets. This condition obviously holds in Euclidean, spherical (and hence elliptic) and hyperbolic geometry since the sets in questions are lines (geodesics); this generalizes to higher dimensions.

Exercise 26.2. Prove that the Dirichlet domain for the action by translations as in Example 26.4 is either a centrally symmetric hexagon or a parallelogram.

Returning to the hyperbolic plane, we see that the set F defined in (26.1) is a topological fundamental domain for the action of $\mathbb{Z}$ on D given by $n \to \psi_A^n$. Figure 5.8 shows the images of F corresponding to $n = -1$, $n = 0$, and $n = 1$. If we take the images corresponding to all other values of $n \in \mathbb{Z}$, we obtain a tiling of D.

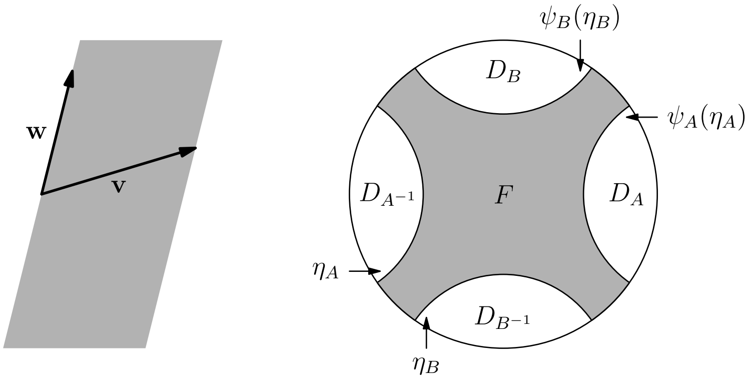


Figure 5.9. Fundamental domains in $\mathbb{R}^2$ and $\mathbb{H}^2$.

b. Existence of free subgroups. Now the fun begins. Given $\mathbf{v} \in \mathbb{R}^2$, the translation $T_{\mathbf{v}}$ induces an action of $\mathbb{Z}$ on $\mathbb{R}^2$; if $\mathbf{w} \in \mathbb{R}^2$ is linearly independent from $\mathbf{v}$, then the strip

$$\{a\mathbf{v} + b\mathbf{w} \mid 0 < a < 1, b \in \mathbb{R}\}$$

is a topological fundamental domain for this action (see Figure 5.9). Furthermore, if $\mathbf{v}, \mathbf{w} \in \mathbb{R}^2$ are linearly independent, then we can obtain a fundamental domain for $\langle T_{\mathbf{v}}, T_{\mathbf{w}} \rangle$ of the form (26.2) by taking the intersection of the fundamental domains for $\langle T_{\mathbf{v}} \rangle$ and $\langle T_{\mathbf{w}} \rangle$.

Let ψ_A and ψ_B be hyperbolic transformations of the unit disc model D of the hyperbolic plane $\mathbb{H}^2$, and let γ_A and γ_B be the corresponding geodesics connecting ideal fixed points. Fix $z_A \in \gamma_A$ and $z_B \in \gamma_B$, and let η_A and η_B be the geodesics through z_A and z_B orthogonal to γ_A and γ_B, respectively. Let F_A and F_B be the corresponding fundamental domains for $\langle \psi_A \rangle$ and $\langle \psi_b \rangle$—that is, F_A is the region between η_A and $\psi_A(\eta_A)$, and similarly for F_B. Finally, let $F = F_A \cap F_B$.

Now we make a crucial assumption: suppose that ψ_A and ψ_B are such that the four geodesics $\eta_A, \eta_B, \psi_A(\eta_A), \psi_B(\eta_B)$ do not intersect each other. This amounts to requiring that ψ_A and ψ_B translate points on γ_A and γ_B by a large enough distance, and guarantees that F has the form shown in Figure 5.9.

Observe that $D \setminus F$ is the union of four disjoint regions, and that each of the four images $\psi_A(F)$, $\psi_A^{-1}(F)$, $\psi_B(F)$, $\psi_B^{-1}(F)$ lies in one of these regions. Denote the region containing $\psi_A(F)$ by D_A, the region containing $\psi_A^{-1}(F)$ by $D_{A^{-1}}$, and similarly for B.

Proposition 26.5. *With A, B as above, the subgroup $\langle A, B\rangle \subset SL(2,\mathbb{R})$ is free.*

Proof. It suffices to show that the natural homomorphism $F_2 \to \langle A, B\rangle \subset SL(2,\mathbb{R})$ has trivial kernel. That is, every reduced word in A and B corresponds to a matrix in $SL(2,\mathbb{R})$, and hence to a Möbius transformation of D. If $w = w_1 \cdots w_n$ is such a word, where each w_i is either A or B, then we must show that $\psi_w = \mathrm{Id}$ if and only if w is the trivial word.

We do this by induction on the length of w, showing that $\psi_w(F) \subset D_{w_1}$ for every reduced word w. The case $n = 1$ is immediate from the observation that $\psi_A(F) \subset D_A$, $\psi_B(F) \subset D_B$, and similarly for the inverses. Now if the result holds for some value of $n \geq 1$, then for any word $w = w_1 \cdots w_{n+1}$, we have

$$\psi_w(F) = \psi_{w_1}(\psi_{w_2 \cdots w_{n+1}}(F)) \subset \psi_{w_1}(D_{w_2}) \subset D_{w_1},$$

using the inductive hypothesis, the assumption that w is reduced, and the fact that $\psi_{w_1}(D \setminus D_{w_1^{-1}}) \subset D_{w_1}$. It follows that $\psi_w \neq \mathrm{Id}$, and hence $\langle A, B\rangle$ is free. □

We will now spend a little time with the question of which groups have free subgroups—although $\mathbb{Z}$ is technically a free group and is isomorphically embedded into any group that contains an element of infinite order, we are concerned with more substantive examples, and so will use "free subgroup" to refer to a free subgroup on at least 2 generators.

Since many matrix groups contain an isomorphic image of $SL(2,\mathbb{R})$, Proposition 26.5 shows that free subgroups occur in many matrix groups—for example, $SL(n,\mathbb{R})$ for any $n \geq 2$. $SO(3)$ is compact, and hence does not contain a copy of $SL(2,\mathbb{R})$. It turns out that $SO(3)$ contains a free subgroup regardless, but this subgroup is not discrete.

Obviously an abelian group does not contain a free subgroup; however, there are other, more interesting examples.

Theorem 26.6. $\mathrm{Isom}(\mathbb{R}^2)$ *does not contain a free subgroup.*

Proof. Recall that $G = \mathrm{Isom}(\mathbb{R}^2)$ is solvable, with $G_1 = [G,G] = \mathrm{Isom}^+(\mathbb{R}^2)$, $G_2 = [G_1, G_1] = \mathcal{T}$ (the subgroup of translations), and $G_3 = [G_2, G_2] = \{\mathrm{Id}\}$. Furthermore, every subgroup of a solvable group is solvable, and so we have reduced the problem to proving the following proposition.

Proposition 26.7. *F_2 is not solvable.*

Proof. How can a group fail to be solvable? One simple way is for G to contain a subgroup H with $[H,H] = H$, and indeed, if G is finite this is the only possibility, since the sequence G_n must eventually stabilise. If

G is infinite, this sequence may not stabilise, and this turns out to be the case for F_2. Indeed, we already saw that $[F_2, F_2]$ is a subgroup of countable index (the factor group being $\mathbb{Z}^2$), and we have no more luck further along.

Thus we use a slightly different approach. Recall that any subgroup of a free group is free; in particular, $[F_2, F_2]$ is a free group. Furthermore, direct inspection shows that $[F_2, F_2]$ is non-abelian, and hence contains an isomorphic image of F_2. It immediately follows that $G = F_2$ cannot be solvable, since every group G_n in the derived sequence contains a free subgroup. □

This completes the proof of Theorem 26.6, since a non-solvable group cannot be contained in a solvable group. □

Theorem 26.6 is really the observation that solvable groups are somehow "too small" to contain a free subgroup. The dichotomy between groups that contain a free subgroup and groups that are too small to do so forms the heart of the *Tits alternative*, which was introduced by Jacques Tits in 1972, and states that within the class of finitely generated matrix groups, the only groups that do not contain a free subgroup are groups with a solvable subgroup of finite index.

Remark. With a little more work, Proposition 26.7 can be strengthened to the statement that $[F_2, F_2] = F_\infty$, the free group on countably many generators. To see this, let $\mathcal{G}$ be the graph in $\mathbb{R}^2$ whose vertices are points on the integer lattice $\mathbb{Z}^2$ and whose edges are all horizontal and vertical unit intervals connecting adjacent vertices. Label each (directed) horizontal edge a and each (directed) vertical edge b, and put the same labels on the two loops in the figure-eight $B_2(S^1)$. Then we have a natural covering map $\rho\colon \mathcal{G} \to B_2(S^1)$, which gives an isomorphism between $\pi_1(\mathcal{G})$ and the subgroup $[F_2, F_2] \subset \pi_1(B_2(S^1))$. Letting $\mathcal{T} \subset \mathcal{G}$ be any maximal tree, we see that $\mathcal{G} \setminus \mathcal{T}$ has infinitely many loops, and so $\pi_1(\mathcal{G}) = F_\infty$.

c. Surfaces as factor spaces. In the next lecture, we will return to the question of obtaining surfaces of higher genus as factor spaces of $\mathbb{H}^2$ by groups of isometries. For now, we recall the situation in the other two metric geometries in two dimensions.

The only non-trivial group of isometries that acts freely and discretely on the sphere S^2 is the two-element group $\mathbb{Z}/2\mathbb{Z}$, and so the only two factor spaces of S^2 by isometries are S^2 itself and the projective plane $\mathbb{R}P(2)$. Both of these have lots of symmetries (the isometry group is $SO(3)$), which require three continuous parameters to specify, and which act transitively on points on the surface and on directions.

Moving to the Euclidean plane, we can construct free and discrete actions of $\mathbb{Z}$ (leading to the cylinder or the Möbius strip) and $\mathbb{Z}^2$ (leading to the torus or the Klein bottle). The resulting factor spaces have fewer symmetries, which are specified by only two continuous parameters, and which act transitively on points, but not on directions.

In the hyperbolic plane, we will see that there are even more possible factor spaces (in fact, infinitely many), but they only have discrete groups of symmetries.

Lecture 27. Friday, November 6

a. A rough sketch of geometric representations of surface groups. Having constructed a geometric representation of the free group, we now do so for the surface groups

$$SG_n = \langle a_1, b_1, \dots, a_n, b_n \mid [a_1, b_1] \cdots [a_n, b_n] = e \rangle; \tag{27.1}$$

thanks to Theorem 25.1 (that we did not quite proved) and our discussion of polygonal complexes, we know that the surface group on n pairs of symbols is the fundamental group of the surface of genus n. We exhibit a free and discrete action of SG_n on the hyperbolic plane $\mathbb{H}^2$, which shows that the surface of genus n can be obtained as a factor space of $\mathbb{H}^2$ by actions of isometries. We go through the details in the case $n = 2$; the other cases are similar.

First we observe that SG_n, $n \geq 2$ cannot act freely and discretely by isometries on the Euclidean plane. For, by Theorem 6.10 any such group contains a finite index subgroup of translations that is abelian hence any two elements have commuting powers, while SG_n contains elements, say a_1 and a_2, none of whose powers commute.

Exercise 27.1. Prove that SG_n for $n \geq 2$ contains a subgroup isomorphic to F_2 and hence cannot be embedded to $Isom(\mathbb{R}^2)$, discretely or not.

Thus we turn our attention to the hyperbolic plane, and use the unit disc model.

The central region in Figure 5.10 is a hyperbolic octagon with angles equal to $\pi/4$. Denote this region by F, and label the edges of F with the symbols $a_1, b_1, c_1, d_1, a_2, b_2, c_2, d_2$, proceeding counterclockwise around the perimeter. Let f_a, f_b, f_c, f_d be the hyperbolic translations along axes through the centre of the circle that map a_1 to a_2, b_1 to b_2, and so on. The regions bordering F in Figure 5.10 (which are also hyperbolic octagons) are the images of F under the eight transformations F_σ, $\sigma \in \{a, b, c, d, a^{-1}, b^{-1}, c^{-1}, d^{-1}\}$.

Recall that in the case $n = 2$, the surface group SG_2 is isomorphic to

$$\langle a, b, c, d \mid abcda^{-1}b^{-1}c^{-1}d^{-1} = e \rangle = F_4/G,$$

where G is the smallest normal subgroup of F_4 that contains the word $abcda^{-1}b^{-1}c^{-1}d^{-1}$.

Using the hyperbolic translations just introduced, we have a natural homomorphism from $\psi\colon F_4 \to \mathrm{Isom}(\mathbb{H}^2)$; given any word $w = \sigma_1\sigma_2 \cdots \sigma_k$, where $\sigma_i \in \{a, b, c, d\}$, we associate to w the fractional linear transformation

$$\psi(w) = f_{\sigma_1} \circ \cdots \circ f_{\sigma_k}.$$

We claim that $\psi(F_4)$ is an isomorphic image of SG_2 that acts freely and discretely on $\mathbb{H}^2$. The first part of this claim requires us to show that $\ker \psi = G$, and then to apply the general result that ψ projects to an isomorphism from $SG_2 = F_4/G$ to $\psi(F_4)$.

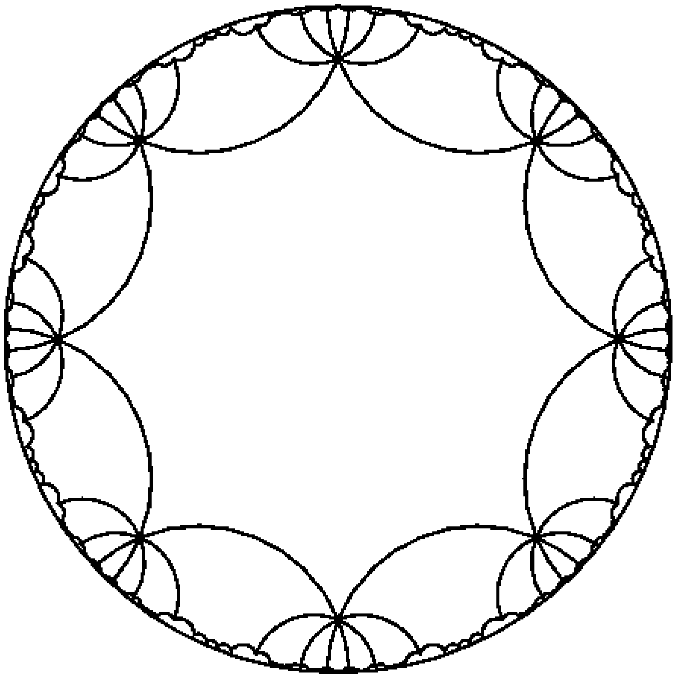


Figure 5.10. Tiling the hyperbolic plane with isometric octagons.

To establish $G \subset \ker \psi$, we must show that

$$f_a \circ f_b \circ f_c \circ f_d \circ f_a^{-1} \circ f_b^{-1} \circ f_c^{-1} \circ f_d^{-1} = \mathrm{Id}\,. \tag{27.2}$$

The reverse inclusion $\ker \psi \subset G$ can be established by induction on the length of w, proving that $\psi(w) = \mathrm{Id}$ implies $w \in G$.

It then remains to show that SG_2 acts freely and discretely on $\mathbb{H}^2$. This will follow once we show that the octagon F is a fundamental domain for SG_2; in fact, it is a Dirichlet domain (or *Voronoi domain*) of the sort defined in the previous lecture, as it comprises all points that are closer to the origin than they are to any of its images under elements of SG_2. To show that F is a fundamental domain, one must show two things:

(1) If $g \in SG_2$ is such that $gF \cap F \neq \emptyset$, then $gF = F$. This can be shown by induction on the length of g in the generators a, b, c, d.
(2) The images of $\overline{F}$ under elements of g fill $\mathbb{H}^2$.

b. Fuchsian groups. The example in the previous section is representative of an important general class.

Definition 27.1. A *Fuchsian group* is a discrete subgroup of $\mathrm{Isom}(\mathbb{H}^2)$—that is, a group that acts discretely on $\mathbb{H}^2$ by isometries.

Remark. A Fuchsian group G also acts on the ideal boundary $S^1 = \partial\mathbb{H}^2$, but this action is not discrete.

To any Fuchsian group $G \subset \mathrm{Isom}(\mathbb{H}^2)$ we can associate a fundamental domain for the action of G on $\mathbb{H}^2$. (Of course, this choice is not canonical, as there are many possible fundamental domains.)

EXAMPLE 27.2. Let $g \in \mathrm{Isom}(\mathbb{H}^2)$ be parabolic (one fixed point on the ideal boundary) or hyperbolic (two fixed points on the ideal boundary). Then $\mathbb{Z} = \{g^n\} \subset \mathrm{Isom}(\mathbb{H}^2)$ is a Fuchsian group; if g is hyperbolic, a fundamental domain for the action of $\mathbb{Z}$ on $\mathbb{H}^2$ is as shown in Figure 5.8.

EXAMPLE 27.3. In the previous lecture, we introduced hyperbolic translations that generate the free group F_2 as a discrete subgroup of $\mathrm{Isom}(\mathbb{H}^2)$. Thus F_2 is a Fuchsian group, with fundamental domain as shown in Figure 5.9.

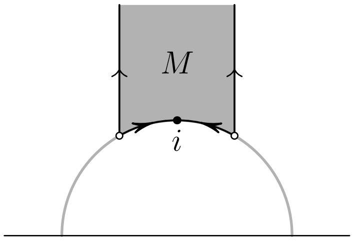


FIGURE 5.11. A fundamental domain for the modular group.

EXAMPLE 27.4. Recall that $\mathrm{Isom}^+(\mathbb{H}^2) = PSL(2,\mathbb{R})$. We can obtain a discrete subgroup by restricting the entries of the matrices to be integer-valued; thus $PSL(2,\mathbb{Z})$ is a Fuchsian group, called the *modular group*. Unlike the previous examples, it does not act freely on $\mathbb{H}^2$, as it contains finite-order elements. For example, $A = \left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$ and $B = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 1 \end{smallmatrix}\right)$ have orders 2 and 3, respectively, as elements of $PSL(2,\mathbb{Z})$, as can be seen by working either with the matrices themselves or with their corresponding fractional linear transformations $\varphi_A(z) = -1/z$ and $\varphi_B(z) = -1/(z+1)$.

In fact, $PSL(2,\mathbb{Z})$ is generated by A and B, and these two elements satisfy no relations aside from the obvious ones $A^2 = B^3 = \mathrm{Id}$. One way to show this is to work with the fundamental domain M shown in Figure 5.11, and to proceed as we did with the surface group SG_2. Identifying edges of the fundamental domain of SG_2 according to the generators of the group yielded a surface of genus two; identifying edges of M according to the generators φ_A and $\varphi_{A^{-1}B}\colon z \mapsto z+1$ yields the *modular surface*.

Topologically, the modular surface is a sphere with a point removed; geometrically, there are three special points. One of these is the point i, at which the angles of the surface add to π, not 2π, and another is the point $e^{i(\pi/3)}$, which is identified with $e^{i(2\pi/3)}$, and at which the angles only add to $2\pi/3$; these two points are called *conic points*, by analogy with the tip of a cone. The other special point is not a point on the surface at all, but rather the point "at infinity", which corresponds to a *cusp* on the surface.

REMARK. The modular group $PSL(2,\mathbb{Z})$ contains a copy of the free group on two symbols that is embedded in a different way than the example we saw earlier. In that case, the elements of the group were hyperbolic translations; it turns out that one can also embed F_2 into $PSL(2,\mathbb{Z})$ in such away that some elements of it are parabolic. We will discuss this together with a modification of the earlier construction that produce a similar effect, in the next lecture.

Once a distance function (a metric) has been properly defined on $\mathbb{H}^2$, it is possible to define an area function as well. One can show that the fundamental domain M shown in Figure 5.11 for the modular group has finite area, despite the fact that it is not compact. This represents a new sort of behaviour compared with what we are used to in the Euclidean case; if $G \subset \mathrm{Isom}(\mathbb{R}^2)$ acts discretely and $F \subset \mathbb{R}^2$ is a fundamental domain for G, then one of the following two things happens.

(1) F has infinite area and hence $\overline{F}$ is non-compact; for example, when $G = \mathbb{Z}$ acts by powers of a single translation and so $\mathbb{R}^2/G$ is a cylinder.
(2) $\overline{F}$ is compact and hence has finite area; for example, when $G = \mathbb{Z}^2$ acts by powers of two linearly independent translations and so $\mathbb{R}^2/G$ is a torus.

Both of these cases occur in Fuchsian groups acting on the hyperbolic plane: the free group F_2 has a non-compact fundamental domain with infinite area, and the surface SG_2 has a compact fundamental domain with finite area. Now, however, there is a new possibility, as exhibited by $PSL(2,\mathbb{Z})$, which has a *non-compact* fundamental domain with *finite* area. Thus we are led to consider not only Fuchsian groups with compact fundamental domain (*cocompact* groups), but also Fuchsian groups with fundamental domains having finite area (groups of *cofinite volume*).

c. More on subgroups of $PSL(2,\mathbb{R})$. The statement that the surface group SG_2 embeds isomorphically as a discrete subgroup of $PSL(2,\mathbb{R})$ is a special case of the Poincaré polygon theorem, which says (more or less) that the isometries of $\mathbb{H}^2$ that realise the edge identifications of a suitable polygon in $\mathbb{H}^2$ generate a Fuchsian group with that polygon as a fundamental domain if (in the case of free action, i.e. a group without elliptic elements) the angles around each geometrically distinct vertex add to 2π. The proof goes in three stages, the first two of which can be done by brute force for any specific case, and the third of which is more abstract:

(1) Check that the isometries identifying edges of the polygon satisfy the appropriate relation; for example, if a, b, c, d are the hyperbolic translations identifying opposite pairs of edges of the octagon, that
$$abcda^{-1}b^{-1}c^{-1}d^{-1} = e.$$
(2) Show that images of the polygon align correctly around a single vertex of the polygon; for example, under the action of the appropriate combinations of a, b, c, d, one obtains eight images of the octagon that are

adjacent to a single vertex of the original octagon, each subtending an angle of $\pi/4$.

(3) Prove and apply a general theorem due to Maskit, showing that these first two conditions imply that the isometries in question generate a discrete subgroup of $PSL(2, \mathbb{R})$—the proof, which we omit here, is essentially an inductive procedure, carried out by "growing" the tesselation beginning with the original polygon.

Before moving on to the final major topic of this course, we return once more to the free group F_2, which we embedded into $PSL(2, \mathbb{R})$ in Proposition 26.5. A fundamental domain F for this embedding was shown in Figure 5.9; F is bounded by four geodesics in $\mathbb{H}^2$ and four arcs on the ideal boundary.

Observe that by decreasing the distance that the generators ψ_A and ψ_B move the geodesics η_A and η_B bounding F, we can decrease the (Euclidean) length of the ideal arcs bounding F. In particular, if we choose ψ_A and ψ_B just right, we can make each of these arcs collapse to a single point on the ideal boundary.

For this particular choice of ψ_A and ψ_B, the fundamental domain F is topologically equivalent to a torus with a single point removed: deforming F into a (Euclidean) square, we see that ψ_A identifies the two vertical edges, and ψ_B the two horizontal edges; thus all four corners are identified into a single point, which is not actually part of F, but lies on the ideal boundary, "at infinity".

The geometric meaning of this can be seen by considering the usual torus and pulling a particular point out to infinity, creating a *cusp*, a sort of infinite "beak" on the torus, which is the same geometrically as the cusp on the modular surface. This "beaked torus" differs from the modular surface in that it has no conic points; in fact, this is a manifestation of a fundamental difference between the topology of the sphere and the topology of the torus, which we shall not get into here.

The proof of Proposition 26.5 goes through for this choice of A and B, showing that $\langle A, B\rangle$ is a free subgroup of $PSL(2, \mathbb{R})$; however, something is different now. Before, every element of $\langle A, B\rangle$ was a hyperbolic transformation, while now, the transformation $\psi_{[A,B]} = \psi_A \circ \psi_B \circ \psi_A^{-1} \circ \psi_B^{-1}$ fixes a single point on the ideal boundary (one of the four vertices of F), and hence is a parabolic transformation. Similarly, all conjugates of $[A, B]$ generate parabolic transformations, as do all the powers of $[A, B]$.

Exercise 27.2. Is this it? Is any parabolic transformation in $\langle A, B\rangle$ is a conjugate of a power of $[A, B]$?

d. The Heisenberg group and nilmanifolds. Now we consider a simplest representative of a species that will play a significant role in the next chapter. The geometric picture here will look simpler that for Fucshian groups, like that appearing in Figure 5.10; the tradeoff being higher dimension and an action that does nor preserve any natural distance function.

Recall that the Heisenberg group H_3 is the group of all 3×3 real matrices of the form

$$\begin{pmatrix}1 & x & y\\ 0 & 1 & z\\ 0 & 0 & 1\end{pmatrix}. \tag{27.3}$$

If we demand that the entries x, y, z in (27.3) be integers, we obtain a discrete subgroup $\Gamma_3 \subset H_3$. As with any subgroup of any group, Γ_3 acts on H_3 by left multiplication:

$$\begin{pmatrix}1 & m & n\\ 0 & 1 & k\\ 0 & 0 & 1\end{pmatrix} : \begin{pmatrix}1 & x & y\\ 0 & 1 & z\\ 0 & 0 & 1\end{pmatrix} \mapsto \begin{pmatrix}1 & m+x & n+y+mz\\ 0 & 1 & k+z\\ 0 & 0 & 1\end{pmatrix}. \tag{27.4}$$

This can be interpreted as an action of Γ_3 on $\mathbb{R}^3$, but *not* by isometries. It is *almost* the same as the action of $\mathbb{Z}^3$ on $\mathbb{R}^3$ by translations, but is “twisted” in the z-coordinate. A natural set of generators for Γ_3 is given by matrices

$$e = \begin{pmatrix}1 & 1 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{pmatrix}, c = \begin{pmatrix}1 & 0 & 1\\ 0 & 1 & 0\\ 0 & 0 & 1\end{pmatrix} \text{ and } f = \begin{pmatrix}1 & 0 & 0\\ 0 & 1 & 1\\ 0 & 0 & 1\end{pmatrix}. \tag{27.5}$$

Matrix c commutes with e and f and hence generates the center of N_3; direct computation shows that

$$[e, f] = c \tag{27.6}$$

providing the remaining generating relation for Γ_3. Notice that this implies that e and f generate Γ_3 In fact, f and c act by translations but e does not. As a fundamental domain for this action, we can take the unit cube $X = \{(x, y, z) \mid x, y, z \in [0, 1]\}$. Two pairs of opposite faces of the cube are identified in the standard way since f and c act by translations: f gives the identification

$$(x, y, 0) \sim (x, y, 1),$$

and c gives

$$(x, 0, z) \sim (x, 1, z).$$

The remaining two faces, however, are identified with a shear by e:

$$(0, y, z) \sim (1, y + z, z).$$

Thus the resulting quotient space $\mathbb{R}^3/\Gamma_3$ is not the three-torus $\mathbb{T}^3 = \mathbb{R}^3/\mathbb{Z}^3$, but something different. Indeed, because $\mathbb{R}^3$ is simply connected and N_3 acts freely and discretely, Theorem 23.16 shows that $\pi_1(\mathbb{R}^3/\Gamma_3) = \Gamma_3$, which is non-abelian.

Remark. Thanks to Theorem 24.2, we know that any metrisable path-connected topological group has abelian fundamental group, and hence $\mathbb{R}^3/\Gamma_3$ does not carry a group structure.

$\mathbb{R}^3/\Gamma_3$ is obtained as the quotient space of a nilpotent Lie group (H_3) by the action of a free discrete subgroup; such quotient spaces are called *nilmanifolds*, and have many nice properties.

Exercise 27.3. Prove that $A : e \to f,\ f \to e,\ c \to c^{-1}$ defines an automorphism of Γ_3 that is also an involution: $A^2 = \mathrm{Id}$.

Thus, intrinsically the non-central elements e and f are equivalent but they clearly play different role with the action on H_3. The reason is that we consider action by *left* multiplications; with the action by right multiplications the roles of n_1 and n_2 will be reversed.

Remark. We can follow the procedure in Theorem 25.1 to construct a polygonal complex whose fundamental group is Γ_3. It turns out that the resulting space is is homeomorphic to the fundamental domain $[0,1]^3$ with the same face identifications, but with the interior $(0,1)^3$ removed. That is, we obtain the two-dimensional skeleton of the above construction. Naturally one needs to add some edges to make this a legitimate polygonal complex.

CHAPTER 6

Groups at large scale

Lecture 28. Monday, November 9

a. Commensurability. In the final chapter of this course, we shall study coarse properties of finitely generated groups. The groups we consider will be countably infinite, and we will examine certain "growth properties" that characterise how these groups look at a "large scale", in a sense that will soon be made more precise. This will give us further insight into the distinction between abelian groups, free groups, and the things that lie in between.

We will be interested in coarse properties of groups, and so we need some way of "throwing away" the insignificant details, and of saying when two groups are equivalent for our purposes.

Definition 28.1. Two groups G and H are *commensurable* if there exist finite index subgroups $G_1 \subset G$ and $H_1 \subset H$ such that G_1 and H_1 are isomorphic. In this case, we write $G \simeq H$.

Proposition 28.2. *Commensurability is an equivalence relation.*

Proof. Symmetry and reflexivity are obvious; the only axiom of an equivalence relation that is not immediate is transitivity. Suppose $G \simeq H$ and $H \simeq K$, so there exist isomorphic finite index subgroups $G_1 \subset G$ and $H_1 \subset H$, and similarly for $H_2 \subset H$ and $K_2 \subset K$. Let $\phi_1 \colon H_1 \to G_1$ and $\phi_2 \colon H_2 \to K_2$ be isomorphism

Now let $H' = H_1 \cap H_2$, and observe that H' is a finite index subgroup of H. Furthermore, $G' = \phi_1(H')$ and $K' = \phi_2(H')$ are finite index subgroups of G and K, respectively. They are isomorphic to H', and hence to each other, thus $G \simeq K$. □

Example 28.3. All finite groups are commensurable to each other and to the trivial group.

Example 28.4. The infinite dihedral group is $D_\infty = \langle a, b \mid b^2 = (ab)^2 = e\rangle$. The subgroup $\langle a\rangle \subset D_\infty$ has index two and is isomorphic to $\mathbb{Z}$, thus $D_\infty \simeq \mathbb{Z}$.

Example 28.5. Let $L \subset \mathbb{R}^2$ be a rank 2 lattice. Then $\text{Isom}(L)$ is a discrete subgroup of $\text{Isom}(\mathbb{R}^2)$, and hence has a finite index subgroup of translations, which is isomorphic to $\mathbb{Z}^2$. It follows that $\text{Isom}(L)$ and $\mathbb{Z}^2$ are commensurable.

In general, every discrete subgroup of $\mathrm{Isom}(\mathbb{R}^2)$ is commensurable to $\{e\}$, $\mathbb{Z}$, or $\mathbb{Z}^2$, and none of these three groups are commensurable to each other.

b. Growth in finitely generated groups. The definition of commensurability gives a sense of the lens through which we will now look at various groups. We now introduce the yardstick by which we will measure those groups.

Let G be a finitely generated group, and let $\Gamma = \{\gamma_1, \dots, \gamma_n\}$ be a set of generators (not necessarily symmetric). Then every element $g \in G$ can be represented by a word

$$w = (\gamma_1^{k_{1,1}}\gamma_2^{k_{1,2}} \cdots \gamma_n^{k_{1,n}})(\gamma_1^{k_{2,1}}\gamma_2^{k_{2,2}} \cdots \gamma_n^{k_{2,n}}) \cdots (\gamma_1^{k_{m,1}}\gamma_2^{k_{m,2}} \cdots \gamma_n^{k_{m,n}}), \tag{28.1}$$

where the exponents $k_{i,j}$ can take any integer values. The *length* of the word w is

$$\ell(w) = \sum_{i,j} |k_{i,j}|\,;$$

of course, there are many different words that represent the element g, and so we define the length of g to be the length of the shortest word representing g:

$$\ell(g) = \min\{\ell(w) \mid w \text{ represents } g\}. \tag{28.2}$$

Remark. If we draw the Cayley graph of G using the set Γ of generators, the length of an element g is just the length of the shortest path in the Cayley graph from the identity element to g. Of course, choosing a different set of generators yields a different Cayley graph, and in our present setting, leads to a potentiall different length for the element g. Thus this notion is very dependent on our choice of Γ.

Definition 28.6. Given a group G and a set of generators Γ, the *growth function* of G and Γ evaluated at $n \in \mathbb{N}$ is the number of elements of G whose length with respect to Γ is bounded by n:

$$\mathcal{G}_{G,\Gamma}(n) = \#\{g \in G \mid \ell(g) \le n\}. \tag{28.3}$$

Example 28.7. If $G = \mathbb{Z}$ and $\Gamma = \{1\}$, then $\mathcal{G}_{G,\Gamma}(n) = 2n + 1$, since $\ell(k) = |k|$.

Example 28.8. If $G = \mathbb{Z}^2$ and $\Gamma = \{(1,0),(0,1)\}$, then $\ell((a,b)) = |a| + |b|$, and an easy computation shows that $\mathcal{G}_{G,\Gamma}(n) = 2n^2 - 2n + 1$. A similar computation and an easy induction shows that

$$\mathcal{G}_{\mathbb{Z}^k,\{\mathbf{e}_1,\dots,\mathbf{e}_k\}}(n) = 2n^k + O(n^{k-1}). \tag{28.4}$$

Exercise 28.1. Find a closed formula for $\mathcal{G}_{\mathbb{Z}^k,\{\mathbf{e}_1,\dots,\mathbf{e}_k\}}(n)$.

Example 28.9. Let a and b be the generators of the free group F_2. Observe that there is only 1 reduced word of length 0 (the identity element), there are 4 reduced words of length 1, and there are $4 \cdot 3^j$ reduced words of length j for $j \geq 2$. Thus we have

$$\mathcal{G}_{F_2,\{a,b\}}(n) = 1 + 4(1 + 3 + 3^2 + \cdots + 3^{n-1}) = 1 + 2(3^n - 1). \tag{28.5}$$

A similar computation for F_k and its natural generators $\{a_1, \dots, a_k\}$ yields the growth function

$$\mathcal{G}_{F_k,\{a_1,\dots,a_k\}}(n) = 1 + k((2k-1)^n - 1). \tag{28.6}$$

In fact, (28.6) gives a universal upper bound on the growth function of *any* group with k generators: whatever group G and generating set Γ we choose, the number of reduced words of length $\leq n$ in elements of Γ, and hence the number of elements of G with length $\leq n$, is bounded above by the expression in (28.6). However, it is important to notice that this growth function can often be made to grow quite quickly by choosing a large set of generators with few relations, and hence a large value of k.

The above examples all yielded explicit expressions for the growth function. However, such explicit expressions are not always so easy to come by, and are also quite sensitive to our choice of generators. A more robust piece of information is the asymptotic growth of $\mathcal{G}_{G,\Gamma}(n)$. So far we have seen two qualitatively different sorts of behaviour:

(1) *Polynomial growth*: There exists α such that $\mathcal{G}_{G,\Gamma}(n) = n^\alpha + o(n^\alpha)$. This is the case for $\mathbb{Z}^k$ with the standard generators and $\alpha = k$.
(2) *Exponential growth*: There exists $\lambda > 1$ such that $\mathcal{G}_{G,\Gamma}(n) = e^{\lambda n + o(n)}$. This is the case for the free group F_k with the standard generators, with

$$\lambda = \lim_{n\to\infty} \frac{1}{n} \log \mathcal{G}_{F_k,\{a_1,\dots,a_k\}}(n) = \log(2k-1). \tag{28.7}$$

The utility of this point of view can be seen by considering the surface group SG_2 with generators a, b, c, d. Explicit computation of the growth function amounts to a relatively intricate combinatorial inspection of the tesselation, which we prefer to avoid. We can work out the asymptotic behaviour with rather less effort: observing first that $\mathcal{G}_{SG_2,\{a,b,c,d\}}(n) \leq \mathcal{G}_{F_4,\{a,b,c,d\}}(n)$, we then note that the subgroup $\langle a, c\rangle \subset SG_2$ is free, and hence

$$\mathcal{G}_{SG_2,\{a,b,c,d\}}(n) \geq \mathcal{G}_{F_2,\{a,c\}}(n) = 1 + 2(3^n - 1).$$

The growth function for SG_2 is bounded above and below by growth functions with exponential growth, and hence itself has exponential growth.

Is this true of SG_2 with *any* set of generators? Or might we find some set of generators with respect to which the growth function has only polynomial growth?

First observe that for any group G, generators Γ, and $m, n \in \mathbb{N}$, every word of length $m + n$ can be decomposed into the concatentation of a word of length m and a word of length n: it follows that

$$\mathcal{G}_{G,\Gamma}(m+n) \leq \mathcal{G}_{G,\Gamma}(m)\mathcal{G}_{G,\Gamma}(n),$$

and so $\log \mathcal{G}_{G,\Gamma}(n)$ is subadditive.

The following fact is one of the most useful that appears in calculus courses. If you did not see a proof it serves as nice exercise.

Proposition 28.10. *Show that if a_n is subadditive—that is, $a_{m+n} \leq a_m + a_n$ for all $m, n \in \mathbb{N}$—then $\lim_{n\to\infty} a_n/n$ exists and is equal to $\inf_n a_n/n$ (it may be $-\infty$).*

Applying Proposition 28.10 to $\log \mathcal{G}_{G,\Gamma}(n)$, we see that the exponential growth rate

$$\lambda(G,\Gamma) = \lim_{n\to\infty} \frac{1}{n} \log \mathcal{G}_{G,\Gamma}(n) \tag{28.8}$$

exists for every group G and generators Γ.

Proposition 28.11. *Let G be a finitely generated group. If $\lambda(G,\Gamma) > 0$ for some set of generators Γ, then $\lambda(G,\mathcal{B}) > 0$ for any other set of generators $\mathcal{B}$.*

Proof. Let $\Gamma = \{\gamma_1, \dots, \gamma_k\}$ and $\mathcal{B} = \{\beta_1, \dots, \beta_m\}$ be two sets of generators of G. Write $\ell_\Gamma(g)$ and $\ell_\mathcal{B}(g)$ for the length of $g \in G$ with respected to Γ and $\mathcal{B}$, respectively. Let $\omega = \max_i \ell_\mathcal{B}(\gamma_i)$, and observe that every word of length $\leq n$ in elements of Γ can be written as a word of length $\leq \omega n$ in elements of $\mathcal{B}$. It follows that

$$\mathcal{G}_{G,\Gamma}(n) \leq \mathcal{G}_{G,\mathcal{B}}(\omega n),$$

and hence

$$\lambda(G,\mathcal{B}) = \lim_{n\to\infty} \frac{1}{\omega n} \log \mathcal{G}_{G,\mathcal{B}}(\omega n) \geq \frac{1}{\omega} \lim_{n\to\infty} \frac{1}{n} \log \mathcal{G}_{G,\Gamma}(n) = \frac{1}{\omega} \lambda(G,\Gamma) > 0.$$

□

As a consequence of Proposition 28.11, the statement that a group G has exponential growth is independent of what set of generators we choose, as is the statement that G has subexponential growth ($\lambda = 0$). However, the exponent λ can vary if we choose a different set of generators.

Within the class of groups with subexponential growth, we have seen polynomial growth occurring for abelian groups. We will see that in fact, any nilpotent group has polynomial growth. Furthermore, a deep result due to Gromov states that a group has polynomial growth if and only if it is *virtually nilpotent*—that is, it is commensurable to a nilpotent group. Unfortunately the proof of this result that features some of the most striking applications of "soft" analysis argument to algebra, lies beyond the scope of these lectures.

Lecture 29. Wednesday, November 11

a. Different growth rates. In the previous lecture, we considered the growth function for finitely generated groups G with generating set Γ. We saw that although the exact value of the exponential growth rate $\lambda(G, \Gamma)$ defined in (28.8) may depend on the choice of generators, the dichotomy between exponential growth ($\lambda > 0$) and subexponential growth ($\lambda = 0$) is valid independently of what generators we choose (Proposition 28.11). Thus we may divide the class of finitely generated groups into two classes: groups with exponential growth and groups with subexponential growth.

So far, we have seen two basic classes of groups with exponential growth up to commensurability: free groups and surface groups. We have seen one basic class of examples of groups with sub-exponential growth, again up to commensurability: the groups $\mathbb{Z}^k$. One way to obtain more examples is to consider direct products. This gives new examples in the exponential case but so far not in the polynomial case. In the exponential case one can also consider free products of previously constructed groups.

The distinction between exponential and subexponential growth is reminiscent of the situation we see when we consider a matrix $A \in SL(n, \mathbb{R})$ acting on $\mathbb{R}^n$. Given a vector $\mathbf{v} \in \mathbb{R}^n$, we may consider the growth rate of the quantity $\|A^k \mathbf{v}\|$—that is, the norm of the image of $\mathbf{v}$ under k iterations of the linear map A. There are three possibilities:

(1) *No growth*: $\|A^k \mathbf{v}\|$ is uniformly bounded for all $k \in \mathbb{Z}$. This occurs if $\mathbf{v}$ lies in the linear span of eigenspaces corresponding to eigenvalues on the unit circle for which there are no Jordan blocks. Thus A acts on a subspace containing $\mathbf{v}$ as a product of rotations; in particular, an isometry.

(2) *Polynomial growth*: $\|A^k \mathbf{v}\| \approx k^\alpha$ for some $\alpha \in \mathbb{N}$. This occurs if $\mathbf{v}$ lies in the linear span of eigenspaces corresponding to eigenvalues on the unit circle for which there *are* Jordan blocks. If m is the size of the largest Jordan block, then $\alpha = m - 1$.

(3) *Exponential growth in either one or both time directions*: This occurs if $\mathbf{v}$ has a non-zero component in the direction of an eigenvector for an eigenvalue not lying on the unit circle.

EXERCISE 29.1. Construct a sequence of real numbers x_n such that both of the following hold simultaneously:

(1) x_n grows *super-polynomially*—that is, for every $\alpha > 0$, $\lim_{n\to\infty} x_n/n^\alpha = \infty$.

(2) x_n grows subexponentially—that is, for every $\lambda > 0$, $\lim_{n\to\infty} x_n/e^{\lambda n} = 0$.

A sequence such as the one in Exercise 29.1 is said to have *intermediate growth*. It follows from the discussion above that intermediate growth does not occur for linear maps—that is, every linear map with subexponential growth actually has polynomial growth. So far, the only examples we have seen of groups with subexponential growth are the abelian groups $\mathbb{Z}^k$, which turn out to have polynomial growth.

There are a number of natural questions to ask at this point.

(1) Is polynomial growth independent of the choice of generators?
(2) Does every abelian group have polynomial growth? And, if yes, are they characterized by the growth rate up to commensurability?
(3) Are there non-abelian groups with polynomial growth?
(4) Does intermediate growth even exist for groups? Or are they like linear maps in displaying only polynomial growth and exponential growth, with nothing in between?
(5) Can we characterise the class of groups with polynomial growth? Do the properties of polynomial growth, exponential growth, and intermediate growth (if such examples exist) have algebraic significance?

The last two questions turn out to be harder than the other three, and so we will postpone them. For the time being, we will address the first three questions in turn.

b. Groups with polynomial growth. We say that (G, Γ) has polynomial growth of degree $\alpha > 0$ if there exists $C > 0$ such that $\mathcal{G}_{G,\Gamma}(n) \leq Cn^\alpha$ for all n, and if this property fails for all smaller values of α.

Remark. If G has an element of infinite order, then we have $\alpha \geq 1$. It is natural to ask if there exist groups with polynomial growth of degree α with $0 < \alpha < 1$: clearly such groups must have infinitely many elements, all of finite order. Groups with these properties do exist, but we do not construct them here.

Proposition 28.11 showed that the property of exponential growth is independent of the choice of generators. A similar statement is true for polynomial growth: in fact, the degree is also independent.

Proposition 29.1. *If Γ is a finite generating set for G and (G, Γ) has polynomial growth of degree α, then $(G, \mathcal{B})$ has polynomial growth of degree α for any other finite generating set $\mathcal{B}$.*

Proof. We use a similar approach to the proof of Proposition 28.11. Consider $\omega = \max_i \ell_\Gamma(\beta_i)$, and observe that $\mathcal{G}_{G,\mathcal{B}}(n) \leq \mathcal{G}_{G,\Gamma}(\omega n)$. Consequently, if $\alpha > 0$ and $C > 0$ are such that $\mathcal{G}_{G,\Gamma}(n) \leq Cn^\alpha$ for all n, then

$$\mathcal{G}_{G,\mathcal{B}}(n) \leq C\omega^\alpha n^\alpha$$

for all n, and hence $(G, \mathcal{B})$ has polynomial growth of degree at most α. A symmetric proof establishes the other inequality. □

This gives a positive answer to the first question above: polynomial growth, and even the degree of that growth, is independent of the choice of generators, so we can speak of “a group with polynomial growth of degree α” without fear of ambiguity due to not specifying a choice of generators.

c. Abelian groups. In the previous lecture, we studied the growth functions for the abelian groups $\mathbb{Z}^k$. It turns out that up to commensurability, these are the only finitely generated abelian groups.

Definition 29.2. Let G be a group. If $g \in G$ has finite order, we say that it is a *torsion element*. If every element of G is a torsion element, then G is a *torsion group*. If the only torsion element is the identity, then G is *torsion-free*.

Theorem 29.3. *Every torsion-free finitely generated abelian group is isomorphic to $\mathbb{Z}^k$ for some k.*

Proof. Given a torsion-free abelian group $G = \langle g_1, \dots, g_n \rangle$, we must show that $G \cong \mathbb{Z}^k$ for some $k \leq n$. Note that k may be strictly less than n; for example, one may consider $\mathbb{Z}^2 = \langle (1,0), (1,1), (0,1) \rangle$, in which case we have $k = 2$ and $n = 3$.

The proof uses geometric methods and goes by induction in n. The case $n = 1$ is easy: if $G = \{g_1\}$ and a_1 has infinite order, then $G \cong \mathbb{Z}$. So now suppose that the result holds for groups with n generators, and suppose $G = \langle g_1, \dots, g_{n+1} \rangle$.

Because G and $\mathbb{Z}^k$ are both abelian, we will use additive notation—that is, given elements $g, h \in G$ and integers a, b, we will write $ag + bh$ instead of $g^a h^b$.

We say that an element $g \in G$ is *prime* in G if it is not obtained from any other single element—that is, if there do not exist $h \in G$ and $a > 1$ such that $ah = g$. Observe that $(m_1, \dots, m_k)$ is prime in $\mathbb{Z}^k$ if and only if the greatest common divisor of $m_1, \dots, m_k$ is 1.

Lemma 29.4. *If $\mathbf{m} \in \mathbb{Z}^k$ is prime in $\mathbb{Z}^k$, then there exists a basis of $\mathbb{Z}^k$ containing $\mathbf{m}$—that is, there exist $\mathbf{m}_1, \dots, \mathbf{m}_k$ such that $\mathbf{m}_1 = \mathbf{m}$ and $\langle \mathbf{m}_1, \dots, \mathbf{m}_k \rangle = \mathbb{Z}^k$. Equivalently, there is a matrix $A \in SL(k, \mathbb{Z})$ such that the first row of A is the vector $\mathbf{m}$.*

Proof. Consider the line P_1 in $\mathbb{R}^k$ given by $P_1 = \{t\mathbf{m} \mid t \in \mathbb{R}\}$. Because the slope of this line is rational, there exists a lattice point $\mathbf{m}_2 \in \mathbb{Z}^k$ such that $d(\mathbf{m}_2, P_1) \leq d(\mathbf{m}', P_1)$ for all $\mathbf{m}' \in \mathbb{Z}^k$.

Now consider the plane $P_2 = \{t_1\mathbf{m}_1 + t_2\mathbf{m}_2 \mid t_1, t_2 \in \mathbb{R}\}$. The lattice $L_2 = \{n_1\mathbf{m}_1 + n_2\mathbf{m}_2 \mid n_1, n_2 \in \mathbb{Z}\} = P_2 \cap \mathbb{Z}^k$. For, otherwise there would a point $\mathbb{Z}^k \ni \mathbf{x} = t_1\mathbf{m}_1 + t_2\mathbf{m}_2, 0, < t_2 < 1$ that is closer to the line P_1 than $\mathbf{m}_2$.

Once again, let $\mathbf{m}_3 \in \mathbb{Z}^k$ be the closest lattice point to P_2 and let
$P_3 = \{t_1\mathbf{m}_1 + t_2\mathbf{m}_2 + t_3\mathbf{m}_3 \mid t_1, t_2, t_3 \in \mathbb{R}\}$ The same argument using fundamental domains shows that

$$L_3 = \{n_1\mathbf{m}_1 + n_2\mathbf{m}_2 + n_3\mathbf{m}_3 \mid n_1, n_2, n_3 \in \mathbb{Z}\} = P_3 \cap \mathbb{Z}^k$$

Proceeding by induction we obtain $P_k = \mathbb{R}^k$ and

$$L_k = \{n_1\mathbf{m}_1 + n_2\mathbf{m}_2 + \dots + n_k\mathbf{m}_k \mid n_1, n_2, \dots, n_k \in \mathbb{Z}\} = P_k \cap \mathbb{Z}^k = \mathbb{Z}^k.$$

□

Armed with Lemma 29.4, we can carry out the inductive step. Given a torsion-free abelian group $G = \langle g_1, \dots, g_{n+1} \rangle$, we have by the inductive hypothesis that $\langle g_1, \dots, g_n \rangle$ is isomorphic to $\mathbb{Z}^k$ for some $k \leq n$. Now there are two cases.

Case 1. Suppose $ag_{n+1} \notin \langle g_1, \dots, g_n \rangle$ for all non-zero integers a. Then mapping g_{n+1} to $\mathbf{e}_{k+1}$ extends the isomorphism between $\langle g_1, \dots, g_n \rangle$ and $\mathbb{Z}^k$ to an isomorphism between G and $\mathbb{Z}^{k+1}$.

Case 2. Suppose $a > 0$ is the smallest positive integer such that $ag_{n+1} \in \langle g_1, \dots, g_n \rangle$. Now let $h \in \langle g_1, \dots, g_n \rangle$ be prime in $\langle g_1, \dots, g_n \rangle$ such that $ag_{n+1} = bh$ for some integer b. Since $\langle g_1, \dots, g_n \rangle \cong \mathbb{Z}^k$, Lemma 29.4 proves the existence of a basis $\{\mathbf{m}_1, \dots, \mathbf{m}_k\}$ for $\mathbb{Z}^k$ such that $\mathbf{m}_1$ corresponds to h.

Now observe that under the correspondence between $\langle g_1, \dots, g_n \rangle$ and $\mathbb{Z}^k$, the element ag_{n+1} corresponds to $b\mathbf{m}_1$. Let $\mathbf{m}_1' = \frac{b}{a}\mathbf{m}_1 \in \mathbb{R}^k$, and observe that $\langle \mathbf{m}_1', \mathbf{m}_2, \dots, \mathbf{m}_k \rangle$ is a subgroup of $\mathbb{R}^k$ that is isomorphic to $\mathbb{Z}^k$ and that is also isomorphic to $\langle g_1, \dots, g_{n+1} \rangle$. □

If G is a finitely generated abelian group that contains elements of finite order, one can consider the *torsion subgroup* $T \subset G$ that comprises all elements of finite order. (Observe that T is a subgroup because two commuting elements of finite order have a product that is also of finite order; this is not necessarily true if the elements do not commute.) Then G/T is a torsion-free finitely generated abelian group, so by Theorem 29.3 it is isomorphic to $\mathbb{Z}^k$. One can show that G/T is also isomorphic to a finite index subgroup of G, and it follows that G and $\mathbb{Z}^k$ are commensurable.

Summarising, Theorem 29.3 tells us that up to commensurability, every finitely generated abelian group is equivalent to $\mathbb{Z}^k$. In particular, every finitely generated abelian group G has polynomial growth of degree k, where k is the so-called "torsion-free rank" of G. Thus we see that the degree of polynomial growth completely classifies finitely generated abelian groups up to commensurability. This is a stark contrast from the case for the free groups F_k, where the growth rate alone cannot distinguish F_2 from F_3, or from any other F_k.

d. Nilpotent groups. Having settled the first and second questions on our original list, we turn our attention to the third: are there non-abelian groups that also have polynomial growth? It turns out that there are.

Theorem 29.5. *Any finitely generated nilpotent group has polynomial growth.*

The proof goes by induction in nilpotent length. The base of induction is nilpotent length one, i.e. abelian groups, which follows from Theorem 29.3 and the results on growth of $\mathbb{Z}^k$.

Proof for nilpotent length ≤ 2. We start with the case of nilpotent length two to explain the ideas and show how they work for the example given above. After that we present the general induction step.

Assume G is such a group with m generators, $g_1, \dots, g_m$. Then the commutator $[G, G]$ is abelian and belongs to the center of G. Take a product of n generators. Exchanging any two generators produces a commutator on the right. Since commutators lie in the center they can be moved to the right. In order to move generators to a canonical order one needs thus no more than n^2 interchanges. Thus we obtain a word of the form $g_1^{k_1} g_2^{k_2} \dots g_m^{k_m} C$, where C is the product of no more that n^2 commutators of the generators. Those commutators are words of bounded length with respect to any system of generators in the abelian group $[G, G]$ with polynomial growth of degree, say, k. Thus the growth of G is polynomial of degree at most $m + 2k$. □

Remark. Notice that $m \geq 2$ since any group with one generator is abelian, and $k \geq 1$; otherwise commutator is finite and G can be shown to have an abelian subgroup of finite index. Hence the minimal value of our above estimate for the growth in a non-abelian nilpotent group is four.

Example 29.6. (See Section 27.d) Let Γ_3 be the group of upper-triangular unipotent 3×3 matrices with integer entries (this is a subgroup of the Heisenberg group). Γ_3 is generated by

$$e = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad f = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix},$$

and has a one-dimensional center generated by

$$[e, f] = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

As our above estimate shows the polynomial growth in this group has degree ≤ 4. Let us show that this growth is actually achieved. Consider a word w that contains k e's and l f's and no inverses, Let $I(w)$ be the number of paris (e, f) in w where e precedes f. Immediate calculation shows that w corresponds to the matrix

$$\begin{pmatrix} 1 & k & I(w) \\ 0 & 1 & l \\ 0 & 0 & 1 \end{pmatrix}.$$

For fixed value of k and l $a(w)$ can take any value between 0 and kl. Thus

$$\mathcal{G}_{\Gamma_3, <e,f>}(n) > \sum_{k,l,k+l \leq n} kl > Cn^4,$$

where the constant C can be easily estimated from below, say $C > 2^{-6}$ by taking k and l between $n/4$ and $3n/4$.

This is the slowest possible growth in a nilpotent group which does not have an abelian subgroup of finite index.

Proof for the general case. Assume G has nilpotent length s and as before has m generators $g_1, \dots, g_m$. Then $[G,G]$ has nilpotent length $\leq s-1$ and hence by the inductive assumption has polynomial growth of degree, say k.

As before, consider a product w of n generators and try to bring it to a form $g_1^{k_1} g_2^{k_2} \dots g_m^{k_m} C$, where $C \in [G,G]$ and estimate the length of C. Exchanging a pair of generators produces a commutator on the right; as before there will be no more than n^2 such commutators in the process of rearranging the generators. But this time when we move generators to the left we need to exchange them with the commutators thus producing elements of the from $[g_{i_1}, [g_{i_2}, g_{i_3}]] \in [G, [G,G]]$, the total of no more of n^3, and so on. Since G has nilpotent length s this process of generating new terms will stop at s-th level, i.e. moving generators through commutators of i-th order with not produce any new terms. Thus the total length of C is estimated from above by $const \cdot n^s$ since there are at most $n^2 + \dots + n^s$ commutators of different orders and each of them is a word of bounded length. Thus the growth of G is polynomial of degree at most $m + sk$. $\square$

The above proof gives an above estimate of the degree of growth of G.

Example 29.7. Consider the group Γ_4 of of upper-triangular unipotent 4×4 matrices with integer entries. It has three generators

$$e_{12} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \ e_{23} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \text{ and } e_{34} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

its nilpotent length is three and its commutant is isomorphic to Γ_3. This gives $3 + 3 \cdot 4 = 15$ as an above estimate on the degree of of growth for Γ_4. However, a more accurate count shows that the actual degree is 10.

Lecture 30. Friday, November 13

a. The automorphism group. Let us return now to a more general discussion. Broadly speaking, one of the goals of the endeavour in which we are presently embroiled is to understand "just what groups are actually out there". One way of doing this is to characterise certain interesting properties that distinguish between various sorts of groups: polynomial growth and exponential growth are examples of this approach.

Another approach is to construct lots of interesting examples of groups with different algebraic structure, and thus to gain our insight from more concrete cases. So how do we construct examples of groups? The technique of group presentations is a powerful way to specify very many interesting groups by giving a set of generators and a set of relations. However, it suffers from a drawback: it is often quite hard to tell just what group we actually have our hands on. For example, here are two finitely presented groups:

$$G = \langle a, b, c \mid a^2 = b^2 = c^2 = e, ac = ca, aba = bab, bcb = cbc \rangle,$$
$$H = \langle x_1, x_2, x_3 \mid x_i^2 = (x_i x_j)^3 = (x_i x_j x_k)^4 = e \text{ for all } i \neq j, i \neq k, j \neq k \rangle.$$

Are G and H isomorphic? It is not at all obvious how we ought to go about answering this question: in fact, it has been proved that there does not exist an algorithm for determing when two finitely presented groups are isomorphic.

One situation where we are on better footing is when a finitely presented group turns out to be isomorphic to a group we have already studied in some more concrete realisation.

Exercise 30.1. Show that G and H above are both isomorphic to S_4, and hence are isomorphic to each other.

Another problem with using the presentation method is that it does not behave well with respect to commensurability. For example, F_2 is isomorphic to $SL(2, \mathbb{Z})$; standard presentation for the latter do not seem to indicate this.

Since group presentations are so slippery, we may well search for more transparent ways of producing new groups from old ones. We have already seen two important constructions: the direct product and the free product. In this section, we consider a third construction, which builds a (possibly new) group not as a "product" of two others, but in terms of a single group. This will be later used in the construction of a "semi-direct" product of two groups (that requires some extra information) that is more flexible than the straightforward direct product but still connects the groups more tightly than free product.

Recall that an automorphism of a group G is a bijective homomorphism from G to itself. The set of all automorphisms of G forms a group, which we denote $\mathrm{Aut}(G)$. Every element $g \in G$ induces an automorphism $\psi_g \in \mathrm{Aut}(G)$ by conjugation: $\psi_g(h) = g^{-1}hg$. Automorphisms of this form are called *inner automorphisms*; the set of all inner automorphisms is a subgroup of $\mathrm{Aut}(G)$, which we denote $\mathrm{Inn}(G)$.

Inner automorphisms are very helpful in understanding the internal structure of G. An outstanding example concerns continuous groups, such as matrix groups, and, more generally, Lie groups. The action of a matrix group on itself by inner automorphisms is called the *adjoint representation*. It, and its derivative, the adjoint representation of the Lie algebra, play the central role in the structural theory of Lie groups and Lie algebras.

However, we do not obtain any new groups by looking at inner automorphisms, since $\mathrm{Inn}(G)$ is a homomorphic image of G itself. Indeed, the map $g \mapsto \psi_g$ is a homomorphism from G onto $\mathrm{Inn}(G)$, whose kernel is $Z(G)$, the center of G, and thus

$$\mathrm{Inn}(G) = G/Z(G). \tag{30.1}$$

Example 30.1. Consider the abelian group $\mathbb{Z}^k$: because the centre is everything, the inner automorphism group is trivial. Nevertheless, there are many automorphisms of $\mathbb{Z}^k$:

$$\mathrm{Aut}(\mathbb{Z}^k) = GL(k,\mathbb{Z}) = \{A \in M_n(\mathbb{Z}) \mid \det A = \pm 1\}. \tag{30.2}$$

If k is odd, then $GL(k,\mathbb{Z})$ is the direct product of $SL(k,\mathbb{Z})$ and $\mathbb{Z}/2\mathbb{Z}$. If k is even, the relationship between $GL(k,\mathbb{Z})$ and $SL(k,\mathbb{Z})$ is slightly more complicated (but only slightly).

Example 30.1 shows that interesting new groups can be constructed as the automorphism groups of already familiar groups. One may ask how $\mathrm{Aut}(\mathbb{Z}^k) = GL(k,\mathbb{Z})$ fits into the theory of group presentations: is it finitely generated? Finitely presented? We will see later on that this turns out to be the case.

b. Outer automorphisms. For non-abelian groups G, the inner automorphism group $\mathrm{Inn}(G)$ is non-trivial, and so we would like to construct a new group by factoring $\mathrm{Aut}(G)$ by $\mathrm{Inn}(G)$. To do this, we must first verify that $\mathrm{Inn}(G)$ is normal.

Proposition 30.2. $\mathrm{Inn}(G)$ *is a normal subgroup of* $\mathrm{Aut}(G)$*.*

Proof. Given $g \in G$, consider the corresponding inner automorphism $\psi_g \colon h \mapsto g^{-1}hg$, and let $\phi \in \mathrm{Aut}(G)$ be an arbitrary automorphism (not necessarily inner). Then for all $h \in G$, we have

$$(\phi \circ \psi_g \circ \phi^{-1})(h) = \phi(g^{-1}\phi^{-1}(h)g) = \phi(g)^{-1}h\phi(g) = \psi_{\phi(g)}(h),$$

and it follows that $\phi \circ \psi_g \circ \phi^{-1} = \psi_{\phi(g)} \in \mathrm{Inn}(G)$. □

Definition 30.3. The quotient group $\mathrm{Out}(G) = \mathrm{Aut}(G)/\mathrm{Inn}(G)$ is called the *outer automorphism group* of G.

Remark. In day-to-day usage, we often refer to an automorphism that is not an inner automorphism as an *outer automorphism*. However, strictly speaking, the elements of the outer automorphism group are actually *cosets* of the inner automorphism group, rather than individual "outer automorphisms".

EXAMPLE 30.4. Since $\mathbb{Z}^k$ is abelian, we have $\mathrm{Out}(\mathbb{Z}^k) = \mathrm{Aut}(\mathbb{Z}^k) = GL(k, \mathbb{Z})$.

The outer automorphism groups of the free groups F_k and the surface groups SG_n turn out to be exceedingly interesting objects.

$\mathrm{Out}(SG_n)$ is the so-called *mapping class group*, which plays a role in *Teichmüller theory.* In the case $n = 1$, this is just $GL(2, \mathbb{Z})$, since $SG_2 = \mathbb{Z}^2$, but for larger values of n, $\mathrm{Out}(SG_n)$ is a completely different beast than the linear groups $GL(k, \mathbb{Z})$. Thus $SL(2, \mathbb{Z})$ has two natural generalisations to "higher dimensions": $SL(k, \mathbb{Z})$ and $\mathrm{Out}(SG_n)$.

EXERCISE 30.2. Let Γ_3 once again be the group of upper-triangular unipotent matrices with integer entries. Show that $\mathrm{Inn}(\Gamma_3) = \mathbb{Z}^2$. What are $\mathrm{Aut}(\Gamma_3)$ and $\mathrm{Out}(\Gamma_3)$?

Lecture 31. Monday, November 16

a. The structure of $SL(2,\mathbb{Z})$. In the previous lecture, we obtained the groups $GL(n,\mathbb{Z})$ (and hence their index 2 subgroups $SL(n,\mathbb{Z})$) by considering the automorphism groups $\mathrm{Aut}(\mathbb{Z}^n)$. In this lecture, we study these groups further, beginning with the case $n=2$.

PROPOSITION 31.1. *$SL(2,\mathbb{Z})$ is generated by the following matrices:*

$$A=\begin{pmatrix}1&1\\0&1\end{pmatrix},\quad B=\begin{pmatrix}0&-1\\1&0\end{pmatrix}. \tag{31.1}$$

PROOF. Recall the classical Euclidean algorithm: one begins with two positive integers a and b, and transforms the pair (a,b) into the pair $(\gcd(a,b),0)$ through successive applications of the transformation

$$(a,b)\mapsto\begin{cases}(a,a-b) & a\geq b,\\ (b,a) & a<b.\end{cases} \tag{31.2}$$

We follow a similar procedure here, replacing the integers a,b with vectors $\mathbf{x},\mathbf{y}\in\mathbb{Z}^2$, and using multiplication by the matrices A,B as the analogue of the two transformations in (31.2).

Let us make this precise. Begin with a matrix $X\in SL(2,\mathbb{Z})$, and let $\mathbf{x}=\left(\begin{smallmatrix}x_1\\x_2\end{smallmatrix}\right),\mathbf{y}=\left(\begin{smallmatrix}y_1\\y_2\end{smallmatrix}\right)\in\mathbb{Z}^2$ be the column vectors of X. Write $p(\mathbf{x})=\min(|x_1|,|x_2|)$ for the distance from $\mathbf{x}$ to the nearest axis, and also write $P(X)=p(\mathbf{x})$. Observe that $P(X)$ does not depend on the second column of X. The proof is by induction on $P(X)$: essentially we perform the Euclidean algorithm on the pair (x_1,x_2), and bring the pair (y_1,y_2) along for the ride.

First observe that the matrices B^iX for $i\in\{0,1,2,3\}$ have the forms

$$\begin{pmatrix}x_1&y_1\\x_2&y_2\end{pmatrix},\quad\begin{pmatrix}-x_2&-y_2\\x_1&y_1\end{pmatrix},\quad\begin{pmatrix}-x_1&-y_1\\-x_2&-y_2\end{pmatrix},\quad\begin{pmatrix}x_2&y_2\\-x_1&-y_1\end{pmatrix}. \tag{31.3}$$

For the base case of the induction, we use the fact that $P(X)\geq 0$ for every matrix $X\in SL(2,\mathbb{Z})$, and that if $P(X)=0$, then either X or BX is upper-triangular. Upper-triangular matrices in $SL(2,\mathbb{Z})$ have integer entries equal to ±1, and hence one of the matrices B^iX is of the form $\left(\begin{smallmatrix}1&k\\0&1\end{smallmatrix}\right)$ for some $k\in\mathbb{Z}$. Furthermore, this matrix is equal to A^k, and it follows that $X=B^{-i}A^k\in\langle A,B\rangle$.

Now for the induction step. For any 2×2 matrix T, we have $TX=T(\mathbf{x}\ \mathbf{y})=(T\mathbf{x}\ T\mathbf{y})$, and consequently, $P(TX)=p(T\mathbf{x})$. Furthermore, $A^k\mathbf{x}=\left(\begin{smallmatrix}x_1+kx_2\\x_2\end{smallmatrix}\right)$. Thus if $|x_1|\geq|x_2|$, then we choose $k\in\mathbb{Z}$ such that $|x_1+kx_2|<|x_2|$, and we see that

$$P(A^kX)=p(A^k\mathbf{x})=|x_1+kx_2|<|x_2|=p(\mathbf{x})=P(X).$$

If $|x_1|<|x_2|$, then $B\mathbf{x}=\left(\begin{smallmatrix}-x_2\\x_1\end{smallmatrix}\right)$, and the same argument shows that

$$P(A^kBX)=p(A^kB\mathbf{x})<p(\mathbf{x})=P(X).$$

By the inductive hypothesis, then, either A^kX or A^kBX lies in $\langle A,B\rangle$, and so $X\in\langle A,B\rangle$ as well. This completes the proof that $\langle A,B\rangle=SL(2,\mathbb{Z})$. □

It turns out that $SL(2,\mathbb{Z})$ is not far from being free. Of course, it has a non-trivial centre $Z(SL(2,\mathbb{Z})) = \{\pm I\}$, and so we consider instead the projective group $PSL(2,\mathbb{Z}) = SL(2,\mathbb{Z})/\{\pm I\}$. Consider the following subgroup of $PSL(2,\mathbb{Z})$:

$$\begin{aligned}\Gamma(2) &= \{X \in PSL(2,\mathbb{Z}) \mid X \equiv I \bmod 2\} \\ &= \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in PSL(2,\mathbb{Z}) \,\middle|\, a, d \text{ are odd and } b, c \text{ are even} \right\}.\end{aligned} \tag{31.4}$$

EXERCISE 31.1. Show that $\Gamma(2) = \langle S, T\rangle$, where $S = \left(\begin{smallmatrix} 1 & 2 \\ 0 & 1 \end{smallmatrix}\right)$ and $T = \left(\begin{smallmatrix} 1 & 0 \\ 2 & 1 \end{smallmatrix}\right)$. *(Hint: mimic the proof of Proposition 31.1).*

EXERCISE 31.2. Show that $\Gamma(2)$ is isomorphic to F_2 with generators S and T—that is, show that there are no non-trivial relations between S and T. *(Hint: mimic the proof of Proposition 26.5, replacing the regions $D_A, D_B, D_{A^{-1}}, D_{B^{-1}} \subset \mathbb{H}^2$ with the four "quadrants" in $\mathbb{R}^2$ bounded by the lines $y = \pm x$.)*

EXERCISE 31.3. Show that $\Gamma(2)$ has index 6 in $PSL(2,\mathbb{Z})$, and that we can take the coset representatives to be

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}, \quad \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}, \quad \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

It follows from the above exercises that $SL(2,\mathbb{Z})$ has a free subgroup of index 12, and hence $GL(2,\mathbb{Z})$ has a free subgroup of index 24. In particular, each of these groups is commensurable to F_2.

Not only does $PSL(2,\mathbb{Z})$ have a free subgroup of index 6, but it itself can be written as a free product. Indeed, using the fact that $SL(2,\mathbb{Z}) = \langle A, B\rangle$, where A, B are as in Proposition 31.1, and abusing notation slightly by also writing A and B for the corresponding elements of $PSL(2,\mathbb{Z})$, we have $PSL(2,\mathbb{Z}) = \langle A, B\rangle = \langle B, BA\rangle$, where

$$BA = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & -1 \\ 1 & 1 \end{pmatrix}.$$

Observe that as elements of $PSL(2,\mathbb{Z})$, B has order 2 and BA has order 3: furthermore, it can be shown (though we do not do so), that A and BA do not satisfy any other relations. Thus we have

$$PSL(2,\mathbb{Z}) \cong (\mathbb{Z}/2\mathbb{Z}) * (\mathbb{Z}/3\mathbb{Z}), \tag{31.5}$$

which is the smallest non-trivial example of a free product (the free product of two copies of $\mathbb{Z}/2\mathbb{Z}$ turns out not to be terribly complicated).

Returning to $SL(2,\mathbb{Z})$ itself, one can show that writing g and h for abstract generators corresponding to B and BA, respectively, we have

$$SL(2,\mathbb{Z}) = \langle g, h \mid g^4 = h^6 = e, g^2 = h^3\rangle. \tag{31.6}$$

This gives a finite presentation of $SL(2,\mathbb{Z})$, and also lets us write it as something which is very nearly the free product of $\mathbb{Z}/4\mathbb{Z}$ and $\mathbb{Z}/6\mathbb{Z}$, but not quite. If we think of the free product of two groups as being obtained by "gluing" the groups together at the identity and nowhere else, then this construction corresponds to gluing the two groups not just at the identity, but also at the element $g^2 = h^3$. The "gluing set" $\{e, g^2 = h^3\}$ corresponds to a subgroup of order two in both $\mathbb{Z}/4\mathbb{Z}$ and $\mathbb{Z}/6\mathbb{Z}$: this is an example of a general construction, where we take a *free product with respect to a subgroup*, what is sometimes called a *free product with amalgamation.* In this case, it amounts to taking groups G and H together with isomorphic normal subgroups $G_1 \subset G$ and $H_1 \subset H$, and then constructing the quotient group $G{*}H/N$, where N is the smallest normal subgroup of $G{*}H$ containing

$$\{g\phi(g)^{-1} \mid g \in G_1\},$$

with ϕ the isomorphism between G_1 and H_1.

b. The space of lattices. Earlier in the course, we spent a decent bit of time with the various algebraic and geometric properties of various two-dimensional tori. These tori were obtained as factors $\mathbb{R}^2/L$, where

$$L = L(\mathbf{u}, \mathbf{v}) = \{a\mathbf{u} + b\mathbf{v} \mid a, b \in \mathbb{Z}\} \tag{31.7}$$

is the lattice generated by two linearly independent vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^2$. Because $\mathbb{R}^2$ is abelian, the lattice L is a normal subgroup of $\mathbb{R}^2$, and hence the torus $\mathbb{R}^2/L$ is a topological group.

The general narrative of all this is repeated in the present setting: $SL(n, \mathbb{Z})$ is a discrete subgroup of $SL(n, \mathbb{R})$, and so we may consider the quotient $SL(n, \mathbb{R})/SL(n, \mathbb{Z})$. In this case, however, $SL(n, \mathbb{Z})$ is *not* a normal subgroup, and so the quotient is not a group. Rather, the collection of cosets of $SL(n, \mathbb{Z})$ in $SL(n, \mathbb{R})$ forms a *homogeneous space*, which inherits a topological and geometric structure from $SL(n, \mathbb{R})$ in the same way that $\mathbb{R}^n/\mathbb{Z}^n$ inherits a topological and geometric structure from $\mathbb{R}^n$.

In fact, the relationship with the example of the torus goes even deeper than this superficial similarity. Given n linearly independent vectors $\mathbf{v}_1, \dots, \mathbf{v}_n \in \mathbb{R}^n$, we have a lattice

$$L = L(\mathbf{v}_1, \dots, \mathbf{v}_n) = \{a_1\mathbf{v}_1 + \dots + a_n\mathbf{v}_n \mid a_1, \dots, a_n \in \mathbb{Z}\}, \tag{31.8}$$

and we can consider the factor torus $\mathbb{R}^n/L$. Of course, different choices of vectors may give the same (geometric) lattice, and hence the same factor torus. For example, the following three pairs of vectors all generate the integer lattice in $\mathbb{R}^2$, and hence give the usual torus $\mathbb{R}^2/\mathbb{Z}^2$:

$$\left\{\begin{pmatrix}1\\0\end{pmatrix}, \begin{pmatrix}0\\1\end{pmatrix}\right\}, \quad \left\{\begin{pmatrix}1\\0\end{pmatrix}, \begin{pmatrix}-1\\1\end{pmatrix}\right\}, \quad \left\{\begin{pmatrix}1\\1\end{pmatrix}, \begin{pmatrix}1\\2\end{pmatrix}\right\}. \tag{31.9}$$

Similarly, the lattice whose points are the vertices of a tiling of the plane by equilateral triangles with side length 2 is generated by any of the following pairs of vectors:

$$\left\{\begin{pmatrix}2\\0\end{pmatrix}, \begin{pmatrix}1\\\sqrt{3}\end{pmatrix}\right\}, \quad \left\{\begin{pmatrix}2\\0\end{pmatrix}, \begin{pmatrix}-1\\\sqrt{3}\end{pmatrix}\right\}, \quad \left\{\begin{pmatrix}3\\\sqrt{3}\end{pmatrix}, \begin{pmatrix}4\\2\sqrt{3}\end{pmatrix}\right\}. \tag{31.10}$$

The astute reader will observe that the second and third pairs of vectors in (31.9) can be obtains from the first pair via left multiplication by the matrices $\left(\begin{smallmatrix} 1 & -1 \\ 0 & 1 \end{smallmatrix}\right)$ and $\left(\begin{smallmatrix} 1 & 1 \\ 1 & 2 \end{smallmatrix}\right)$, respectively. The even more astute reader will observe that the same thing is true in (31.10).

Indeed, this is quite a general fact. Given a lattice $L = L(\mathbf{v}_1, \dots, \mathbf{v}_n) \subset \mathbb{R}^n$, the parallelepiped spanned by $\mathbf{v}_1, \dots, \mathbf{v}_n$ is a fundamental domain for the action of $\mathbb{Z}^n$ on $\mathbb{R}^n$. If the n-dimensional volume of this fundamental domain is 1, we say that L is *unimodular*. Because the determinant of a matrix is the volume of the parallelepiped spanned by its row (or column) vectors, the row vectors of any matrix $X \in SL(n, \mathbb{R})$ generate a unimodular lattice. Furthermore, two bases $\{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ and $\{\mathbf{w}_1, \dots, \mathbf{w}_n\}$ generate the same (unimodular) lattice if and only if there exists a matrix $A \in SL(n, \mathbb{Z})$ such that $\mathbf{v}_i A = \mathbf{w}_i$ for all i, and so two matrices $X, Y \in SL(n, \mathbb{R})$ generate the same lattice if and only if $Y = XA$ for some $A \in SL(n, \mathbb{Z})$—that is, if and only if X and Y represent the same coset in $SL(n, \mathbb{R})/SL(n, \mathbb{Z})$. The upshot of all this is that the homogeneous space $SL(n, \mathbb{R})/SL(n, \mathbb{Z})$ can be equated to the space of all unimodular lattices in $\mathbb{R}^n$.

For $n = 2$ this homogeneous space is closely related to the modular surface $\mathbb{H}^2/PSL(2, \mathbb{Z})$, that can be identified with the double coset space $SO(2)\{\}SL(2, \mathbb{R})/SL(2, \mathbb{Z})$ and like that surface, it is non-compact. This non-compactness has important implications in number theory, but we do not consider them here. For larger values of n it is still non-compact and the role of the hyperbolic plane $\mathbb{H}^2$ is played by the *symmetric space* $SO(n)\{\}SL(n\mathbb{R})$ a fascinating geometric object with an even more fascinating boundary that is a far cry from the plain round sphere of the appropriate dimension. The role of the modular surface is then played by the double coset space $SO(n)\{\}SL(n, \mathbb{R})/SL(n, \mathbb{Z})$, non-compact again in a much more interesting way that the plain cusp of the modular surface.

c. The structure of $SL(n, \mathbb{Z})$. In our investigation of the algebraic structure of $SL(2, \mathbb{Z})$, we saw that it is finitely generated, finitely presented, and commensurable to a free group. It is natural to ask if the same properties hold for $SL(n, \mathbb{Z})$ with $n \geq 3$; it turns out that the first two do, but the third does not. In fact, we will see along the way that $SL(n, \mathbb{Z})$ with $n \geq 3$ is a representative of a totally new species of finitely generated groups: "large" in the sense of having exponential growth and many free subgroups but much more "rigid" or structured" that free groups or free products in the sense of having many "small" infinite subgroups such as abelian subgroups or rank ≥ 2 and nilpotent groups.

First we examine how far $SL(n, \mathbb{Z})$ is from being free. Observe that if a group G is commensurable to a free group (that is, if G is *virtually free*), then so is every subgroup of G. In particular, G cannot have any solvable subgroups that are not virtually cyclic.

c.1. *Nilpotent subgroups.* However, such subgroups exist in $SL(3,\mathbb{Z})$, and hence in $SL(n,\mathbb{Z})$ for any $n \geq 3$. Indeed, the subgroup $\mathcal{U}_n(\mathbb{Z})$ of upper-triangular unipotent $n \times n$ matrices with integer entries is nilpotent, and hence solvable, but is not virtually cyclic, and it follows that $SL(n,\mathbb{Z})$ does not have a finite index free subgroup for any $n \geq 3$.

Notice that this is not just an isolated freak, such as a center or a commutator. The subgroup $\mathcal{U}_n(\mathbb{Z})$ is very far from being normal in $SL(n,\mathbb{Z})$; it has lots of conjugates and in fact its conjugates generate the whole group $SL(n,\mathbb{Z})$

c.2. *Abelian unipotent subgroups.* Actually, we can do even better than nilpotent. $SL(n,\mathbb{Z})$ contains an abelian subgroup of rank $\frac{n^2}{4}$ if n is even, and rank $\frac{n^2-1}{4}$ if n is odd. To see this, consider the subgroup

$$G = \{I + A \mid A_{ij} = 0 \text{ whenever } i \leq n/2 \text{ and } j > n/2\}, \tag{31.11}$$

which contains matrices whose only non-zero off-diagonal entries occur in the upper right quadrant of the n^2 total entries.

To see that introduce the *elementary matrices* $E_{ij} = \mathrm{Id} + e_{ij}$, $i \neq j$ where e_{ij} if the matrix with entry 1 in the intersection of the ith row and jth column, and zeroes elsewhere. We encountered elementary matrices in Lecture 18.a where we proved in particular that E_{ij} and E_{kl} commute unless $k = j$ or $i = l$. Since elementary matrices have infinite order and G is generated by E_{ij}, $i \leq n/2$, $j > n/2$ we see that G is isomorphic to $\mathbb{Z}^{n^2/4}$ for n even and to $\mathbb{Z}^{(n^2-1)/4}$ for n odd. We will extensively use elementary matrices later.

c.3. *Semisimple(diagonalizable) abelian subgroups.* The last construction relies on the presence of Jordan blocks: one may ask if such an example can be found where the matrices in question are diagonalisable. That is, does there exist a matrix $C \in SL(n,\mathbb{R})$ such that $C\mathcal{D}_nC^{-1} \cap SL(n,\mathbb{Z})$ is an abelian subgroup of rank greater than 1? Observe that

$$\mathcal{D}_n = \{\mathrm{diag}(e^{t_1}, \dots, e^{t_n}\} \mid \sum t_i = 0\}$$

is isomorphic to $\mathbb{R}^{n-1}$, and so such a subgroup can have rank at most $n-1$. To answer this question, we need the following definition.

Definition 31.2. A matrix $A \in SL(n,\mathbb{Z})$ is *irreducible over* $\mathbb{Q}$ if its characteristic polynomial $p(\lambda) = \det(A - \lambda I)$ does not factor as the product of two polynomails with rational coefficients and smaller degree.

Remark. Irreducibility over $\mathbb{Q}$ implies that all eigenvalues are different, and hence the matrix is diagonalisable (over $\mathbb{C}$). Geometrically, irreducibility is the statement that there does not exist an invariant rational subspace (one that is defined by a linear equation with rational coefficients). Thus matrices in $\mathcal{U}_n$ are in some sense the *opposite* of irreducible, since they have invariant subspaces of every dimension.

The following is a particular case of the famous 1846 Dirichlet Unit Theorem that plays a central role in the the theory of algebraic number fields:

THEOREM 31.3. *Let $A \in SL(n, \mathbb{Z})$ be irreducible over $\mathbb{Q}$, and suppose all the eigenvalues of A are real. Then there exists a subgroup $S \subset SL(n, \mathbb{Z})$ such that $A \in S$ and S is abelian with rank $n - 1$.*

We omit the proof of this result, and instead turn our attention to the first question we asked at the beginning of this section: Is $SL(n, \mathbb{Z})$ (or equivalently, $GL(n, \mathbb{Z})$) finitely generated?

Lecture 32. Wednesday, November 18

a. Generators and generating relations for $SL(n, \mathbb{Z})$. Fix $n \in \mathbb{N}$, and recall that e_{ij} is the $n \times n$ matrix whose (i, j)-th entry is 1, and which has all other entries equal to 0. For all $i \neq j$, write $E_{ij} = I + e_{ij}$, and recall from Lecture 18.a that we proved that

$$[E_{ij}, E_{k\ell}] = \begin{cases} E_{i\ell} & j = k, i \neq \ell, \\ -E_{kj} & j \neq k, i = \ell, \\ I & j \neq k, i \neq \ell \end{cases} \tag{32.1}$$

PROPOSITION 32.1. *$SL(n, \mathbb{Z})$ is generated by the matrices E_{ij}, $1 \leq i, j \leq n$, $i \neq j$..*

PROOF. First observe that for $n = 2$, we have

$$E_{12} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \quad E_{21} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix},$$

and that

$$E_{21} E_{12}^{-1} E_{21} = B = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}. \tag{32.2}$$

Thus we have in particular

$$(E_{21} E_{12}^{-1} E_{21})^4 = \mathrm{Id}\,. \tag{32.3}$$

By Proposition 31.1, the matrix B, together with E_{12}, generates $SL(2, \mathbb{Z})$: it follows that $\langle E_{12}, E_{21} \rangle = SL(2, \mathbb{Z})$.

We now prove that the matrices E_{ij} generate $SL(n, \mathbb{Z})$. By induction, we may assume that the elementary $(n-1) \times (n-1)$ matrices E_{ij} generate $SL(n-1, \mathbb{Z})$. Our goal is to reduce to this case by proceeding along the same lines as the proof of Proposition 31.1.

Begin by letting $P(X) = |x_{11}| + |x_{21}| + \cdots + |x_{n1}|$ be the sum of the absolute values of the entries in the first column of $X \in SL(n, \mathbb{Z})$. Observe that if all but one of these entries vanish, then the remaining entry divides the determinant of X, and hence must be ± 1; in this case $P(X) = 1$. Thus for $P(X) > 1$, there exist $1 \leq i, j \leq n$, $i \neq j$, such that x_{i1} and x_{j1} are both non-zero. Let $P_0 = P(X) - |x_i| - |x_j|$ be the sum of the absolute values of the entries in the first column of X apart from x_i and x_j. Then a simple matrix multiplication shows that

$$\begin{aligned} P(E_{ij} X) &= P_0 + |x_i + x_j| + |x_j|\,, \\ P(E_{ij}^{-1} X) &= P_0 + |x_i - x_j| + |x_j|\,, \\ P(E_{ji} X) &= P_0 + |x_i| + |x_j + x_i|\,, \\ P(E_{ji}^{-1} X) &= P_0 + |x_i| + |x_j - x_i|\,, \end{aligned}$$

Now there are four possibilities:

(1) If $|x_i| \geq |x_j|$ and $x_i x_j < 0$, then $P(E_{ij}X) < P(X)$.
(2) If $|x_i| \geq |x_j|$ and $x_i x_j > 0$, then $P(E_{ij}^{-1}X) < P(X)$.
(3) If $|x_i| \leq |x_j|$ and $x_i x_j < 0$, then $P(E_{ji}X) < P(X)$.
(4) If $|x_i| \leq |x_j|$ and $x_i x_j > 0$, then $P(E_{ji}^{-1}X) < P(X)$.

It follows that there exists a matrix $C \in \langle \{E_{ij}\} \rangle$ such that $P(CX) = 1$. Furthermore, by the above remarks, the first column of CX is the vector $\pm \mathbf{e}_i$ for some i.

Now we observe that for every $i \neq j$, a computation analogous to (32.2) shows that the matrix $R_{ij} = E_{ij}E_{ji}^{-1}E_{ij}$ has the following action:

$$R_{ij}\mathbf{e}_i = -\mathbf{e}_j, \quad R_{ij}\mathbf{e}_j = \mathbf{e}_i, \quad R_{ij}\mathbf{e}_k = \mathbf{e}_k \text{ for all } k \neq i, j.$$

(Similar relations hold if R_{ij} acts from the right.)

In particular, choosing $R = R_{1i}$ if $(CX)_{i1} = 1$, and $R = R_{i1}$ if $(CX)_{i1} = -1$, we see that

$$RCX = \begin{pmatrix} 1 & \mathbf{b} \\ \mathbf{0} & X' \end{pmatrix}, \tag{32.4}$$

where $\mathbf{b} \in \mathbb{R}^{n-1}$ is a row vector, $\mathbf{0}$ is the $(n-1) \times 1$ zero vector, and $X' \in SL(n-1, \mathbb{R})$.

Now observe that if we carry out the procedure described above but let the elementary matrices act from the right instead of the left, we can produce a matrix $C' \in \langle \{E_{ij}\} \rangle$ such that

$$RCXC' = \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & X' \end{pmatrix}. \tag{32.5}$$

Finally, by the induction hypothesis $X' \in \langle \{E_{ij}\} \rangle$, and so $RCXC' \in \langle \{E_{ij}\} \rangle$ as well. It follows that $X = C^{-1}R^{-1}(RCXC')(C')^{-1}$ is a product of elementary matrices E_{ij}. □

Relations (32.1) (called *commutator relations*) and the additional relation (32.3) that does not follow from the commutator relations, form a system of generating relations in the group $SL(n, \mathbb{Z})$.

Elementary matrices form a very natural and convenient systems of generators for $SL(n, \mathbb{Z})$, however their number grows with n. One can easily construct the system of three generators using the following observations:

(1) All elementary matrices are conjugate via permutations of coordinates;
(2) permutations are realized by matrices in $GL(n, Z)$; even permutations by matrices in $SL(n, \mathbb{Z})$;
(3) permutation group S_n is generated by a cyclic permutation and a single transposition;
(4) for a matrix of an odd permutation changing the sign of one of one non -zero entry brings it into $SL(n, \mathbb{Z})$.

The generators then are: The generators A and B of $SL(2, \mathbb{Z})$ extended by ones on the rest of the main diagonal (the first one is simply E_{12}), and either the matrix of a cyclic permutation of coordinates (for odd n or such a matrix with one of the entries changed to -1 (for odd n).

b. Semi-direct products. In Lecture 30, we observed that $SL(n,\mathbb{Z})$ appears as the automorphism group $\mathrm{Aut}(\mathbb{Z}^n)$, illustrating that we can find new and interesting groups by consider automorphisms of already familiar examples. We now introduce another use of automorphisms to construct new groups from old.

As motivation, consider the group $\mathrm{Isom}^+(\mathbb{R}^n)$ of even isometries of Euclidean space $\mathbb{R}^n$. There are two naturally occurring subgroups of $\mathrm{Isom}^+(\mathbb{R}^n)$: first, the group $SO(n)$, which comprises all isometries fixing the origin,and second, the translation group $\mathbb{R}^n$, which simply comprises all translations. The following features of this example are of particular importance:

(1) The two subgroups $SO(n)$ and $\mathbb{R}^n$ generate the group $\mathrm{Isom}^+(\mathbb{R}^n)$: every even isometry of $\mathbb{R}^n$ can be written in the form $S\colon \mathbf{x} \mapsto A\mathbf{x} + \mathbf{v} = T \circ R(\mathbf{x})$, where $R\colon \mathbf{x} \mapsto A\mathbf{x}$ is in $SO(n)$ and $T\colon \mathbf{x} \mapsto \mathbf{x} + \mathbf{v}$ is in $\mathbb{R}^n$.
(2) This decomposition is unique: $\mathbf{v}$ is uniquely determined by the fact that $\mathbf{v} = S\mathbf{0}$, and A is uniquely determined by the fact that $A\mathbf{x} = S(\mathbf{x} - S\mathbf{0})$.
(3) The translation subgroup $\mathbb{R}^n$ is normal, and hence "canonical" in some sense: there is only one subgroup of $\mathrm{Isom}^+(\mathbb{R}^n)$ that is isomorphic to $\mathbb{R}^n$. Since it is normal, we can take the quotient group, and we find that $\mathrm{Isom}^+(\mathbb{R}^n)/\mathbb{R}^n$ is isomorphic to $SO(n)$.
(4) The subgroup $SO(n)$ is not normal in $\mathrm{Isom}^+(\mathbb{R}^n)$, and is isomorphic to all of its conjugates. Indeed, given an even isometry $S\colon \mathbb{R}^n \to \mathbb{R}^n$ and an isometry $R \in SO(n)$, we see that $S \circ R \circ S^{-1}$ is an even isometry of $\mathbb{R}^n$ that fixes $\mathbf{p} = S\mathbf{0}$. Thus $G_{\mathbf{p}} = S \circ SO(n) \circ S^{-1}$ is the group of all even isometries that fix $\mathbf{p}$. $G_{\mathbf{p}}$ is isomorphic to $SO(n)$, and thus we see that there is one isomorphic copy of $SO(n)$ for every $\mathbf{p} \in \mathbb{R}^n$.

The first two properties above are reminiscent of the direct product construction, where we construct a group G as $H \times K$, and observe that isomorphic copies of H and K sit inside the direct product G as normal subgroups. The third property is also reminiscent of that situation, but the fourth is new.

The difference between the direct product case and our current situation can also be seen as follows: thanks to the first two properties, there is a one-to-one correspondence between elements of G (or of $\mathrm{Isom}^+(\mathbb{R}^n)$) and ordered pairs $h \star k$ (or $R \star T$), where we use this notation rather than (h,k) to emphasise the relationship with the free product. To obtain the multiplication rule for elements of the product, we need a way of going from the concatenation $h_1 \star k_1 \star h_2 \star k_2$ to something of the form $h \star k$. That is, we need to equate the expression $k_1 \star h_2$ with a pair $h \star k$ (it suffices to do this since we already know how to multiply elements within H and K).

For a direct product $H \times K$, this is given by declaring that elements of H and K commute, and so $k{\star}h = h{\star}k$. For the product in the example, we have a different rule, which is specified for us by function composition. Taking

$H = \mathbb{R}^n$ and $K = SO(n)$, we have the following rule for $S_1 = T_{\mathbf{v}_1} \circ R_{A_1}$ and $S_2 = T_{\mathbf{v}_2} \circ R_{A_2}$:

$$(S_1 \circ S_2)(\mathbf{x}) = S_1(A_2\mathbf{x} + \mathbf{v}_2) = A_1A_2\mathbf{x} + A_1\mathbf{v}_2 + \mathbf{v}_1. \tag{32.6}$$

Observe that the map $\psi_A \colon \mathbf{v} \mapsto A\mathbf{v}$ is an automorphism of $\mathbb{R}^n$. Writing $k_i = A_i$ and $h_i = \mathbf{v}_i$, we may write (32.6) as

$$h_1 \star k_1 \star h_2 \star k_2 = (h_1 \star \psi_{k_1}(h_2)) \star (k_1 \star k_2), \tag{32.7}$$

where $\star$ denotes both the binary operation of H and the binary operation of K, and $\psi_{k_1} \colon H \to H$ is the automorphism described above.

This procedure is very general. Let G be any group, and suppose H, K are subgroups of G with the following properties:

(1) H is normal.
(2) $G = HK$. That is, for every $g \in G$ there exist $h \in H$ and $k \in K$ such that $g = hk$.
(3) $H \cap K = \{e\}$, and consequently the decomposition in (2) is unique.

Then we see that the binary operation in G can be (almost) reconstructed from the binary operations in H and K as follows:

$$g_1g_2 = h_1k_1h_2k_2 = (h_1k_1h_2k_1^{-1})(k_1k_2). \tag{32.8}$$

Because H is normal, the product khk^{-1} is in H for all $h \in H$, $k \in K$, and so $\psi_k \colon h \mapsto khk^{-1}$ defines an automorphism of H (which comes from an inner automorphism of G). This automorphism (or rather, collection of automorphisms) is all the extra information needed to reconstruct G from its subgroups H and K, via (32.7). We say that G is a *semi-direct product* of H and K, and write $G = H \ltimes K$. (Note that this is not the same thing as $K \ltimes H$.)

This description of semi-direct products assumes that we already know the group G, and are interested in decomposing it into a product of two subgroups, one of which is normal. This is an *internal* semi-direct product, since everything happens within a setting that is already known. One may also define an *external* semi-direct product: given two groups H and K and a homomorphism $\psi \colon k \mapsto \psi_k \in \mathrm{Aut}(H)$, the semi-direct product of H and K with respect to the family of automorphisms ψ_k is the group whose elements are ordered pairs (h, k), and whose binary operation is given by

$$(h_1, k_1) \star (h_2, k_2) = (h_1\psi_{k_1}(h_2), k_1k_2). \tag{32.9}$$

Exercise 32.1. Show that H is isomorphic to a normal subgroup of $H \ltimes K$. Show that the isomorphic image $\{e_H\} \times K \subset H \ltimes K$ is not normal unless ψ_k is the identity automorphism for all k.

Remark. We have now seen three sorts of products of groups: direct, semi-direct, and free. These three constructions are related. To begin with, a direct product is a special case of a semi-direct product, where each automorphism ψ_k is taken to be the identity automorphism. Furthermore, both direct and semi-direct products can be obtained as factor groups of the free product, as follows:

$$\begin{aligned} H \times K &= H * K/\langle \{[h,k]\}\rangle, \\ H \ltimes K &= H * K/\langle \{\psi_k(h)kh^{-1}k^{-1}\}\rangle. \end{aligned}$$

The generating relations for $H \times K$ ensure that $hk = kh$, while the generating relations for $H \ltimes K$ ensure that $kh = \psi_k(h)k$.

c. Examples and properties of semi-direct products. We are now in a position to construct many of our familiar examples using very simple building blocks. For example, returning to the example of $\mathrm{Isom}^+(\mathbb{R}^n)$, we see that

$$\mathrm{Isom}^+(\mathbb{R}^n) = \mathbb{R}^n \ltimes SO(n). \tag{32.10}$$

In particular, because $SO(2)$ can be identified with $S^1 = \{z \in \mathbb{C} \mid |z| = 1\}$, we have

$$\mathrm{Isom}^+(\mathbb{R}^2) = \mathbb{R}^2 \ltimes S^1. \tag{32.11}$$

If we broaden our horizons to the group $\mathrm{Sim}^+(\mathbb{R}^2)$ of orientation-preserving similarity transformations—that is, maps of the form $\mathbf{x} \mapsto \rho A\mathbf{x} + \mathbf{v}$, where $A \in SO(2)$ and $\rho > 0$—then we may observe that every such transformation can be written as an affine transformation of $\mathbb{C}$, as $z \mapsto w_1 z + w_2$, where $w_1 \in \mathbb{C}^* = \mathbb{C} \setminus \{0\}$ and $w_2 \in \mathbb{C}$. Thus

$$\mathrm{Sim}^+(\mathbb{R}^2) = \mathrm{Aff}^+(\mathbb{C}) = \mathbb{R}^2 \ltimes \mathbb{C}^* = \mathbb{R}^2 \ltimes (S^1 \times \mathbb{R}^+). \tag{32.12}$$

Example 32.2. A more intricate example occurs if we consider the subgroup G of $Aff^+(\mathbb{R})$ generated by $E\colon x \mapsto 2x$ and $T\colon x \mapsto x + 1$. This group consists of all transformations of the form

$$x \mapsto 2^n x + p2^k x, \ \ n, k, p \in \mathbb{Z}. \tag{32.13}$$

Its translation group is not finitely generated. In fact, this translation group is isomorphic to the additive group $\mathbb{Z}(\frac{1}{2}) = \{k/2^n \mid k \in \mathbb{Z}, n \in \mathbb{N}\}$ of dyadic rationals. On the one hand, any two of its elements have a common multiple, hence it does not contain a copy of $\mathbb{Z}^2$.On the other hand it is torsion free, and if it were finitely generated, then by Theorem29.3 it would have to be isomorphic to $\mathbb{Z}$ that is obviously not the case. The factor group $G/\mathbb{Z}(\frac{1}{2})$ is isomorphic to $\mathbb{Z}$ Let $\psi : \mathbb{Z}^n \to \mathbb{Z}(\frac{1}{2})$ be defined by $\psi(n) = 2^n$. Identify the element (32.13) with the pair $(p2^k, 2^n)$. Then multiplication in G is given by (32.9) hence

$$G = \mathbb{Z}\left(\frac{1}{2}\right) \ltimes \mathbb{Z}.$$

An important property of the semi-direct product is its behaviour with respect to solvability.

Proposition 32.3. *The semi-direct product $H \ltimes K$ where H and K are solvable groups, is solvable.*

PROOF. It follows for (32.9) that $[G,G] \subset H \ltimes [K,K]$. By iterating the commutator we get down to $[H,H]$ and hence eventually to the identity. $\square$

The fact that semi-direct products of solvable groups are themselves solvable allows us to build many solvable groups from familiar building blocks. We can use elements and, more generally subgroups of $GL(n,\mathbb{Z})$ as automorphisms of $\mathbb{Z}^n$ to construct interesting finitely generated nilpotent and solvable groups all of whose subgroups are finitely generated.

In general even if both groups H and K are abelian the semi-direct product $H \ltimes K$ is not nilpotent as Example 32.2 or the following example show.

EXAMPLE 32.4. Let $\psi(n) = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}^n$ and let G be the corresponding semi-direct product $\mathbb{Z}^2 \ltimes \mathbb{Z}$. One sees form the multiplication formula that $[G,G] = \mathbb{Z}^2$ and hence $[[G,G],G] = [G,G]$.

However sometimes one gets more lucky.

EXERCISE 32.2. Consider a homomorphism $\psi : \mathbb{Z}^n \to \Gamma_N$, the group of upper-diagonal $N \times N$ unipotent matrices with integer entries. Prove that the semi-direct product $\mathbb{Z}^N \ltimes \mathbb{Z}^n$ with multiplication rule (32.9) is a finitely generated nilpotent group.

We finish our discussion of semi-direct products with the following observation. Given a group G and a normal subgroup $H \subset G$, one can consider the factor group $K = G/H$ and try to construct G as a semi-direct product of H and K. After all, this is more or less exactly what we did in the case $G = \text{Isom}^+(\mathbb{R}^n)$, $H = \mathbb{R}^n$. However, that example had one further key property: the factor group $G/H = SO(n)$ occurs *as a subgroup of G that sits "orthogonally" to H*. We say that the group G *splits* over H. This is not always the case, and when it fails, it is impossible to write $G = H \ltimes G/H$.

EXERCISE 32.3. Let Γ_3 be the group of upper-triangular unipotent matrices with integer entries, and let $H = Z(\Gamma_3)$. Show that $H = \mathbb{Z}$ and $\Gamma_3/H = \mathbb{Z}^2$, and show that there is no subgroup $K \subset \Gamma_3$ such that

(1) K is isomorphic to $\mathbb{Z}^2$.
(2) $HK = \Gamma_3$.
(3) $H \cap K = \{I\}$.

Finally, show that Γ_3 cannot be written as a semi-direct product of $\mathbb{Z}$ and $\mathbb{Z}^2$.

Lecture 33. Friday, November 20

a. Quasi-isometries. Having already seen a number of connections between groups and geometry, we now look at groups *themselves* as geometric objects. We will be interested not in local geometric properties—those having to do with small scales—but rather with global properties, the *coarse geometry* associated to a group. These methods have their origins in the pioneering work of Mikhail Gromov, which uses soft arguments to derive powerful results about both the algebraic structure and the large-scale structure of broad classes of groups.

We begin with some purely metric definitions. Given metric spaces (X, d) and (X', d'), we want to weaken the notion of an isometry between X and X' to allow maps that may lose the small-scale structure, but still capture large-scale structure. For example, one could consider bi-Lipschitz maps—that is, maps $f\colon X \to X'$ for which there exists a constant $L > 0$ such that

$$\frac{1}{L}d(x,y) \leq d'(f(x), f(y)) \leq Ld(x,y) \tag{33.1}$$

for every $x, y \in X$. However, this definition is still too restrictive for our purposes: in particular, every invertible bi-Lipschitz map is a homeomorphism and therefore relates the local topological structure of X and X'. We are interested in maps that allow us to disregard this structure: the appropriate definition is as follows.

Definition 33.1. A map $f\colon X \to X'$ is a *quasi-isometric embedding* if there exist constants $A, B > 0$ such that

$$d'(f(x), f(y)) \leq Ad(x,y) + B \tag{33.2}$$

for all $x, y \in X$. The metric spaces (X, d) and (X', d') are *quasi-isometric* if there exist quasi-isometric embeddings $f\colon X \to X'$ and $g\colon X' \to X$ and a constant $C > 0$ such that

$$\begin{aligned} d(g \circ f(x), x) &\leq C, \\ d'(f \circ g(x'), x') &\leq C \end{aligned} \tag{33.3}$$

for all $x \in X$, $x' \in X$.

Remark. If we require $B = C = 0$, then this definition reduces to the definition of an invertible bi-Lipschitz map. By allowing B and C to be positive (indeed, arbitrarily large), we can ignore all finite-scale properties of the space and focus on the global (coarse) properties.

Example 33.2. Observe that the natural embedding $f\colon \mathbb{Z} \to \mathbb{R}$ given by $f(n) = n$ is a quasi-isometric embedding with $A = 1$, $B = 0$. The nature of the definition is shown more clearly by the map

$$\begin{aligned} g\colon \mathbb{R} &\to \mathbb{Z} \\ x &\mapsto \lfloor x \rfloor, \end{aligned}$$

which is a quasi-isometric embedding with $A = 1$, $B = 1$. (In particular, quasi-isometric embeddings need not be continuous!) Furthermore, $g \circ f \colon \mathbb{Z} \to \mathbb{Z}$ is the identity map, and $f \circ g \colon \mathbb{R} \to \mathbb{R}$ is given by $f \circ g(x) = \lfloor x \rfloor$, and hence satisfies (33.3) with $C = 1$. Thus $\mathbb{Z}$ and $\mathbb{R}$ are quasi-isometric.

Example 33.3. A similar argument shows that $\mathbb{Z}^n$ is quasi-isometric to $\mathbb{R}^n$ for any $n \in \mathbb{N}$. More generally, given a metric space (X, d), we say that a subset $\mathcal{N} \subset X$ is an *ε-net* for some $\varepsilon > 0$ if every $x \in X$ is within a distance ε of a point $n \in \mathcal{N}$. We may also suppress the precise value of ε and simply say that $\mathcal{N}$ is a net in X. Observe that $\mathbb{Z}^n$ is a net in $\mathbb{R}^n$ (here $\varepsilon > \sqrt{n}/2$).

Given an ε-net $\mathcal{N} \subset X$, let $f \colon \mathcal{N} \to X$ be the natural inclusion map, and let $g \colon X \to \mathcal{N}$ be the map that takes a point in X to the nearest point in $\mathcal{N}$. Then f and g are quasi-isometric embeddings and their compositions satisfy (33.3), so X and $\mathcal{N}$ are quasi-isometric.

Example 33.4. Let $G \subset PSL(2, \mathbb{R})$ be a cocompact Fuchsian group, and let $\mathrm{Orb}(z)$ be the orbit of a point $z \in \mathbb{H}^2$. Then $\mathrm{Orb}(z)$ is a net in $\mathbb{H}^2$, and hence $\mathrm{Orb}(z)$ and $\mathbb{H}^2$ are quasi-isometric.

Example 33.5. Suppose X has finite diameter—that is, there exists $C > 0$ such that $d(x, y) < C$ for all $x, y \in X$. Then X is quasi-isometric to a point—that is, to a metric space with a single element. In particular, every compact metric space is quasi-isometric to a point.

Proposition 33.6. *Quasi-isometry is an equivalence relation.*

Proof. Symmetry and reflexivity are immediate from the definition. For transitivity, we first observe that the composition of two quasi-isometric embeddings is again a quasi-isometric embedding. Indeed, if $f \colon X \to X'$ and $f' \colon X' \to X''$ satisfy (33.2) for constants $A_f, B_f, A_{f'}, B_{f'}$, then

$$d''(f' \circ f(x), f' \circ f(y)) \leq A_{f'} d'(f(x), f(y)) + B_{f'} \leq A_{f'} A_f d(x, y) + A_{f'} B_f + B_{f'},$$

and so $f' \circ f$ is a quasi-isometric embedding. Now if X' is quasi-isometric to both X and X'', then there exist quasi-isometric embeddings $f \colon X \to X'$, $g \colon X' \to X$, $f' \colon X' \to X''$, and $g' \colon X'' \to X'$ such that $f \circ g$, $g \circ f$, $f' \circ g'$, and $g' \circ f'$ all satisfy (33.3). The compositions $f' \circ f$ and $g \circ g'$ give quasi-isometric embedings between X and X'', and it remains only to check (33.3):

$$\begin{aligned} d(g \circ g' \circ f' \circ f(x), x) &\leq d(g \circ g' \circ f' \circ f(x), g \circ f(x)) + d(g \circ f(x), x) \\ &\leq A_g d(g' \circ f' \circ f(x), f(x)) + C_{g \circ f} \\ &\leq A_g C_{g' \circ f'} + C_{g \circ f}. \end{aligned}$$

□

b. Quasi-isometries and growth properties. Having seen several examples of metric spaces that *are* quasi-isometric, we now give some examples of spaces that are *not* quasi-isometric. In general, given some abstract equivalence relation (such as quasi-isometry) the best way to show that two objects are not equivalent is to give a property that is invariant under equivalence but differs for the two objects in question. This is the approach we take here, and the property in question is the growth rate.

To be precise, given a discrete metric space (X, d), a point $x \in X$, and a radius $r > 0$, we let $\mathcal{G}_X(x, r)$ be the number of points in X that lie within a distance r of x. Observe that if we consider integer lattices in $\mathbb{R}^n$, then $\mathcal{G}_{\mathbb{Z}^n}(\mathbf{x}, r) \approx Cr^n$ for some constant $C > 0$, while if we consider a cocompact Fuchsian group $G \subset PSL(2, \mathbb{R})$ and a corresponding net $\text{Orb}(z) \subset \mathbb{H}^2$, then $\mathcal{G}_{\text{Orb}(z)}(z, r) \approx e^{\lambda n}$ for some $\lambda > 0$.

Exercise 33.1. Show that if $\mathcal{G}_X(x, r)$ grows exponentially in r, then $\mathcal{G}_X(y, r)$ does as well for any $y \in X$. Similarly, show that if $\mathcal{G}_X(x, r)$ grows polynomially in r, then $\mathcal{G}_X(y, r)$ does as well for any $y \in X$, and the degree of polynomial growth is the same.

As a consequence of Exercise 33.1, we may speak without ambiguity of the growth rate of a metric space, since this growth rate does not depend on the choice of centre.

Exercise 33.2. Show that if X and X' are quasi-isometric and X has exponential growth, then X' does as well. Similarly, show that polynomial growth is a quasi-isometric invariant, as is the degree of polynomial growth.

It follows from Exercise 33.2 that $\mathbb{Z}^m$ and $\mathbb{Z}^n$ are not quasi-isometric for $m \neq n$, as they have different degrees of polynomial growth. Consequently, because quasi-isometry is an equivalence relation, we see that $\mathbb{R}^m$ and $\mathbb{R}^n$ are not quasi-isometric for $m \neq n$. Similarly, $\mathbb{H}^2$ has a net with exponential growth, and so is not quasi-isometric to $\mathbb{R}^n$ for any n.

Remark. We could attempt to define an analogue of these growth rates for non-discrete metric spaces such as $\mathbb{R}^n$ and $\mathbb{H}^2$ by using the volume; however, we run into the problem that quasi-isometries are not necessarily smooth or even continuous, and so it is not obvious how they transform volume. Thus the discrete case is the better playground for these techniques.

Remark. In the next section we will apply the notion of quasi-isometry to groups themselves. One can show (though we do not do so yet) that F_2 is *not* quasi-isometric to $\mathbb{H}^2$ (and hence to the surface groups SG_n), despite the fact that they have the same growth rate. Thus growth rate is not a *complete* invariant for quasi-isometry.

c. Geometric properties of groups. Let G be a finitely generated group, and let Γ be a generating set for G. Let d_Γ be the word metric on G—that is, $d_\Gamma(g, g')$ is the minimum length of a word w in the generators Γ such that $g' = gw$. Then the discussion of growth rates in the previous section reduces to our familiar notion of the growth rate of $\mathcal{G}_{G,\Gamma}(n)$.

Proposition 33.7. *(G, d_Γ) is quasi-isometric to its Cayley graph.*

Proof. Exactly as for $\mathbb{Z}$ and $\mathbb{R}$: one quasi-isometric embedding is the natural embedding of G into its Cayley graph, and the other direction is given by the map from a point on the Cayley graph to the nearest vertex. □

The metric space (G, d_Γ) is discrete, rendering it well-suited for the techniques of the previous section. Cayley graphs have an important advantage over this space. Namely, they are *geodesic spaces*—that is, they admit isometric embeddings of the interval [1] and that any two points are endpoints for such an embedding.

As always, we must ask what happens to the metric space (G, d_Γ) if we pass to a different generating set. It follows immediately from the estimates in the proofs of Propositions 28.11 and 29.1 that (G, d_Γ) and $(G, d_{\Gamma'})$ are quasi-isometric whenever Γ and Γ' are finite generating sets. However, the ballgame changes entirely if the generating set is allowed to be infinite, and so we steer clear of this case for the time being.

Example 33.8. Suppose G acts discretely on a metric space X, and suppose that there exists a fundamental domain in X of finite diameter. Then G is quasi-isometric to X.

Example 33.9. Commensurable groups are quasi-isometric: it suffices to show that if $H \subset G$ is a subgroup of finite index, then H and G are quasi-isometric. This is the same principle as the statement that a net is quasi-isometric to the space it is embedded in: the embedding map $G \to H$ is the natural one, and for the map $H \to G$ we may take the map that assigns to each element of G the nearest element of H. That this satisfies (33.2) and (33.3) follows from the fact that the coset representatives are at most some fixed finite distance from the identity.

Example 33.10. Any two free groups are commensurable, and hence quasi-isometric. Similarly, every surface group SG_n, $n \geq 2$, is quasi-isometric to $\mathbb{H}^2$, and hence these groups are all quasi-isometric.

Now we can come to the heart of the matter. We say that *geometric properties of groups* are precisely those properties that are invariant under quasi-isometry. For example, the rate of growth of a group is a geometric property: more precisely, the *presence* of exponential growth is a geometric property, but the *exponent* of exponential growth is not, while both the presence and the degree of polynomial growth are geometric properties.

Given a finitely generated group G, one can ask if it is finitely presented. There are many finitely generated groups that are *not* finitely presented. It turns out that the property of being finitely presented is also geometric, in that it is preserved by quasi-isometries.

We end this lecture with a brief discussion of geometrically different groups that are *not* distinguishable by growth rates alone. Many solvable groups with exponential growth; for example those from Examples 32.2 and 32.4 these are not quasi-isometric to either F_2 or $\mathbb{H}^2$ (which as we already said, are not quasi-isometric to each other). Similarly, $SL(n, \mathbb{Z})$ has exponential growth for $n \geq 2$ (it contains a copy of F_2) but is not quasi-isometric

[1] This is a purely metric definition of geodesic, with no assumption of smoothness on the embedding or the space.

to any of these examples. In order to distinguish these four types of groups with exponential growth one introduces geometric invariants of very different sorts that together provide the basic toolkit for geometric group theory.

(1) We mention briefly the case of solvable groups. One way of characterising a group or metric space with exponential growth is to observe that the boundary of a ball is "fat" compared to the interior of the ball. Consider a ball of radius r that contains roughly $e^{\lambda r}$ points of X (or elements of G). Fixing a constant $a > 0$, the ball of radius $r - a$ contains roughly $e^{\lambda(r-a)}$ points, and thus the proportion of points that are within a of the boundary is roughly

$$1 - \frac{e^{\lambda(r-a)}}{e^{\lambda r}} = 1 - e^{-\lambda a}.$$

This proportion stays constant as $r \to \infty$, in marked contrast with the polynomial case, where

$$1 - \frac{(r-a)^n}{r^n} \to 0.$$

Thus in some sense, spaces or groups with sub-exponential growth are characterised by the property that we can cover them with balls whose boundaries are "thin". One can generalise this notion to allow covers with other sets besides balls, and by doing so one obtains the definition of *amenable groups*. We do not prove anything about amenability here, or even give its precise definition, but we note the following points:

(a) Amenability is a geometric property.
(b) Solvable groups are amenable.
(c) Free groups are not amenable.

(2) To distinguish free groups and surface groups from $SL(n, \mathbb{Z})$, $n \geq 3$ one uses the notion of a *hyperbolic group*. Any geodesic triangles in such groups are very thin; they may only contain a ball of bounded radius. This is clearly a geometric property and one shows that the first two groups are hyperbolic while the third is not.

(3) For a hyperbolic group one introduces the notion of *group boundary*, a compact topological space such that quasi-isometric groups have homeomorphic boundaries. The boundary of a surface group is, not surprisingly, a circle as oneguesses form the quasi-isometry with $\mathbb{H}^2$, but the boundary of a free group is a Cantor set.

Within polynomial growth, we observe that Γ_3 and $\mathbb{Z}^4$ have the same degree of polynomial growth; nevertheless, one may show that they are not quasi-isometric.

Is there a geometric property that lets us distinguish these examples?