MTH 461: Survey of Modern Algebra

University of Miami

Chris Scaduto

Abstract

These are lecture notes for an introductory course on Modern Algebra, originally taught in the spring semester of 2021 at the University of Miami. The notes were given some minor updates after I taught the same course in the spring semesters of 2024 and 2026. Each lecture note is a near faithful recording of the content covered in a one 50 minute lecture of the first time the material was taught, and was originally typed during the week of the given lecture.

The course is designed for students that have a background in basic proofs and linear algebra, but may not go on to study more advanced mathematics that uses algebra. For this reason, the material, while somewhat systematic, attempts to maximize the number of interesting topics and applications from abstract algebra. Proofs are sometimes only sketched, or entirely omitted. There is no claim of originality and most if not all the material is standard and can be found elsewhere. A (possibly incomplete) reference list appears at the end of the document.

Contents

Definition of a Group . . . 3
Integers modulo n . . . 9
More modular arithmetic . . . 12
Orders of group elements . . . 16
Symmetries . . . 19
Symmetric groups . . . 22
Alternating groups . . . 24
Cosets and Lagrange's Theorem . . . 28
Consequences of Lagrange's Theorem . . . 30
Applications to cryptography . . . 33
Normal subgroups and quotient groups . . . 36
Homomorphisms and Isomorphisms . . . 39
Complex numbers . . . 42
More isomorphisms, and kernels . . . 45
First isomorphism theorem, and symmetries of a cube . . . 48
More isomorphism theorems . . . 52
Cayley's Theorem . . . 55
Classification of finite groups . . . 58
The Rubik's cube group . . . 61
Definition of a ring . . . 65
More ring basics . . . 68
Kernels, ideals and quotient rings . . . 71
Principal ideals . . . 74
Principal ideal domains . . . 76
Rings and geometry . . . 79
Prime and maximal ideals . . . 83
Euclidean domains and UFDs . . . 85
Primes as the sum of two squares . . . 89
Constructions of fields . . . 92
Algebraic field extensions . . . 96
Degrees of field extensions . . . 99
Straightedge and compass constructions . . . 103
Splitting fields and Galois groups . . . 108
The Galois Theorem . . . 112

Definition of a Group

A *group* $(G, \circ)$ is a set G together with a binary operation

$$\circ : G \times G \to G$$

written $(a, b) \mapsto a \circ b$, such that the following properties (axioms) hold:

1. (Associativity) For every $a, b, c \in G$ we have $a \circ (b \circ c) = (a \circ b) \circ c$.
2. (Identity) There is some $e \in G$ such that $a \circ e = e \circ a = a$ for all $a \in G$.
3. (Inverses) For every $a \in G$ there exists $a^{-1} \in G$ such that $a \circ a^{-1} = a^{-1} \circ a = e$.

Given a group $(G, \circ)$ we sometimes omit the notation of the operation $\circ$ and simply write G. However, keep in mind that the group operation is essential information.

In axiom (2), the element $e \in G$ is called an *identity element*. In axiom (3), the element $a^{-1} \in G$ is called an *inverse* of the element a.

Examples

1. The pair $(\mathbb{Z}, +)$, the integers with the operation of addition, is a group:

$$G = \mathbb{Z} = \{\ldots, -3, -2, -1, \quad 0, \quad 1, \quad 2, \quad 3, \ \ldots\}, \qquad \text{“}a \circ b\text{”} = a + b$$

Let us check the 3 axioms of a group. Associativity is the familiar identity $a + (b + c) = (a + b) + c$. An identity “$e$” is given by $0 \in \mathbb{Z}$, as $a + 0 = 0 + a = a$. Finally, given an integer $a \in \mathbb{Z}$, the negative $-a \in \mathbb{Z}$ serves as “a^{-1}” because $a + (-a) = (-a) + a = 0$.

In the same way, the rational numbers with addition $(\mathbb{Q}, +)$, and the real numbers with addition $(\mathbb{R}, +)$, are both examples of groups.

For each of these examples we sometimes just write $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$ to mean the above groups, where the operation is addition.

2. Consider $(\mathbb{Q}^\times, \times)$, the *non-zero* rational numbers $\mathbb{Q}^\times = \mathbb{Q} \setminus \{0\}$ with the operation of *multiplication*. This is a group. Associativity is the familiar identity $a \times (b \times c) = (a \times b) \times c$. An identity element is given by $1 \in \mathbb{Q}^\times$, as $a \times 1 = 1 \times a = a$ for any $a \in \mathbb{Q}^\times$. Finally, an inverse “a^{-1}” of $a \in \mathbb{Q}^\times$ is $1/a \in \mathbb{Q}$. Here we see why the element zero had to have been expelled, as it has no inverse with respect to multiplication. Similarly, the non-zero real numbers with multiplication $(\mathbb{R}^\times, \times)$ is a group.

3. Define $\mathrm{GL}_2(\mathbb{R})$ to be the set of 2×2 matrices with real entries and non-zero determinant:

$$\mathrm{GL}_2(\mathbb{R}) = \left\{ A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} : \quad a, b, c, d \in \mathbb{R}, \quad \det(A) = ad - bc \neq 0 \right\}$$

Then $\mathrm{GL}_2(\mathbb{R})$ together with the operation of matrix multiplication is a group. This example is called the *general linear group* of degree 2 over $\mathbb{R}$. An inverse of $A \in \mathrm{GL}_2(\mathbb{R})$ is given by

$$A^{-1} = \frac{1}{ad-bc}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

Similarly, we can define $\mathrm{GL}_2(\mathbb{Q})$ by requiring the entries a, b, c, d to be rational.

4. Let $G = \{e\}$ and defined $e \circ e = e$. This is a group, called the *trivial* group.

5. The subset $\{1, -1\} \subset \mathbb{Z}$ with the operation of multiplication is a group.

Some terminology

A group $(G, \circ)$ is *abelian* (or *commutative*) if for all $a, b \in G$ we have $a \circ b = b \circ a$.

The groups $(\mathbb{Z}, +)$, $(\mathbb{Q}, +)$, $(\mathbb{R}, +)$, $(\mathbb{Q}^\times, \times)$, $(\mathbb{R}^\times, \times)$, $(\{1, -1\}, \times)$ and the trivial group are all abelian groups.

The general linear group $\mathrm{GL}_2(\mathbb{R})$ is *not* abelian. For example, consider

$$A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}, \qquad B = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$$

Then you can check that $A, B \in \mathrm{GL}_2(\mathbb{R})$ and $AB \neq BA$.

The *order* of a group G is equal to $|G|$, the cardinality (number of elements) of the set G. G has *infinite* order if $|G| = \infty$. If G is a finite set, i.e. $|G| < \infty$, then G is called a *finite group*.

All of the examples given above are groups of infinite order, except for the trivial group (order 1) and $(\{1, -1\}, \times)$ (order 2).

Cayley tables

All of the information of a group $(G, \circ)$ can be encoded in what's called the *Cayley table*. The Cayley table has columns and rows labelled by the elements of G. Given $a, b \in G$, the entry in row a and column b is the element $a \circ b$. For example, if $G = \{a, b, c\}$, it looks like:

	a	b	c
a	$a \circ a$	$a \circ b$	$a \circ c$
b	$b \circ a$	$b \circ b$	$b \circ c$
c	$c \circ a$	$c \circ b$	$c \circ c$

Here is the Cayley table for the finite group $(\{1, -1\}, \times)$ of order 2:

	1	-1
1	1	-1
-1	-1	1

If G is not a finite group, the Cayley table will of course be "infinite", having as many rows and columns as the elements of G.

Whether or not the group is abelian can be read off from the Cayley table: if G is abelian, the Cayley table is symmetric with respect to reflection across the diagonal line that passes through the entries $a \circ a$, $b \circ b$, ...

All of the groups we have seen so far are examples of structures you have encountered in various mathematical settings. Here is an example of an "abstract" group. Let the set G be

$$G = \{e, r, b, g, y, o\}$$

We define the group operation on this set by writing out the Cayley table:

	e	r	b	g	y	o
e	e	r	b	g	y	o
r	r	e	o	y	g	b
b	b	y	e	o	r	g
g	g	o	y	e	b	r
y	y	b	g	r	o	e
o	o	g	r	b	e	y

For example, the table says $r \circ b = o$ and $y \circ r = b$. You can check by brute force that the operation defined by the table satisfies the axioms of a group. The Cayley table is not symmetric with respect to the diagonal line, so this group is not abelian. Thus this group is a non-abelian group of order 6.

Basic properties of groups

In this lecture we discuss some basic properties of groups which follow directly from the definition. To begin, a few words on notation. Up to now we have considered a typical group $(G, \circ)$ with operation $a \circ b$. It is convenient to omit "$\circ$" from the notation and write

$$ab = a \circ b$$

Although this is a valid convention for any group, we will not always want to use it. For example, for the group $(\mathbb{Z}, +)$, writing "ab" for $a \circ b = a + b$ has the shortcoming of looking like integer multiplication. But for an arbitrary abstract group it is very convenient.

The associativity property of a group tells us that $(ab)c = a(bc)$. This continues on for more complicated operations. For example, we have

$$((ab)c)d = (a(bc))d = a((bc)d) = a(b(cd)) = (ab)(cd)$$

Each equality uses one use of the associativity axiom. What associativity is really telling us is that we can forget about those pesky parantheses: no matter where we put them, we get the same answer. The above group element can just be written $abcd$.

For what follows we let G be any group, with the conventions above.

▶ **The identity element in G is unique.**

Proof. Let $e, e' \in G$ be two identity elements. Because e is an identity element, $ee' = e$. Because e' is an identity element, $ee' = e'$. Together we get $e = e'$. □

▶ **The inverse of any element in G is unique.**

Proof. Let $a \in G$ be be any element. Let b and c be two inverses of a. (Let us avoid calling either one a^{-1} for now.) Because b is an inverse of a we have $ba = e$. Multiply both sides of this equation on the right by c to get $bac = c$. Because c is an inverse for a, we have $ac = e$. Thus $bac = c$ becomes $be = c$, and finally $b = c$. □

▶ **For every $a \in G$, we have $(a^{-1})^{-1} = a$.**

Proof. The element a satisfies $aa^{-1} = a^{-1}a = e$ and thus is an inverse of a^{-1}. It then makes sense to say $a = (a^{-1})^{-1}$ because inverses are unique. □

▶ **For all $a, b \in G$ we have $(ab)^{-1} = b^{-1}a^{-1}$.**

Proof. We only need check that $b^{-1}a^{-1}$ satisfies the property of being an inverse for ab. To this end: $(ab)(b^{-1}a^{-1}) = abb^{-1}a^{-1} = aea^{-1} = aa^{-1} = e$. Similarly $(b^{-1}a^{-1})(ab) = e$. □

This last property can be used any number of times to show the following relation:

$$(a_1 a_2 \cdots a_{n-1} a_n)^{-1} = a_n^{-1} a_{n-1}^{-1} \cdots a_2^{-1} a_1^{-1}$$

We introduce some more convenient notation. Suppose n is a positive integer. Then we define the symbol a^n to mean the element $aa \cdots a = a \circ a \circ \cdots \circ a$ formed by applying the group operation to n copies of the element a. If n is negative, define $a^n = a^{-1} \cdots a^{-1}$ for $-n$ copies of a^{-1}. If $n = 0$, define $a^0 = e$, the identity element. The following is straightforward to verify:

▶ **For all $a, b \in G$ and $n, m \in \mathbb{Z}$ we have the following properties:**

$$a^n a^m = a^{n+m}, \qquad (a^n)^m = a^{nm}, \qquad (ab)^n = (b^{-1}a^{-1})^{-n}$$

The above discussion shows that in an arbitrary group G, we have all the properties we are used to with, say, matrix multiplication of invertible matrices. Furthermore:

▶ **If G is abelian, then for all $a, b \in G$ and $n \in \mathbb{Z}$ we have $(ab)^n = a^n b^n$.**

To see this, write $(ab)^n = abab \cdots ab$ and use that $ab = ba$, since G is abelian, to move the terms past one another, yielding $a \cdots ab \cdots b = a^n b^n$. However, for a general group which is not necessarily abelian, just like for matrices, we do not always have $(ab)^n = a^n b^n$. To further illustrate this point:

▶ **If a group G has $(ab)^2 = a^2 b^2$ for all $a, b \in G$ then G is abelian.**

Proof. Suppose $(ab)^2 = a^2 b^2$ for all $a, b \in G$, i.e. $abab = aabb$. Multiply both sides of this equation by a^{-1} on the left and b^{-1} on the right to obtain $ba = ab$. Thus G is abelian. □

In the argument just made, we used the following cancellation property, which again follows by multiplying both sides of the equation on the left or right by the appropriate element:

▶ **Let $a, b, c \in G$. If $ab = ac$ then $b = c$. If $ba = ca$ then $b = c$.**

Let us illustrate how to solve equations in an abstract group. Suppose we are given

$$(xax)^2 = abx^2 \qquad x^2 a = (xa)^{-1}$$

where $a, b \in G$ are known and we would like to solve for $x \in G$. We do this as follows:

$$\begin{aligned}
(xax)^2 &= abx^2 \\
xaxxax &= abx^2 \\
xa(x^2 a)x &= abx^2 \\
xa(xa)^{-1}x &= abx^2 \\
x &= abx^2 \\
e &= abx \\
x &= (ab)^{-1} = b^{-1}a^{-1}
\end{aligned}$$

Subgroups

A subset $H \subset G$ of a group G is called a *subgroup* if the set H with the group operation restricted from G makes H a group. If we spell this out, we see that a subset $H \subset G$ is a subgroup if and only if the following properties hold:

1. The identity element e is in H.
2. For all $a, b \in H$ we have $ab \in H$.
3. For all $a \in H$ we have $a^{-1} \in H$.

You might like to verify that these properties imply H is a subgroup. The key point is that given these properties, the axioms of a group for H are inherited from those of G. A subgroup $H \subset G$ is *proper* if $H \neq G$. Another good exercise is to check:

▶ **The intersection of two subgroups $H, K \subset G$ is again a subgroup.**

Examples

1. The group $(\mathbb{Z}, +)$ is a subgroup of $(\mathbb{Q}, +)$ and $(\mathbb{R}, +)$, and $(\mathbb{Q}, +)$ is a subgroup of $(\mathbb{R}, +)$.

2. The group $(\mathbb{Q}^\times, \times)$ is a subgroup of $(\mathbb{R}^\times, \times)$. Note that $(\mathbb{Q}^\times, \times)$ is *not* a subgroup of $(\mathbb{Q}, +)$, even though $\mathbb{Q}^\times \subset \mathbb{Q}$, because the group operations are not the same.

3. Define $\mathrm{SL}_2(\mathbb{R})$ to be the set of 2×2 matrices with real entries and determinant 1:

$$\mathrm{SL}_2(\mathbb{R}) = \left\{ A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} : \quad a, b, c, d \in \mathbb{R}, \quad \det(A) = ad - bc = 1 \right\}$$

This is called the *special linear group* of degree 2 over $\mathbb{R}$. This is a subgroup of $\mathrm{GL}_2(\mathbb{R})$.

4. A non-trivial group G has at least two subgroups: G itself and $\{e\} \subset G$.

5. Consider $G = \{e, r, b, g, y, o\}$ of order 6 from Lecture 1. Then $\{e, r\}$, $\{e, b\}$, $\{e, g\}$ are subgroups of order 2, while $\{e, y, o\}$ is a subgroup of order 3. Here are their Cayley tables:

	e	r
e	e	r
r	r	e

	e	b
e	e	b
b	b	e

	e	g
e	e	g
g	g	e

	e	y	o
e	e	y	o
y	y	o	e
o	o	e	y

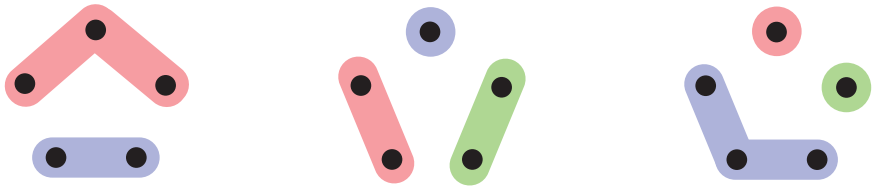
Integers modulo n

In this lecture we discuss some of the most important examples of groups: ones which come from taking certain equivalence classes of integers. To begin, we recall the notion of equivalence relations. Let S be a set, and $R \subset S \times S$ a subset. Write $a \sim b$ if and only if $(a, b) \in R$. Then R is an *equivalence relation* on the set S if the following hold:

1. (Reflexivity) $a \sim a$ for all $a \in S$.
2. (Symmetry) $a \sim b$ implies $b \sim a$.
3. (Transitivity) $a \sim b$ and $b \sim c$ implies $a \sim c$.

Given an equivalence relation on S as above, we write $[a] = \{b \in S : b \sim a\}$ for the *equivalence class of* a, which is a subset of S.

A *partition* of a set S is a collection of non-empty subsets $\{S_i\}_{i \in I}$ of S such that the union of all S_i over $i \in I$ is equal to S, and the subsets are pairwise disjoint: $S_i \cap S_j = \emptyset$ if $i \neq j$. For example, for a set with 5 elements represented by dots, here are depicted a few different partitions of S, where the subsets are encoded by colors:



Equivalence relations on sets and partitions of sets are essentially the same thing. Given an equivalence relation on S, the equivalence classes form a partition of S. Conversely, if we have a partition $\{S_i\}_{i \in I}$ of S, then the relation $a \sim b$ if and only if "a and b belong to some common subset S_i" defines an equivalence relation on S.

The group $\mathbb{Z}_n$

Fix a positive integer n. Define a relation on $\mathbb{Z}$ as follows: $a \sim b$ if and only if $a - b = nk$ for some $k \in \mathbb{Z}$. We check that this is an equivalence relation:

1. (Reflexivity) $a \sim a$ because $a - a = n0$.
2. (Symmetry) $a \sim b$ implies $a - b = nk$. Then $b - a = n(-k)$, implying $b \sim a$.
3. (Transitivity) $a \sim b$ and $b \sim c$ imply $a - b = nk$ and $b - c = nl$. Consequently we have $a - c = (a - b) + (b - c) = nk + nl = n(k + l)$. This implies $a \sim c$.

This equivalence relation partitions the set $\mathbb{Z}$ into n equivalence classes.

$$\mathbb{Z}_n = \{\text{equivalence classes of the relation } \sim\} = \{[0], [1], \ldots, [n-1]\}$$

For example, if $n = 3$, then $\mathbb{Z}_3$ consists of the equivalence classes $[0], [1], [2]$ where

$$[0] = \{0 + 3k : k \in \mathbb{Z}\}$$
$$[1] = \{1 + 3k : k \in \mathbb{Z}\}$$
$$[2] = \{2 + 3k : k \in \mathbb{Z}\}$$

and these partition the integers into 3 subsets. More generally, $[0], [1], \ldots, [n-1]$ are the equivalence classes of this relation. The set $\mathbb{Z}_n$ is called the *integers modulo n* or the *integers mod n*. Another notation for $a \sim b$ is: $a \equiv b \pmod{n}$. In summary we have:

$$[a] = [b] \qquad \Longleftrightarrow \qquad a - b = nk \text{ for some } k \in \mathbb{Z} \qquad \Longleftrightarrow \qquad a \equiv b \pmod{n}$$

Next, we define a binary operation "+" on the set $\mathbb{Z}_n$ as follows:

$$[a] + [b] = [a + b]$$

We first check this is well-defined. That is, suppose $[a'] = [a]$ and $[b'] = [b]$, i.e. $a' - a = nk$ and $b' - b = nl$. Then $(a' + b') - (a + b) = (a' - a) + (b' - b) = nk + nl = n(k + l)$. We conclude that $[a' + b'] = [a + b]$, and the operation is well-defined.

▶ **The set $\mathbb{Z}_n$ with the operation $+$ is an abelian group.**

To verify this we check the group axioms. First, we have associativity:

$$\begin{aligned}[a] + ([b] + [c]) = [a] + [b + c] &= [a + (b + c)] \\ &= [(a + b) + c] = [a + b] + [c] = ([a] + [b]) + [c].\end{aligned}$$

Note that associativity of $(\mathbb{Z}, +)$ was used from one line to the next. Next, $e = [0]$ is an identity because $[a] + [0] = [a + 0] = [a]$ and similarly $[0] + [a] = [a]$. An inverse for $[a] \in \mathbb{Z}_n$ is $[-a]$ because $[a] + [-a] = [a + (-a)] = [a - a] = [0]$ and similarly $[-a] + [a] = [0]$. Thus $(\mathbb{Z}_n, +)$ is a group. It is abelian because $[a] + [b] = [a + b] = [b + a] = [b] + [a]$.

The group $\mathbb{Z}_n$ is sometimes written $\mathbb{Z}/n$ or $\mathbb{Z}/n\mathbb{Z}$. When working in $\mathbb{Z}_n$ we often drop the brackets from the equivalence classes and write "a" instead of "$[a]$". The context should make it clear that "a" means the equivalence class of a mod n, and not the integer a. Using this convention, the following is the Cayley table for the group $\mathbb{Z}_6 = \{0, 1, 2, 3, 4, 5\}$:

	0	1	2	3	4	5
0	0	1	2	3	4	5
1	1	2	3	4	5	0
2	2	3	4	5	0	1
3	3	4	5	0	1	2
4	4	5	0	1	2	3
5	5	0	1	2	3	4

For example, in $\mathbb{Z}_6$ we have $1 + 1 = 2$, $3 + 3 = 0$ and $4 + 4 = 2$. We can also write these relations as $1 + 1 \equiv 2 \pmod{6}$, $3 + 3 \equiv 0 \pmod{6}$ and $4 + 4 \equiv 2 \pmod{6}$.

Cyclic groups

The group $(\mathbb{Z}_n, +)$ is a finite abelian group of order n. It is also very special because it is a cyclic group. An arbitrary group G is called *cyclic* if there is some $a \in G$ such that

$$G = \{a^k : k \in \mathbb{Z}\}.$$

The element a is called a *generator* of the group G. The group $(\mathbb{Z}, +)$ is cyclic with generator $1 \in \mathbb{Z}$, because any integer $a \in \mathbb{Z}$ can be written as $a = 1+\cdots+1$. For a similar reason:

▶ **The group $(\mathbb{Z}_n, +)$ is a cyclic group.**

To spell this out, take $a = [1]$. Then "a^k" in the group $(\mathbb{Z}_n, +)$ is none other than $[1]+\cdots+[1]$, where $[1]$ appears k times, which is equal to $[k]$. Now $\mathbb{Z}_n$ consists exactly of the classes $[k]$ as k runs over the integers; in fact, as we saw above, k need only run over $0, 1, \ldots, n-1$. Thus every element of $\mathbb{Z}_n$ is of the form "a^k" and so $\mathbb{Z}_n$ is cyclic with generator $[1]$.

In turns out the groups $(\mathbb{Z}, +)$ and $(\mathbb{Z}_n, +)$ for positive integers n are essentially the "only" cyclic groups, in a sense that we will make precise later.

More modular arithmetic

Last time, for any given positive integer n, we introduced the set $\mathbb{Z}_n$, which is obtained from $\mathbb{Z}$ by identifying integers which differ by a multiple of n. We defined a group operation on $\mathbb{Z}_n$ called "+" which is inherited from addition in $\mathbb{Z}$. In this lecture we continue studying further algebraic properties and structures on $\mathbb{Z}_n$.

Another natural operation to consider on $\mathbb{Z}_n$ is multiplication: writing $a \pmod n$ for the equivalence class of $a \in \mathbb{Z}$, we define the product of $a \pmod n$ and $b \pmod n$ to be $ab \pmod n$. Alternatively, if we write $[a]$ for the equivalence class of $a \in \mathbb{Z}$, our definition is:

$$[a][b] = [ab]$$

This is well-defined: if $[a'] = [a]$ and $[b'] = [b]$, then $a' - a = nk$ and $b' - b = nl$ for some $k, l \in \mathbb{Z}$, so $a'b' - ab = a'(b' - b) + b(a' - a) = n(la' + kb)$, and therefore $[a'b'] = [ab]$.

▶ **The operations of addition and multiplication on $\mathbb{Z}_n$ satisfy the properties:**

(Associativity for $+$)	$a + (b + c) \equiv (a + b) + c \pmod n$
(Associativity for $\times$)	$a(bc) \equiv (ab)c \pmod n$
(Identity for $+$)	$a + 0 \equiv 0 + a \equiv a \pmod n$
(Identity for $\times$)	$1 \cdot a \equiv a \cdot 1 \equiv a \pmod n$
(Inverses for $+$)	$a + (-a) \equiv (-a) + a \equiv 0 \pmod n$
(Distributivity)	$a(b + c) \equiv ab + ac \pmod n$
(Commutativity for $+$)	$a + b \equiv b + a \pmod n$
(Commutativity for $\times$)	$ab \equiv ba \pmod n$

In short, all the formal properties you are familiar with in $\mathbb{Z}$ hold in $\mathbb{Z}_n$. A structure with two operations ("addition" and "multiplication") satisfying all of the above formal properties is called a (commutative) *ring*. However we will hold off on studying general rings.

Included in the above list are the axioms for $(\mathbb{Z}_n, +)$ to be an abelian group. However, except in the degenerate case $n = 1$ for which $0 \equiv 1 \pmod 1$, the set $\mathbb{Z}_n$ with *multiplication* does not form a group. This is because $0 \in \mathbb{Z}_n$ does not have a multiplicative inverse. The problem is not just 0, however. For example, in $\mathbb{Z}_4 = \{0, 1, 2, 3\}$ we have:

$$2 \cdot 0 \equiv 0 \pmod 4, \quad 2 \cdot 1 \equiv 2 \pmod 4, \quad 2 \cdot 2 \equiv 0 \pmod 4, \quad 2 \cdot 3 \equiv 2 \pmod 4$$

Thus there is no element of $\mathbb{Z}_4$ which is a multiplicative inverse for 2; such an element $x \in \mathbb{Z}_4$ would have to have $2 \cdot x \equiv 1 \pmod 4$. After reviewing some basic properties of the integers we will show how to "correct" this problem, by eliminating certain elements from $\mathbb{Z}_n$ so that it becomes a group with multiplication.

The group $\mathbb{Z}_n^\times$

An integer d is a *divisor* of an integer a if $a = dk$ for some $k \in \mathbb{Z}$. A positive integer d is a *greatest common divisor* of a and b, written $d = \gcd(a, b)$, if it is a divisor of a and b, and for any integer d' dividing both a and b, we have that d' divides d. A good exercise is to check that there is a unique greatest common divisor for any $a, b \in \mathbb{Z}$. As an example, we have $\gcd(9, 24) = 3$. The key property we are after is:

▶ **Let $a, b \in \mathbb{Z}$ be non-zero integers. Then there exist integers $r, s \in \mathbb{Z}$ such that**

$$\gcd(a, b) = ar + bs$$

As an illustration, we can see directly that this is true for $a = 9, b = 24$:

$$9(3) + 24(-1) = 3$$

so upon choosing $r = 3, s = -1$ we have the desired relation.

To prove the statement in general, we proceed by considering the subset of $\mathbb{Z}$ defined by

$$S = \{am + bn : \ \ m, n \in \mathbb{Z},\ am + bn > 0\}$$

As a, b are non-zero, the set S is non-empty. For example, with $m = a$ and $n = b$ we have $am + bn = a^2 + b^2 > 0$, so $a^2 + b^2$ is in S. Let d be the smallest element in S. Then $d = ar + bs$ for some $r, s \in \mathbb{Z}$. The claim is that $d = \gcd(a, b)$. To prove this we use division with remainder for integers, applied to a divided by d: that is, we can write

$$a = dq + r'$$

where $q, r' \in \mathbb{Z}$ and $0 \leqslant r' < d$ (the remainder). We can then write

$$r' = a - dq = a - (ar + bs)q = a(1 - rq) - b(sq)$$

If $r' > 0$ then $r' \in S$ and is less than d, a contradiction to our minimality assumption on d. Thus $r' = 0$, and $a = dq$, so d divides a. A similar argument shows d divides b. Thus d divides both a and b. Finally, we must show that if d' divides both a and b then it divides d. If d' divides a then $a = d'k$ and if d' divides b then $b = d'l$. So $d = ar + bs = d'kr + d'ls = d'(kr + ls)$, and thus d' divides d. Therefore $d = \gcd(a, b)$. This completes the proof.

▶ **In $\mathbb{Z}_n$, $a \pmod n$ has a multiplicative inverse if and only if** $\gcd(a, n) = 1$.

To see this, first suppose $a \pmod n$ has a multiplicative inverse, i.e. there is some integer b such that $ab \equiv 1 \pmod n$, or equivalently $ab - 1 = nk$ for some integer k. Thus $ab - nk = 1$. If d divides a and n, it divides $ab - nk = 1$, so d divides 1. Thus $\gcd(a, n) = 1$. Conversely, suppose $\gcd(a, n) = 1$. There exists $r, s \in \mathbb{Z}$ such that $\gcd(a, n) = ar + ns$. Then $ar - 1 = -ns$, so $ar \equiv 1 \pmod n$. Thus $r \pmod n$ is a multiplicative inverse for $a \pmod n$.

If $\gcd(a,b) = 1$ then a, b are said to be *relatively prime*. This motivates the following definition: we let $\mathbb{Z}_n^\times$ denote the subset of $\mathbb{Z}_n$ consisting of elements relatively prime to n:

$$\mathbb{Z}_n^\times = \{\ a \pmod n : \quad \gcd(a,n) = 1\ \}$$

▶ **$\mathbb{Z}_n^\times$ equipped with the operation of multiplication defines an abelian group.**

Let us look at some examples. First, consider the integers mod 4, i.e. $\mathbb{Z}_4$. Then we have

$$\mathbb{Z}_4^\times = \{1,3\} \subset \mathbb{Z}_4 = \{0,1,2,3\}$$

As is often done, we have just written "1" etc. for the equivalence class "1 (mod 4)" which we also previously wrote as "[1]". In conclusion, $\mathbb{Z}_4^\times$ is a finite abelian group of order 2.

Next, consider $\mathbb{Z}_{10}$, the integers modulo 10. In this case we find:

$$\mathbb{Z}_{10}^\times = \{1,3,7,9\} \subset \mathbb{Z}_{10} = \{0,1,2,3,4,5,6,7,8,9\}$$

Thus $\mathbb{Z}_{10}^\times$ is a finite abelian group of order 4. For example, $3 \cdot 7 \equiv 21 \equiv 1 \pmod{10}$, so the inverse of $3 \pmod{10}$ is $7 \pmod{10}$, and conversely. Here is the Cayley table for $(\mathbb{Z}_{10}^\times, \times)$:

	1	3	7	9
1	1	3	7	9
3	3	9	1	7
7	7	1	9	3
9	9	7	3	1

When n is a prime number p, then $\mathbb{Z}_p^\times = \{1, \ldots, p-1\}$. For example, $\mathbb{Z}_7^\times = \{1,2,3,4,5,6\}$.

▶ **Suppose $\gcd(a,n) = 1$ and b any integer. Then the equation in $\mathbb{Z}_n$ given by**

$$ax \equiv b \pmod n$$

can always be solved for x, and the solution x is unique as an element of $\mathbb{Z}_n$.

To see that this statement is true, we use the fact that $a \pmod n$ has an inverse in $\mathbb{Z}_n^\times$ and multiply both sides of the equation by this inverse.

For example, consider the equation $7x \equiv 6 \pmod{10}$ in $\mathbb{Z}_{10}$. Then multiply both sides by $3 \pmod{10}$ to get $x \equiv 18 \equiv 8 \pmod{10}$. On the other hand, we saw earlier that $2x \equiv 1 \pmod 4$ has no solutions, but in this case $\gcd(2,4) = 2 \neq 1$.

Euclidean Algorithm

Above we have seen that $a \pmod n$ is invertible in $\mathbb{Z}_n$ if and only if $\gcd(a, n) = 1$. In practice, how can we find the inverse? We use the Euclidean algorithm.

The algorithm takes two integers a, b and computes $\gcd(a, b)$. By recording each step of the algorithm one has the information to find $r, s \in \mathbb{Z}$ such that

$$ar + bs = \gcd(a, b).$$

In the case that $b = n$ and $\gcd(a, n) = 1$, the inverse of $a \pmod n$ is given by $r \pmod n$.

The algorithm is as follows. Assume $a > b > 0$. Set $r_1 = a$, $r_2 = b$. Divide a by b to obtain

$$r_1 = r_2 q_1 + r_3$$

where $0 \leq r_3 < r_2$ is the remainder. Continue in this fashion to obtain a sequence of non-negative integers $r_1, r_2, r_3, \ldots$: if one has computed up to r_k, then divide r_{k-1} by r_k to obtain

$$r_{k-1} = r_k q_{k-1} + r_{k+1}$$

where $0 \leq r_{k+1} < r_k$ is the remainder. As each r_k is non-negative and smaller than the previous entry r_{k-1}, this process must eventually stop. The last non-zero r_k obtained is $\gcd(a, b)$!

Let us do a simple example: $a = 17$, $b = 11$. We compute:

$$\begin{aligned} 17 &= 11 \cdot 1 + 6 & 6&= 17 - 11 \\ 11 &= 6 \cdot 1 + 5 & 5&= 11 - 6 \\ 6 &= 5 \cdot 1 + 1 & 1&= 6 - 5 \end{aligned}$$

The left column shows the algorithm as described; it terminates at "1" which is $\gcd(17, 11)$. But we can say more! In the right column we have rearranged each equation to solve for the remainder. Now starting from "1" (the gcd) we continually substitute the expressions in the right column to obtain an end result in terms of the original $a = 17$ and $b = 11$:

$$\begin{aligned} \gcd(17, 11) = 1 &= 6 - 5 \\ &= (17 - 11) - (11 - 6) \\ &= 17 - 2(11) + 6 \\ &= 17 - 2(11) + (17 - 11) \\ &= 2(17) - 3(11) \end{aligned}$$

In particular, we see that $-3(11) \equiv 1 \pmod{17}$, and so $-3 \pmod{17}$, which is the same as $14 \pmod{17}$, is the inverse of $11 \pmod{17}$.

Orders of group elements

Let G be an arbitrary group, and let $a \in G$ be an element. The *order* of a, written $\mathrm{ord}(a)$, is the smallest positive integer k such that a^k is the identity:

$$\mathrm{ord}(a) = \min\left\{k \in \mathbb{Z} :\ k > 0,\ a^k = e\right\}$$

Our convention is that if the set appearing in this definition is empty, in other words if a^k is never the identity, then $\mathrm{ord}(a) = \infty$.

▶ **If $a \in G$ and G is a finite group, then $\mathbf{ord}(a) \leqslant |G|$.**

To prove this, consider the sequence $a, a^2, a^3, \ldots$. Because this is a sequence of elements in G and G has $|G|$ elements, after $|G| + 1$ terms in this sequence we must have a repeated entry. In other words, there are positive integers i, j such that $i < j$, and $j \leqslant |G| + 1$, and $a^i = a^j$. Then $e = a^j a^{-i} = a^{j-i}$ and therefore $\mathrm{ord}(a) \leqslant j - i \leqslant |G|$.

Let's consider some examples. First, note if $\mathrm{ord}(a) = 1$, then $a = a^1 = e$, so a is the identity.

Consider the group $(\mathbb{Z}_n, +)$. If $a \in \mathbb{Z}_n$ is equal to $m \pmod n$, then "$a^k = a \cdots a$" is given by

$$m + \cdots + m \equiv km \pmod n$$

In this case $|G| = n$ and $n \cdot m \equiv 0 \pmod n$, which exhibits $\mathrm{ord}(a) \leqslant |G|$. Let us consider the case $n = 8$. The element $6 \pmod 8$ has $4 \cdot 6 \equiv 24 \equiv 0 \pmod 8$, while $3 \cdot 6 \equiv 18 \equiv 2 \pmod 8$, $2 \cdot 6 \equiv 12 \equiv 4 \pmod 8$. Thus the order of $6 \pmod 8$ in the group $\mathbb{Z}_8$ is equal to 4.

▶ **The order of $m \pmod n$ in the group $(\mathbb{Z}_n, +)$ is equal to $n/\mathbf{gcd}(n, m)$.**

Let us prove this. Note $n/\gcd(n, m)$ and $m/\gcd(n, m)$ are integers. Then we have

$$\frac{n}{\gcd(n, m)} \cdot m \equiv n \cdot \frac{m}{\gcd(n, m)} \equiv 0 \pmod n$$

Therefore $\mathrm{ord}(m) \leqslant n/\gcd(n, m)$. On the other hand, letting $\mathrm{ord}(m) = k$, we have $km \equiv 0 \pmod n$, i.e. $km = nl$ for some $l \in \mathbb{Z}$. Then we have the following relation:

$$k \cdot \frac{m}{\gcd(n, m)} = \frac{n}{\gcd(n, m)} \cdot l$$

Each factor is an integer. From the definition of the greatest common divisor it follows that $n/\gcd(n, m)$ and $m/\gcd(n, m)$ are relatively prime. From this it follows that k is divisible by $n/\gcd(n, m)$, and in particular $\mathrm{ord}(m) = k \geqslant n/\gcd(n, m)$. Thus $\mathrm{ord}(m) = n/\gcd(n, m)$.

For example, $8/\gcd(8, 6) = 8/2 = 4$ and this agrees with our previous determination of the order of $6 \pmod 8$ inside $(\mathbb{Z}_8, +)$ from above.

▶ **If $\gcd(n, m) = 1$, then $m \pmod n$ generates the group $(\mathbb{Z}_n, +)$.**

This follows from the fact that $\gcd(n, m) = 1$ implies m has order $n = |\mathbb{Z}_n|$. (See also the next proposition below.) Recall that in general $a \in G$ *generates* G if all elements in G can be written as a^k for some $k \in \mathbb{Z}$. To illustrate this for $3 \in (\mathbb{Z}_{10}, +)$, we compute in $\mathbb{Z}_{10}$:

$$1 \cdot 3 = 3, \qquad 2 \cdot 3 = 6, \qquad 3 \cdot 3 = 9, \qquad 4 \cdot 3 = 2, \qquad 5 \cdot 3 = 5,$$

$$6 \cdot 3 = 8, \qquad 7 \cdot 3 = 1, \qquad 8 \cdot 3 = 4, \qquad 9 \cdot 3 = 7, \qquad 10 \cdot 3 = 0.$$

Thus we get all of the group $(\mathbb{Z}_{10}, +)$ by successively adding 3 to itself.

We next turn to the group $(\mathbb{Z}_n^\times, \times)$. Unlike the previous examples, the notation in this group agrees nicely with the general notation $a^k = a \cdots a$ for the k^{th} power of an element.

Suppose $n = 10$. Then $\mathbb{Z}_{10}^\times = \{1, 3, 7, 9\}$. We have $3^2 \equiv 9 \pmod{10}$ and $3^3 \equiv 27 \equiv 7 \pmod{10}$ while $3^4 \equiv 3 \cdot 7 \equiv 21 \equiv 1 \pmod{10}$. Thus the order of $3 \in \mathbb{Z}_{10}^\times$ is equal to 4. On the other hand, $9^2 \equiv 81 \equiv 1 \pmod{10}$ so the order of 9 is equal to 2.

It is important, as always, to know what group we are working in and what the group operation is. For example, $3 \pmod{10}$ can also be viewed as an element of the group $(\mathbb{Z}_{10}, +)$, but in this case it has order $10/\gcd(10, 3) = 10$ as we saw above.

▶ **Given $a \in G$, let $\langle a \rangle = \{a^k : k \in \mathbb{Z}\} \subset G$. Then $\langle a \rangle$ is a subgroup, and $\mathbf{ord}(a) = |\langle a \rangle|$. If $\mathbf{ord}(a)$ is finite, then $\langle a \rangle$ consists of the distinct elements**

$$e = a^0,\ a^1,\ a^2,\ a^3,\ \ldots,\ a^{\mathbf{ord}(a)-1}$$

If $\mathbf{ord}(a) = \infty$ then $a^k = a^l$ for $k, l \in \mathbb{Z}$ if and only if $k = l$.

The verification that $\langle a \rangle$ is a subgroup is straightforward and left as an exercise. So let us suppose $\mathrm{ord}(a)$ is finite. Let $m \in \mathbb{Z}$. We aim to show that $a^m \in \langle a \rangle$ is among the list of elements shown above. For this we use division with remainder: write

$$m = \mathrm{ord}(a) \cdot q + r$$

where $q, r \in \mathbb{Z}$ and $0 \leqslant r < \mathrm{ord}(a)$, the remainder. Then we compute

$$a^m = a^{\mathrm{ord}(a)\cdot q + r} = a^{\mathrm{ord}(a)\cdot q} a^r = (a^{\mathrm{ord}(a)})^q a^r = e^q a^r = a^r$$

Furthermore, a^r is on the list, because $0 \leqslant r < \mathrm{ord}(a)$. This shows $\langle a \rangle$ consists of the elements in the list given, and in particular $|\langle a \rangle| \leqslant \mathrm{ord}(a)$. Finally, we know $\mathrm{ord}(a) \leqslant |\langle a \rangle|$, and so $|\langle a \rangle| = \mathrm{ord}(a)$, and consequently the elements in the list must all be distinct. The case in which $\mathrm{ord}(a) = \infty$ is left as an exercise.

Recall that a group G is *cyclic* if there is an $a \in G$ such that $\langle a \rangle = G$. Thus G is cyclic if and only if there is an element of order $|G|$. Recall $(\mathbb{Z}, +)$ and $(\mathbb{Z}_n, +)$ are cyclic.

Let us illustrate these concepts with a few more examples. Consider $\mathbb{Z}_5^\times = \{1, 2, 3, 4\}$. This is a group of order 4 and we display its Cayley table below. Note $2^0 = 1$, $2^1 = 2$, $2^2 = 4$, $2^3 = 3$, where of course all operations are modulo 5. Thus $2 \in \mathbb{Z}_5^\times$ generates the group. The same is true for 3 (but not for 1 or 4). Thus $\mathbb{Z}_5^\times$ is cyclic.

	1	2	3	4
1	1	2	3	4
2	2	4	1	3
3	3	1	4	2
4	4	3	2	1

Next, consider the group $(\mathbb{Z}_8^\times, \times)$. We have $\mathbb{Z}_8^\times = \{1, 3, 5, 7\}$ and so like the previous example, this is a group of order 4. Its Cayley table is displayed below.

	1	3	5	7
1	1	3	5	7
3	3	1	7	5
5	5	7	1	3
7	7	5	3	1

However we see that every element squares to the identity. In particular, no element is of order 4. Therefore $\mathbb{Z}_8^\times$ is *not cyclic*.

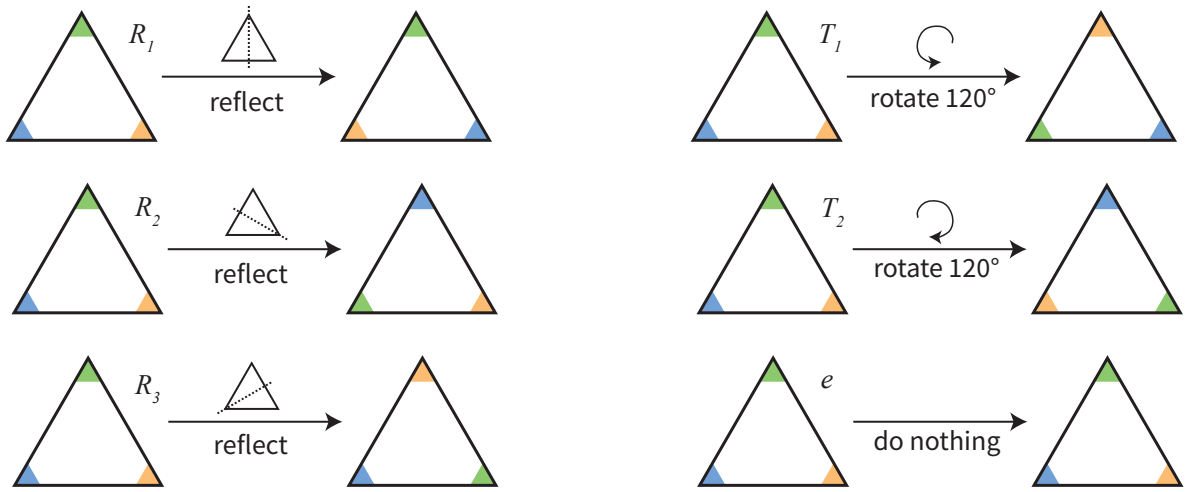
Finally, consider the following element of the general linear group $\mathrm{GL}_2(\mathbb{R})$:

$$A = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$

A short computation shows $A^2 = -I$, $A^3 = -A$, $A^4 = I$ where I is the 2×2 identity matrix. Thus $\mathrm{ord}(A) = 4$. This is an example of an element of finite order in an infinite group.

Symmetries

A *symmetry* of an object is a rearrangement of the object such that its geometric properties are preserved. If the object is a subset X of points in space, then a symmetry is a 1-1 map $f : X \to X$ that preserves distances and angles. For example, consider an equilateral triangle $X \subset \mathbb{R}^2$ in 2-dimensional space. A symmetry maps the triangle to itself in a way which sends vertices to vertices, and edges to edges. All symmetries of the triangle are as follows:



Let us give each such symmetry the name as indicated in the figure. Thus R_1 is "reflection across the vertical line" and so on. We use the notation

$$R_1 \circ T_1$$

to mean "first T_1 then R_1." In this case this means "rotate 120° counter-clockwise, then reflect across the vertical line." (The order is just as for functions, or maps.) We see that

$$R_1 \circ T_1 = R_3$$

We also have $R_1 \circ R_1 = e$, $R_1 \circ R_2 = T_2$, $T_1 \circ T_2 = e$. Eventually we obtain a Cayley table:

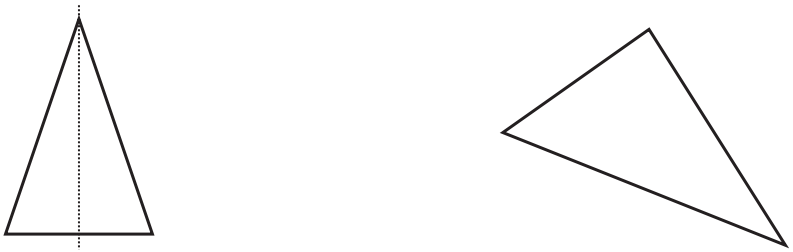
	e	R_1	R_2	R_3	T_1	T_2
e	e	R_1	R_2	R_3	T_1	T_2
R_1	R_1	e	T_2	T_1	R_3	R_2
R_2	R_2	T_1	e	T_2	R_1	R_3
R_3	R_3	T_2	T_1	e	R_2	R_1
T_1	T_1	R_2	R_3	R_1	T_2	e
T_3	T_2	R_3	R_1	R_2	e	T_1

This should look familiar: it is exactly the Cayley table for the non-abelian group of order 6 that we discussed in Lecture 1 (after renaming symbols).

▶ **The symmetries of an object forms a group.**

This can be seen explicitly in the example above, but the general case is also straightforward. The key is to think of symmetries as 1-1 maps from a set X to itself which satisfy certain properties (the maps preserve distances and angles, for example).

Let us consider other objects. Take an isosceles triangle in the plane. Then apart from the identity symmetry e, there is only one symmetry R, which is a reflection across the line indicated in the figure below on the left. Further, $R \circ R = e$. Thus the group of symmetries of an isosceles triangle is $\{e, R\}$ with $R^2 = e$. This is a cylic group of order 2.



On the other hand, if we take an arbitrary triangle with 3 different angles, so that no two sides are the same length, then there are no symmetries apart from the identity. See the figure above on the right. The group of symmetries of such a triangle is $\{e\}$, a trivial group.

Permutations

Let X be a finite non-empty set. Suppose $|X| = n$. We think of X as an abstract collection of n points, with no geometric information regarding the points. In this case the "symmetries of X" are simply the 1-1 maps $\sigma : X \to X$, i.e. the ways of rearranging the points in the set. These are called *permutations* of the set X. We will not care about the names of the elements of X, so we may as well let $X = \{1, \ldots, n\}$.

▶ **The group of permutations on $\{1, \ldots, n\}$ is called S_n, the n^{th} symmetric group.**

An element of S_n, which is a map from $\{1, \ldots, n\}$ to itself, may be represented as:

$$\sigma = \begin{pmatrix} 1 & 2 & \cdots & n \\ \sigma(1) & \sigma(2) & \cdots & \sigma(n) \end{pmatrix}$$

We also write $i \mapsto j$ to mean σ sends i to j, i.e. $\sigma(i) = j$. For example, let $\sigma \in S_6$ map $1 \mapsto 2$, $2 \mapsto 3$, $3 \mapsto 5$, $4 \mapsto 6$, $5 \mapsto 1$, $6 \mapsto 4$. Then we write this permutation as

$$\sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 2 & 3 & 5 & 6 & 1 & 4 \end{pmatrix}$$

Let us consider another permutation $\tau \in S_6$, given as follows:

$$\tau = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 3 & 1 & 4 & 2 & 5 & 6 \end{pmatrix}$$

Let us compute $\tau \circ \sigma$. Note that the notation means "do σ, then τ." We find:

$$(\tau \circ \sigma)(1) = \tau(\sigma(1)) = \tau(2) = 1 \qquad (\tau \circ \sigma)(2) = \tau(\sigma(2)) = \tau(3) = 4$$
$$(\tau \circ \sigma)(3) = \tau(\sigma(3)) = \tau(5) = 5 \qquad (\tau \circ \sigma)(4) = \tau(\sigma(4)) = \tau(6) = 6$$
$$(\tau \circ \sigma)(5) = \tau(\sigma(5)) = \tau(1) = 3 \qquad (\tau \circ \sigma)(6) = \tau(\sigma(6)) = \tau(4) = 2$$

Writing the result in our notation introduced above we then have:

$$\tau \circ \sigma = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 1 & 4 & 5 & 6 & 3 & 2 \end{pmatrix}$$

There is a more compact notation, called cycle notation. We illustrate this notation in the current example, and also comment on the general case in the next lecture. Starting with the element "1" we can successively apply the permutation σ until we get back to "1":

$$1 \overset{\sigma}{\longmapsto} 2 \overset{\sigma}{\longmapsto} 3 \overset{\sigma}{\longmapsto} 5 \overset{\sigma}{\longmapsto} 1$$

This forms a *cycle* and is denoted (1235). The notation is meant to be cyclic, and can be written also as (2351), (3512) or (5123). To describe the rest of the permutation σ we must also know what it does to "4" and "6". Note that

$$4 \overset{\sigma}{\longmapsto} 6 \overset{\sigma}{\longmapsto} 4$$

and this forms another cycle $(46) = (64)$. We then write σ in cycle notation as follows:

$$\sigma = (1235)(46) = (46)(1235)$$

We can also write τ in cycle notation, as $\tau = (1342)$. Note that "5" and "6" are fixed by τ and they do not appear in the notation, by convention. We then compute

$$\tau \circ \sigma = (1342)(1235)(46) = (246)(35)$$

by reading cycles from right to left, and this is the same answer we obtained above.

▶ **The n^{th} symmetric group S_n is non-abelian if $n \geqslant 3$.**

Note $S_1 = \{e\}$ is the trivial group, and $S_2 = \{e, (12)\}$ is cyclic of order 2. For $n = 3$ we have

$$S_3 = \{e, (12), (23), (31), (123), (132)\}$$

This is a non-abelian group of order 6. In fact if you write out its Cayley table you may notice it looks just like the symmetries of an equilateral triangle. This is no accident, for S_3 is essentially the same group: any symmetry of the triangle is determined by how the 3 vertices are permuted. To see that S_n is non-abelian for $n \geqslant 4$, note that we can view $S_3 \subset S_n$ as the subgroup which fixes the last $n - 3$ elements.

Symmetric groups

In this lecture we continue our study of the symmetric groups S_n introduced last lecture, and we then introduce the notion of *even* and *odd* permutations.

▶ **The order of the group S_n is given by $|S_n| = n!$.**

This is a basic counting exercise. Each permutation σ is a way of rearranging the elements $\{1, \ldots, n\}$. There are n possibilities for $\sigma(1) \in \{1, \ldots, n\}$. Once the value of $\sigma(1)$ is fixed, there are $n-1$ possibilities left for $\sigma(2)$, since it can be anything except for $\sigma(1)$. Continuing in this fashion, we see that the number of possibilities for the permutation $\sigma \in S_n$ is equal to $n(n-1)(n-2)\cdots 2 \cdot 1 = n!$.

For example, we have seen that $|S_1| = 1 = 1!$, $|S_2| = 2 = 2!$ and $|S_3| = 6 = 3!$. Recall that S_3 is essentially the symmetries of an equilateral triangle, and there are 6 such symmetries.

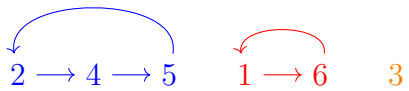
A *cycle* is a permutation $\sigma \in S_n$ such that there exists an $i \in \{1, \ldots, n\}$ with the property that for each $j \in \{1, \ldots, n\}$ either $\sigma(j) = j$ or $j = \sigma^k(i)$ for some $k \in \mathbb{Z}$. In other words, every element of $\{1, \ldots, n\}$ is either fixed by σ or can be obtained from i by successively applying σ. Our notation for such a cycle, which we introduced last lecture, is as follows:

$$(i \quad \sigma(i) \quad \sigma^2(i) \quad \cdots \quad \sigma^{l-1}(i))$$

Here l is the smallest positive integer such that $\sigma^l(i) = i$.

▶ **Every $\sigma \in S_n$ can be written as a product of disjoint cycles.**

Two cycles are *disjoint* if they share no common elements in their cycle notations. For example, (13) and (246) are disjoint cycles, while (12) and (254) are not. To prove the statement, define a relation on the set $\{1, \ldots, n\}$ as follows: $i \sim j$ if and only if $\sigma^k(i) = j$ for some $k \in \mathbb{Z}$. You can check that this defines an equivalence relation, and hence partitions the set $\{1, \ldots, n\}$, and each equivalence class forms a cycle with respect to σ.



The content of the above is that any permutation $\sigma \in S_n$ partitions the set $\{1, 2, \ldots, n\}$ into subsets each of which σ permutes in a cyclic fashion. For example, $\sigma = (245)(16) \in S_6$ is depicted above. Note that σ fixes "3" and this does not appear in cycle notation.

▶ **Disjoint cycles in S_n commute.**

The *length* of a cycle is the number of elements appearing in it: (146) is a cycle of length 3, and (4263) is length 4. A *transposition* is a cycle of length 2, such as (12).

▶ **For $n \geqslant 2$, every $\sigma \in S_n$ is a product of transpositions.**

We first note that an arbitrary cycle $(a_1 a_2 \cdots a_k)$ can be written

$$(a_1 a_2 \cdots a_{k-1} a_k) = (a_3 a_2)(a_4 a_3) \cdots (a_k a_{k-1})(a_1 a_k)$$

For example, $(245) = (45)(24)$. The general case follows from the case of cycles, because any $\sigma \in S_n$ is a product of cycles.

The way in which a cycle is written as a product of transpositions is not unique. As an example, we have $(12)(13) = (132) = (23)(12)$ in the group S_3. In fact, another formula for writing an arbitrary cycle as a product of transpositions is as follows:

$$(a_1 a_2 \cdots a_{k-1} a_k) = (a_1 a_k)(a_1 a_{k-1}) \cdots (a_1 a_3)(a_1 a_2)$$

Even and odd permutations

Let $f = f(x_1, \ldots, x_n)$ be a polynomial in n variables. Given a permutation $\sigma \in S_n$ define a new polynomial $\sigma(f)$ to be $f(x_{\sigma(1)}, \ldots, x_{\sigma(n)})$. Define the special polynomial

$$\Delta_n = \prod_{1 \leqslant i < j \leqslant n} (x_i - x_j)$$

to be the product of all factors $(x_i - x_j)$ where i, j range over elements of $\{1, \ldots, n\}$ with $i < j$. Then we may apply any permutation σ to this to obtain a new polynomial $\sigma(\Delta_n)$. It turns out that this new polynomial is always either Δ_n or $-\Delta_n$. This lets us divide permutations into two types, as follows:

▶ **If $\sigma(\Delta_n) = \Delta_n$ then σ is called *even*. If $\sigma(\Delta_n) = -\Delta_n$ then σ is called *odd*.**

For example, consider $\sigma = (12) \in S_3$. Note $\Delta_3 = (x_1 - x_2)(x_1 - x_3)(x_2 - x_3)$. Then

$$\begin{aligned} \sigma(\Delta_3) &= (x_{\sigma(1)} - x_{\sigma(2)})(x_{\sigma(1)} - x_{\sigma(3)})(x_{\sigma(2)} - x_{\sigma(3)}) \\ &= (x_2 - x_1)(x_1 - x_3)(x_2 - x_3) \\ &= -(x_1 - x_2)(x_2 - x_3)(x_1 - x_3) \\ &= -\Delta_3 \end{aligned}$$

Thus $\sigma = (12)$ is odd. In the next lecture we will show that the *even* permutations form a subgroup of S_n called the *alternating group* A_n.

Alternating groups

In this lecture we continue studying even and odd permutations. We introduce and study the *alternating groups* A_n which consist of *even* permutations. We then consider the rotational symmetries of the tetrahedron, which are closely related the group A_4.

Recall that a permutation $\sigma \in S_n$ is *even* if $\sigma(\Delta_n) = \Delta_n$ and *odd* if $\sigma(\Delta_n) = -\Delta_n$. Here Δ_n is the polynomial $\prod_{1 \leqslant i < j \leqslant n}(x_i - x_j)$ introduced last lecture. Define

$$A_n = \{\sigma \in S_n : \sigma \text{ is even}\} \subset S_n$$

▶ **The subset $A_n \subset S_n$ is a subgroup, called the n^{th} *alternating group*.**

To prove this we first record a useful relation. Given any permutations $\sigma, \sigma' \in S_n$ we have

$$(\sigma\sigma')(\Delta_n) = \sigma(\sigma'(\Delta_n))$$

This just follows by writing out what each side means explicitly:

$$\begin{aligned}(\sigma\sigma')(\Delta_n) &= \prod_{1 \leqslant i < j \leqslant n} (x_{(\sigma\sigma')(i)} - x_{(\sigma\sigma')(j)}) = \prod_{1 \leqslant i < j \leqslant n} (x_{\sigma(\sigma'(i))} - x_{\sigma(\sigma'(j))}) \\ &= \sigma\left(\prod_{1 \leqslant i < j \leqslant n} (x_{\sigma'(i)} - x_{\sigma'(j)})\right) = \sigma(\sigma'(\Delta_n))\end{aligned}$$

Similarly, $\sigma(-\Delta_n) = -\sigma(\Delta_n)$. Now suppose $\sigma(\Delta_n) = (-1)^k\Delta_n$ and $\sigma'(\Delta_n) = (-1)^l\Delta_n$. Then

$$\begin{aligned}(\sigma\sigma')(\Delta_n) &= \sigma(\sigma'(\Delta_n)) = \sigma((-1)^l\Delta_n) \\ &= (-1)^l\sigma(\Delta_n) = (-1)^l(-1)^k\Delta_n \\ &= (-1)^{l+k}\Delta_n\end{aligned}$$

From this computation we see the same rules as for adding even and odd integers:

σ	σ'	$\sigma\sigma'$
even	even	even
odd	even	odd
even	odd	odd
odd	odd	odd

In particular, if $\sigma, \sigma' \in A_n$ (σ, σ' are both even) then $\sigma\sigma' \in A_n$ ($\sigma\sigma'$ is even). Also the identity is even, so it is in A_n. Further, if $\sigma \in A_n$ (σ is even), then since $\sigma\sigma^{-1} = e \in A_n$ ($\sigma\sigma^{-1}$ is even) we must have $\sigma^{-1} \in A_n$ (σ^{-1} is even). Thus A_n is a subgroup of S_n.

An alternative characterization of the parity of a permutation is as follows:

▶ **If $\sigma = \tau_1 \cdots \tau_k$ where τ_i are transpositions, then σ is odd if and only if k is odd.**

To prove this we first show that every transposition is odd. First consider the transposition $\sigma = (12) \in S_n$. There are four kinds of factors in Δ_n:

$$(x_1 - x_2), \qquad (x_1 - x_j) \ \ (j > 2), \qquad (x_2 - x_j) \ \ (j > 2), \qquad (x_i - x_j) \ \ (j > i > 2)$$

Now $\sigma = (12)$ only swaps 1 and 2. So it sends the first type of factor to its negative $(x_2 - x_1) = -(x_2 - x_1)$. It interchanges the second and third types (preserving signs), and fixes all factors of the fourth type. Taking the product we conclude $\sigma(\Delta_n) = -\Delta_n$, where the sign comes from the effect of $\sigma = (12)$ on the factor $(x_1 - x_2)$. Next, we use:

▶ **Let $\sigma = (a_1\ a_2\ \cdots\ a_k) \in S_n$ be a cycle and $\tau \in S_n$ any other permutation. Then**

$$\tau\sigma\tau^{-1} = (\tau(a_1)\ \ \tau(a_2)\ \ \cdots\ \ \tau(a_k))$$

A special case is when $\sigma = (i\,j)$ a transposition different from (12) with $j > i$. Setting $\tau = (i\ 1)(j\ 2)$ we get $\tau^{-1}\sigma\tau = (12)$. If $i = 1$, interpret $(i\ 1)$ as e.

Now let σ be any transposition and choose τ as above such that $\tau\sigma\tau^{-1} = (12)$. Then $\sigma = \tau^{-1}(12)\tau$. Let $\tau(\Delta_n) = (-1)^k\Delta_n$. Note also $\tau^{-1}(\Delta_n) = (-1)^k\Delta_n$. We then compute

$$\begin{aligned}
\sigma(\Delta_n) &= (\tau^{-1}(12)\tau)(\Delta_n) \\
&= \tau^{-1}((12)(\tau(\Delta_n))) \\
&= \tau^{-1}((12)(-1)^k\Delta_n)) \\
&= (-1)^k\tau^{-1}((12)(\Delta_n)) \\
&= (-1)^k\tau^{-1}(-\Delta_n) \\
&= (-1)^{k+1}\tau^{-1}(\Delta_n) \\
&= (-1)^{2k+1}\Delta_n = -\Delta_n
\end{aligned}$$

This completes our claim that every transposition is odd. Then to prove the claim about $\sigma = \tau_1 \cdots \tau_k$ for a product of transpositions, we use the rules of the table we determined above.

Let us look at some examples. As $S_1 = \{e\}$ we of course have $A_1 = \{e\}$. Next, $S_2 = \{e, (12)\}$, and (12) is odd, so in fact $A_2 = \{e\}$ as well. The 3rd symmetric group is

$$S_3 = \{e, (12), (23), (31), (123), (132)\}$$

The three transpositions (12), (23), (31) are odd, so they are not in A_3. On the other hand $(123) = (13)(12)$ and $(132) = (12)(13)$, so these are even. Thus

$$A_3 = \{e, (123), (132)\}$$

Note that $(123)^2 = (132)$ and $(123)^3 = e$, so A_3 is a cyclic (hence abelian) group of order 3. This is in contrast to S_3. However:

▶ **The alternating group A_n is non-abelian if and only if $n \geqslant 4$.**

Symmetries of the tetrahedron

The first non-abelian alternating group, A_4, is closely related to the rotational symmetries of the tetrahedron in 3-dimensional space.

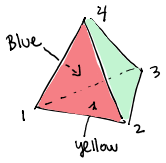
A tetrahedron is a solid in 3-dimensional Euclidean space which has 4 vertices and 4 sides, each an equilateral triangle. On the next page we list the symmetries of the tetrahedron. There are 2 types of non-identity symmetries. The first type $(R_1^{\pm 1}, R_2^{\pm 1}, R_3^{\pm 1}, R_4^{\pm 1})$ fixes a vertex and rotates the tetrahedron around an axis passing through the fixed vertex by 120° in one of two directions. The second kind of symmetry (A, B, C) is a 180° rotation through an axis which passes through the centers of two opposite edges.

If we label the vertices of the tetrahedron by $\{1, 2, 3, 4\}$ we can associate a permutation to each symmetry. Magically, the subgroup of S_4 corresponding to the symmetries of the tetrahedron is A_4! Below we include the Cayley table.

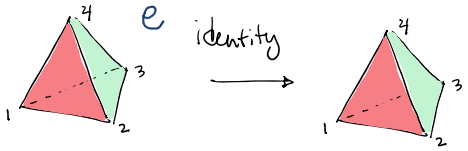
	e	R_1	R_1^{-1}	R_2	R_2^{-1}	R_3	R_3^{-1}	R_4	R_4^{-1}	A	B	C
e	e	R_1	R_1^{-1}	R_2	R_2^{-1}	R_3	R_3^{-1}	R_4	R_4^{-1}	A	B	C
R_1	R_1	R_1^{-1}	e	A	R_4	B	R_2	C	R_3	R_3^{-1}	R_4^{-1}	R_2^{-1}
R_1^{-1}	R_1^{-1}	e	R_1	R_3^{-1}	C	R_4^{-1}	A	R_2^{-1}	B	R_2	R_3	R_4
R_2	R_2	C	R_4^{-1}	R_2^{-1}	e	R_1	B	R_3^{-1}	A	R_1^{-1}	R_4	R_3
R_2^{-1}	R_2^{-1}	R_3	A	e	R_2	C	R_4	B	R_1^{-1}	R_4^{-1}	R_3^{-1}	R_1
R_3	R_3	A	R_2^{-1}	R_4^{-1}	B	R_3^{-1}	e	R_1	C	R_4	R_1^{-1}	R_2
R_3^{-1}	R_3^{-1}	R_4	B	C	R_1^{-1}	e	R_3	A	R_2	R_1	R_2^{-1}	R_4^{-1}
R_4	R_4	B	R_3^{-1}	R_1	A	R_2^{-1}	C	R_4^{-1}	e	R_3	R_2	R_1^{-1}
R_4^{-1}	R_4^{-1}	R_2	C	B	R_3	A	R_1^{-1}	e	R_4	R_2^{-1}	R_1	R_3^{-1}
A	A	R_2^{-1}	R_3	R_4	R_1	R_1^{-1}	R_4^{-1}	R_2	R_3^{-1}	e	C	B
B	B	R_3^{-1}	R_4	R_3	R_4^{-1}	R_2	R_1	R_1^{-1}	R_2^{-1}	C	e	A
C	C	R_4^{-1}	R_2	R_1^{-1}	R_3^{-1}	R_4	R_2^{-1}	R_3	R_1	B	A	e

Symmetries of the tetrahedron.

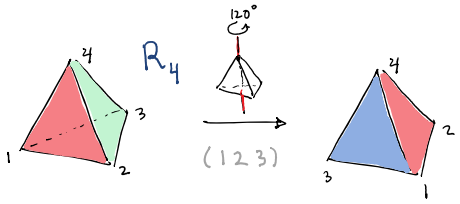
Tetrahedron :



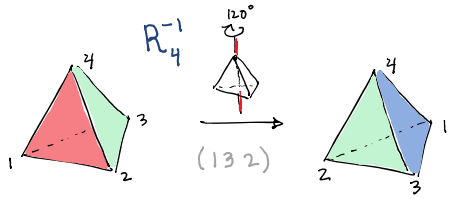
e identity



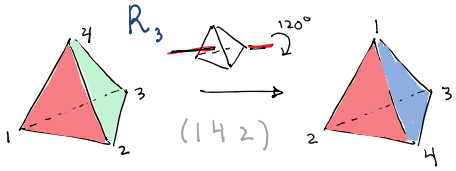
R_4 (1 2 3)



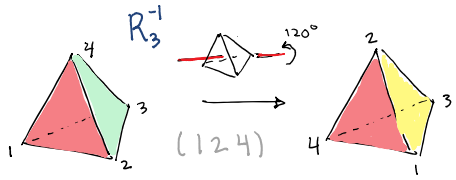
R_4^{-1} (1 3 2)



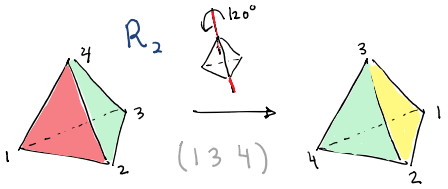
R_3 (1 4 2)



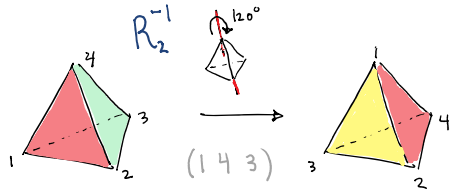
R_3^{-1} (1 2 4)



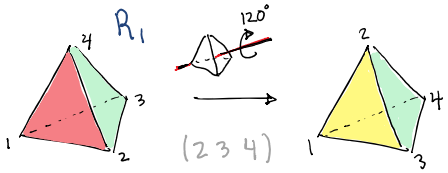
R_2 (1 3 4)



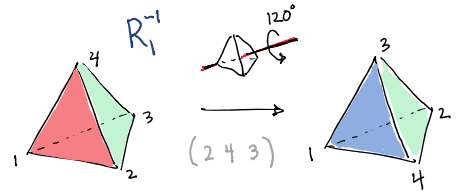
R_2^{-1} (1 4 3)



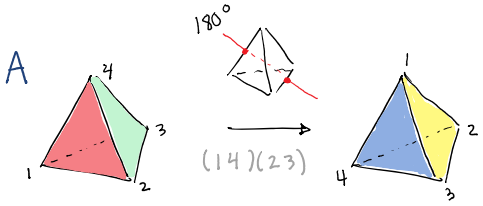
R_1 (2 3 4)



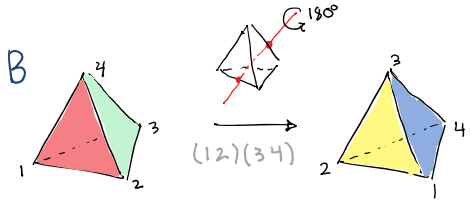
R_1^{-1} (2 4 3)



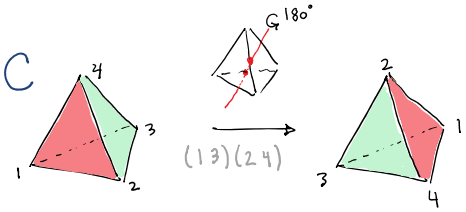
A (1 4)(2 3)



B (1 2)(3 4)



C (1 3)(2 4)



Cosets and Lagrange's Theorem

In this lecture we introduce the notion of a coset and prove the famous result of Lagrange regarding the divisibility of the order of a group by the orders of its subgroups.

Fix a group G and a subgroup $H \subset G$. Define a relation $\sim$ on the set G such that for $a, b \in G$:

$$a \sim b \iff ab^{-1} \in H$$

Keep in mind that $\sim$ depends on H. We show that this is an equivalence relation.

1. (Reflexivity) $a \sim a$ because $aa^{-1} = e$ and the identity is in any subgroup.
2. (Symmetry) $a \sim b$ implies $ab^{-1} \in H$. Since H is a subgroup, the inverse of this element is also in H: we have $(ab^{-1})^{-1} = ba^{-1} \in H$. Thus $b \sim a$.
3. (Transitivity) $a \sim b$ and $b \sim c$ imply $ab^{-1} \in H$ and $bc^{-1} \in H$. Since H is a subgroup, it is closed under the group operation. Thus $(ab^{-1})(bc^{-1}) = ac^{-1} \in H$, and $a \sim c$.

We have seen this construction before in a special case. Let $G = (\mathbb{Z}, +)$ and for a fixed positive integer n take the subgroup $H = n\mathbb{Z} = \{nk : k \in \mathbb{Z}\} \subset \mathbb{Z}$. Then $a \sim b$ if and only if "ab^{-1}" $= a - b \in n\mathbb{Z}$, i.e. $a \equiv b \pmod{n}$. This motivates the following general notation:

$$a \sim b \qquad \iff \qquad a \equiv b \pmod{H}$$

▶ **For $a \in G$, let $Ha = \{ha : h \in H\}$. Then Ha is called a *right coset* of H in G.**

The right cosets of H in G are the equivalence classes of the above relation:

$$Ha = \{b \in G : \ a \equiv b \pmod{H}\}$$

To see this, consider some $b \in Ha$. Then $b = ha$ where $h \in H$. From this we then find $ab^{-1} = h^{-1} \in H$ and so $a \equiv b \pmod{H}$. Thus Ha is a subset of the equivalence class of a. Conversely consider any $b \in G$ such that $a \equiv b \pmod{H}$. Then $ab^{-1} \in H$, so $ab^{-1} = h$ for some $h \in H$, and so $b = h^{-1}a \in Ha$.

▶ **There is a 1-1 correspondence between any two right cosets of H in G.**

Let Ha be a right coset. It suffices to show that Ha is in 1-1 correspondence with H itself. For this, we note that each $h \in H$ determines the element $ha \in Ha$, and every element in Ha is of this form. Thus the only thing to check is that if $ha = h'a$ then $h = h'$, and this just follows from multiplying by a^{-1} on the right.

We define the *index* of a subgroup H in G, written $[G : H]$, as follows:

$$[G : H] = \#\{\text{distinct right cosets of } H \text{ in } G\}$$

Of course it is possible that $[G : H]$ is infinite.

▶ **(Lagrange's Theorem) If G is a finite group, and H is a subgroup of G, then**

$$[G : H] = |G|/|H|$$

In particular, if G is finite, the order of any subgroup H divides the order of the group G. The proof follows from our discussion above: the right cosets in G are equivalence classes, and partition the set G into $[G : H]$ distinct subsets, each of which has size $|H|$. From this it follows that $|G| = [G : H] \cdot |H|$.

Let's see all of this in action. Take the symmetric group S_3 of orer 6:

$$S_3 = \{e, (12), (23), (31), (123), (132)\}$$

Let H be the order 2 cyclic subgroup $\{e, (12)\}$. Then the right cosets are

$$He = \{e, (12)\}, \qquad H(23) = \{(23), (123)\}, \qquad H(31) = \{(31), (132)\}$$

Any other right coset is one of the above 3: we have $H(12) = He = H$, $H(123) = H(23)$ and $H(132) = H(31)$. The number of distinct right cosets is $[S_3 : H] = 3$. We directly observe Lagrange's Theorem: $6/2 = |S_3|/|H| = [S_3 : H] = 3$.

For another example, consider the symmetric group S_4. This has order $|S_4| = 4! = 24$. We saw last lecture that the alternating group $A_4 \subset S_4$ has order 12. Thus

$$[S_4 : A_4] = |S_4|/|A_4| = 24/12 = 2$$

In particular, there are exactly two right cosets: $A_4 = A_4e$ and $A_4\sigma$ where σ is any odd permutation, say, a transposition.

Assume $n \geq 2$. For the symmetric group S_n, and the subgroup $A_n \subset S_n$, there are exactly two right cosets. To see this, we note that $a \equiv b \pmod{A_n}$ if and only if ab^{-1} is even. Thus the two equivalence classes, i.e. right cosets, are the sets of even and odd permutations. (Assuming $n \geq 2$ ensures that these two cosets are both nonempty.) We conclude

$$|A_n| = |S_n|/[S_n : A_n] = n!/2.$$

Thus the alternating group A_n has order $n!/2$.

Consequences of Lagrange's Theorem

Last lecture we discussed Lagrange's Theorem: for any finite group G, and subgroup $H \subset G$, we have $[G : H] = |G|/|H|$. Recall here that $[G : H]$ is the number of right cosets of H in G. The most useful consequence of this theorem is the following:

▶ **If G is a finite group, and H is a subgroup of G, then $|H|$ divides $|G|$.**

This of course greatly constrains the possibilities for which subsets of G can be subgroups. A particular case is the following. Let $a \in G$ and consider the cyclic subgroup $\langle a \rangle \subset G$ generated by a. Recall that $\text{ord}(a)$ is equal to the size of this subgroup. We obtain:

▶ **If G is a finite group and $a \in G$ then $\text{ord}(a)$ divides $|G|$.**

For example, S_3 can only have elements of orders $\{1, 2, 3, 6\}$, and 6 does not occur because S_3 is not cyclic. In fact, we know all of this from direct computation. But now we understand more about why the orders of elements are constrained to these numbers.

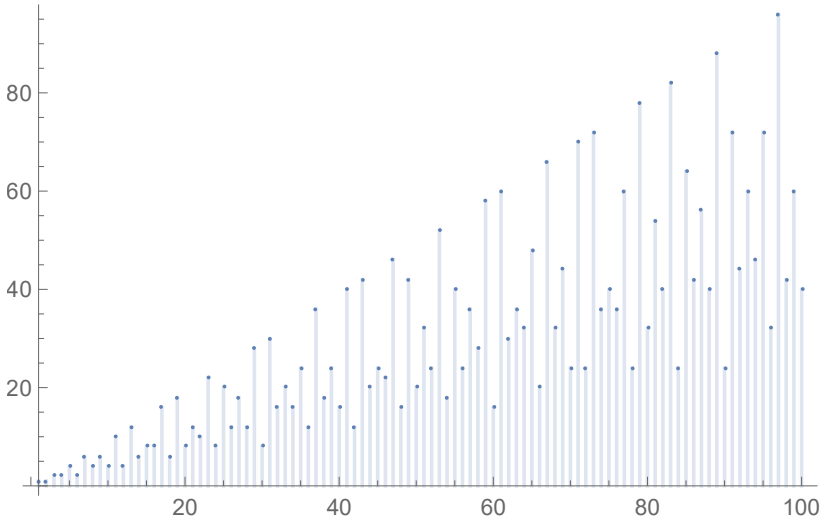
▶ **If G is a finite group and $a \in G$ then $a^{|G|} = e$.**

Indeed, writing $|G| = \text{ord}(a) \cdot n$, we have $a^{|G|} = a^{\text{ord}(a) \cdot n} = (a^{\text{ord}(a)})^n = e^n = e$, as claimed.

Next, we apply this last result to the group $(\mathbb{Z}_n^\times, \times)$ where n is a positive integer. Define

$$\phi(n) = |\mathbb{Z}_n^\times| = \#\{k \in \mathbb{Z} : 1 \leqslant k \leqslant n, \gcd(k, n) = 1\}$$

The function $\phi(n)$ is called *Euler's ϕ-function*, and sometimes *Euler's totient function*. For example, $\mathbb{Z}_7^\times = \{1, 2, 3, 4, 5, 6\}$ so $\phi(7) = 6$, while $\mathbb{Z}_{10}^\times = \{1, 3, 7, 9\}$ and so $\phi(10) = 4$. Below we show a graph of Euler's ϕ-function.



Euler's ϕ-function

▶ **(Euler's Theorem) For any integer k relatively prime to n, we have**

$$k^{\phi(n)} \equiv 1 \pmod{n}$$

This result follows from the previous one: just view $k \pmod{n}$ as an element of $\mathbb{Z}_n^\times$, and note that the order of the group is by definition $\phi(n)$.

For example, let $n = 30$. We list the integers from 1 to 30 which are relatively prime to 30:

$$\mathbb{Z}_{30}^\times = \{1, 7, 11, 13, 17, 19, 23, 29\}$$

Thus $\phi(30) = |\mathbb{Z}_{30}^\times| = 8$. Furthermore, Euler's Theorem tells us that for any one of the above 8 integers k (and their congruence classes mod 30) we have $k^8 \equiv 1 \pmod{30}$.

A special case of Euler's Theorem is when n is a prime number p. For in this case we have

$$\mathbb{Z}_p^\times = \{1, 2, \cdots, p-1\}$$

so in particular $\phi(p) = p - 1$. Therefore we obtain:

▶ **(Fermat's Little Theorem) For a prime p and integer k relatively prime to p:**

$$k^{p-1} \equiv 1 \pmod{p}$$

The conclusion of this result is often written as $k^p \equiv k \pmod{p}$.

For example, 97 is a prime number. Let's compute $5^{99} \pmod{97}$. Fermat's Little Theorem tells us that $5^{96} \equiv 1 \pmod{97}$. Using this we compute:

$$5^{99} \equiv 5^{96+3} \equiv 5^{96}5^3 \equiv 1 \cdot 5^3 \equiv 125 \equiv 28 \pmod{97}$$

Without the help of Fermat's Little Theorem, this would have taken much longer!

Another important consequence of Lagrange's Theorem is the following.

▶ **Suppose G is a finite group of prime order. Then G is cyclic.**

Let $H \subset G$ be a subgroup of G. Then Lagrange's Theorem tells us that $|H|$ divides $|G|$. Since $|G|$ is prime, $|H|$ must be 1 or $|G|$. In the first case, we must have $H = \{e\}$, and in the latter case, $H = G$. In particular, G has no non-trivial proper subgroups. Let $a \in G$ be a non-identity element. Then $\langle a \rangle$ is a non-trivial subgroup and thus must be all of G. In particular, $G = \langle a \rangle$ and so G is cyclic and generated by a.

We make two important remarks about Lagrange's Theorem. First, we could have used the notion of a *left* coset instead of a right coset: these are subsets $aH = \{ah : h \in H\}$. Lagrange's Theorem holds for left cosets, by the same arguments. A consequence is that the number of left cosets is equal to $[G : H]$, the number of right cosets.

Second, the converse to Lagrange's Theorem is false: if a positive integer d divides $|G|$, then it is not necessarily true that there is a subgroup of order d within G. The first instance of this phenomenon is the following:

▶ **In the alternating group A_4 of order 12, there is no subgroup of order 6.**

Let us prove this. First we write out the 12 elements of A_4:

$$A_4 = \{e, (123), (132), (124), (142), (134), (143), (234), (243), (12)(34), (13)(24), (14)(23)\}$$

Note we have 8 cycles of length 3, which have order 3, and 3 elements which are pairs of disjoint transpositions, each of order 2. Now suppose there is a subgroup $H \subset A_4$ of order 6. Let $\sigma \in A_4$ be a cycle of length 3. Consider the right cosets

$$H, \quad H\sigma, \quad H\sigma^2$$

Lagrange's Theorem tells us that $[A_4 : H] = |A_4|/|H| = 12/6 = 2$, so there are exactly 2 right cosets. So two of the cosets above must be equal. If $H = H\sigma$, then $\sigma \in H$, and similarly if $H = H\sigma^2$ then $\sigma^2 \in H$. But since $\sigma^2 = \sigma^{-1}$ and H is a subgroup, we must have $\sigma \in H$. The other possibility is that $H\sigma = H\sigma^2$. Multiplying on the right by σ gives $H\sigma^2 = H$, and again we conclude $\sigma \in H$. In conclusion, every length 3 cycle in A_4 must be in H. But there are 8 such cycles. Thus $6 = |H| \geqslant 8$, which is a contradiction. Thus A_4 cannot have a subgroup of order 6, as we claimed.

Applications to cryptography

In this lecture we apply some of the theory we have developed to Cryptography, which is the subject of sending secret messages. Our main application is to describe the RSA cryptosystem, a widely used method for secure data transmission.

We begin with a simple example of a *private key cryptosystem*, which most people are familiar with in some form. Let us write each letter from A to Z as a number in $\mathbb{Z}_{26}$. For convenience we also write each number with 2 digits. We have the following correspondence:

A	B	C	D	E	F	G	H	I	J	K	L	M
00	01	02	03	04	05	06	07	08	09	10	11	12

N	O	P	Q	R	S	T	U	V	W	X	Y	Z
13	14	15	16	17	18	19	20	21	22	23	24	25

Now, suppose we have a message we would like to send in secret: ALGEBRA. We first write this in numbers as 00110604011700, by replacing A with 00, and so on. Now we encrypt the message using our private encryption key: the function $f : \mathbb{Z}_{26} \to \mathbb{Z}_{26}$ given by $f(k) = k + 10 \pmod{26}$, say. Thus 00 gets sent to $f(0) = 0 + 10 \equiv 10 \pmod{26}$, 11 is sent to $f(11) = 11 + 10 \equiv 21 \pmod{26}$, and so on. Our message 00110604011700 transforms to

$$10211614110110$$

We may write this in letters as KVQOLBK. We send the message in this form to a friend. Once our friend receives the message, to decode it they only need apply f^{-1}, i.e. the function on $\mathbb{Z}_{26}$ which is $k \pmod{26} \mapsto k - 10 \pmod{26}$.

The above cryptosystem, which is ancient and used for example by Caesar, is called a *private key* cryptosystem because the key (in this case the function f or its inverse) must be kept private for the system to work effectively. This scheme also has the property that anyone who has the key, i.e. knowledge of the function f, can easily encode and also decode messages. This is because the function f^{-1} is easy to find if f is known, and vice versa. In particular, anyone who can encode messages can also decode messages.

We now describe a more sophisticated cryptosystem which is used in the modern world and relies on some of the algebra we have developed thus far. It is called *RSA*, named after Rivest, Shamir, and Adleman, who described the method in 1978. The algorithm was actually discovered earlier in 1973 by Clifford Cocks working for British intelligence, but this information was only declassified in 1997.

To begin, choose two large distinct prime numbers at random, say p and q. These numbers are meant to be kept secret from the public. Let $n = pq$ and $N = (p-1)(q-1)$. Choose an integer e which satisfies $1 < e < N$ and $\gcd(e, N) = 1$. We know that there exists an integer d that satisfies $ed \equiv 1 \pmod{N}$, and we compute this. The scheme is as follows:

Public information (encryption key): the number $n = pq$ and integer e.

Private information (decryption key): the prime numbers p, q and the integer d.

Note that although $n = pq$ is known to the public, the factorization of n into the prime numbers p and q is private information. The robustness of RSA is based on the principle that factoring numbers is hard!

Now we describe the simplest version of the algorithm. Suppose someone from the public, "Bob", wants to send an encrypted message to "Alice", who holds the private information of the decryption key. Bob knows only the information n, e while Alice knows p, q and d. Assume Bob's message is in the format of a number $m \in \mathbb{Z}_n^\times$. The algorithm is as follows:

1. Bob encodes his message $m \in \mathbb{Z}_n^\times$ by computing $c = m^e \pmod{n}$.
2. Bob sends his encrypted message $c \in \mathbb{Z}_n^\times$ to Alice.
3. Alice then uses d to compute $c^d \pmod{n}$, which recovers $m \in \mathbb{Z}_n^\times$.

Why does this work? First, we should check that $c^d \equiv m \pmod{n}$, i.e. that Alice's decryption actually recovers the original message. We need the following:

▶ **Let a, b be relatively prime. Then $\phi(ab) = \phi(a)\phi(b)$.**

To see this, recall that $\phi(a) = |\mathbb{Z}_a^\times|$. We define a map from $\mathbb{Z}_{ab}^\times$ to the Cartesian product $\mathbb{Z}_a^\times \times \mathbb{Z}_b^\times$ be sending $k \pmod{ab}$ to the pair $(k \pmod{a}, k \pmod{b})$. This is well-defined: if k is relatively prime to ab it is then certainly relatively prime to each of a and b. It is a 1-1 map: if $k, k' \in \mathbb{Z}_{ab}^\times$ are sent to the same pairs, then $k \equiv k' \pmod{a}$ and $k \equiv k' \pmod{b}$, and so $k - k'$ is divisible by both a and b. Since $\gcd(a, b) = 1$, it follows that $k - k'$ is divisible by ab, and so $k \equiv k' \pmod{ab}$. It is onto: let r, s be such that $ar \equiv 1 \pmod{b}$ and $bs \equiv 1 \pmod{a}$. Then for any pair $(k_1 \pmod{a}, k_2 \pmod{b}) \in \mathbb{Z}_a^\times \times \mathbb{Z}_b^\times$ consider

$$k = k_1 bs + k_2 ar$$

This gives an element of $\mathbb{Z}_{ab}^\times$ and it reduces mod a to k_1, and reduces mod b to k_2. Thus the map is onto. We have thus produced a 1-1 onto map from $\mathbb{Z}_{ab}^\times$ to $\mathbb{Z}_a^\times \times \mathbb{Z}_b^\times$. In particular these sets have the same size. Thus $\phi(ab) = |\mathbb{Z}_{ab}^\times| = |\mathbb{Z}_a^\times \times \mathbb{Z}_b^\times| = \phi(a)\phi(b)$.

Now return to the RSA algorithm. Note that $N = (p-1)(q-1) = \phi(p)\phi(q) = \phi(pq) = \phi(n)$. Recall that Alice has Bob's encrypted message $c \in \mathbb{Z}_n^\times$ and she computed $c^d \in \mathbb{Z}_n^\times$. Since $ed \equiv 1 \pmod{N}$ we have $ed - 1 = kN$ for some $k \in \mathbb{Z}$. Then

$$c^d \equiv (m^e)^d \equiv m^{ed} \equiv m^{1+kN} \equiv m \cdot (m^N)^k \equiv m \cdot (m^{\phi(n)})^k \equiv m \cdot 1 \equiv m \pmod{n}$$

which is of course Bob's original message $m \in \mathbb{Z}_n^\times$. Note that in this computation we used the fact that $m^{\phi(n)} \equiv 1 \pmod{n}$, which is Euler's Theorem!

Example: Suppose $p = 43$ and $q = 97$. Then $n = pq = (43)(97) = 4171$. Note here $N = \phi(n) = (43-1)(97-1) = 4032$. Alice chooses the encryption number $e = 13$, which is relatively prime to $N = 4032$. Alice releases the encryption key information

$$n = 4171, \qquad e = 13$$

to the public. However, only she knows $p = 43$ and $q = 97$. Alice needs to compute her private decryption key: this is a number d such that $ed \equiv 1 \pmod{4171}$. She computes $d = 1861$ using the Euclidean algorithm.

Next, Bob wants to send the message $m = 3161$ to Alice. He wants to compute $m^e \equiv 3161^{13} \pmod{4171}$. To do this he first writes $13 = 2^3 + 2^2 + 1$ and then

$$c \equiv m^e \equiv 3161^{13} = 3161^{(2^3+2^2+1)} \equiv ((3161^2)^2)^2 \cdot (3161^2)^2 \cdot 3161 \pmod{4171}$$

Then $3161^2 = 9991921 \equiv 2376 \pmod{4171}$. Next, $(3161^2)^2 \equiv 2376^2 = 5645376 \equiv 2013 \pmod{4171}$, and $((3161^2)^2)^2 \equiv 2013^2 = 4052169 \equiv 2128 \pmod{4171}$. He finally computes

$$c \equiv m^e = (2128)(2013)(3161) \equiv 2582 \pmod{4171}$$

Thus $c = 2582$ is the encrypted message, and Bob sends this to Alice. She can then decipher the message by using $d = 1861$ to compute $2582^d \equiv m \pmod{4171}$. To do this she can write 1861 in binary as $1861 = 2^{10} + 2^9 + 2^8 + 2^6 + 2^2 + 1$ and proceed as above:

$$2582^{2^1} \equiv 1466, \quad 2582^{2^2} \equiv 1091, \quad 2582^{2^3} \equiv 1456, \quad 2582^{2^4} \equiv 133, \quad 2582^{2^5} \equiv 1005,$$

$$2582^{2^6} \equiv 643, \quad 2582^{2^7} \equiv 520, \quad 2582^{2^8} \equiv 4356, \quad 2582^{2^9} \equiv 2363, \quad 2582^{2^{10}} \equiv 2971.$$

Thus $c^d \equiv 2582^{1861} \equiv (2971)(2363)(3456)(643)(1091)(2582) \equiv 3161 \pmod{4171}$, which is exactly Bob's original message m.

How can someone intercept the message and decipher it? They would have to know $d \pmod{N}$. You can compute $d \pmod{N}$ from the knowledge of e and N using the Euclidean algorithm, and of course $N = (p-1)(q-1)$. So knowing the prime numbers p and q will allow for anyone to decipher encrypted messages. However, the public only knows e and $n = pq$. And it is very difficult to factor large numbers! So if p and q (and hence n) are very large, it will be very difficult for people to decipher intercepted messages.

Normal subgroups and quotient groups

In this lecture we discuss an important class of subgroups called *normal subgroups*. Given a normal subgroup N in a group G, we then construct the *quotient group* G/N. The construction is a generalization of our construction of the groups $(\mathbb{Z}_n, +)$.

Fix a group G and a subgroup H. If we have a Cayley table for G, then it is easy to find the right and left cosets of H in G. Let us illustrate this with an example we have encountered before. Consider the group S_3, and choose the subgroup $H = \{e, (12)\}$. To find the right cosets of H in the Cayley table, first find H in the left column. It's shaded gray below.

	e	(12)	(23)	(31)	(132)	(123)
e	e	(12)	(23)	(31)	(132)	(123)
(12)	(12)	e	(123)	(132)	(31)	(23)
(23)	(23)	(132)	e	(123)	(12)	(31)
(31)	(31)	(123)	(132)	e	(23)	(12)
(132)	(132)	(23)	(31)	(12)	(123)	e
(123)	(123)	(31)	(12)	(23)	e	(132)

Each column, restricted to the rows determined by H, gives a right coset. In our example above, each such column has 2 elements. Equal cosets are given the same color. We see in our example we have 3 distinct right cosets, compatible with Lagrange's Theorem.

The description of left cosets from the Cayley table is similar, except the roles of column and row are switched. We illustrate this for our example:

	e	(12)	(23)	(31)	(132)	(123)
e	e	(12)	(23)	(31)	(132)	(123)
(12)	(12)	e	(123)	(132)	(31)	(23)
(23)	(23)	(132)	e	(123)	(12)	(31)
(31)	(31)	(123)	(132)	e	(23)	(12)
(132)	(132)	(23)	(31)	(12)	(123)	e
(123)	(123)	(31)	(12)	(23)	e	(132)

We see in this example that the right cosets and left cosets are not the same.

For another example, consider the subgroup $H = \{e, (123), (132)\}$ in S_3. The same procedure above carried out for H shows that in this case, the right cosets and left cosets are actually the same. More precisely, $aH = Ha$ for each group element a. This is illustrated in the tables below, where right and left cosets are shown.

	e	(12)	(23)	(31)	(132)	(123)
e	e	(12)	(23)	(31)	(132)	(123)
(12)	(12)	e	(123)	(132)	(31)	(23)
(23)	(23)	(132)	e	(123)	(12)	(31)
(31)	(31)	(123)	(132)	e	(23)	(12)
(132)	(132)	(23)	(31)	(12)	(123)	e
(123)	(123)	(31)	(12)	(23)	e	(132)

	e	(12)	(23)	(31)	(132)	(123)
e	e	(12)	(23)	(31)	(132)	(123)
(12)	(12)	e	(123)	(132)	(31)	(23)
(23)	(23)	(132)	e	(123)	(12)	(31)
(31)	(31)	(123)	(132)	e	(23)	(12)
(132)	(132)	(23)	(31)	(12)	(123)	e
(123)	(123)	(31)	(12)	(23)	e	(132)

This latter kind of subgroup is called a *normal subgroup* and is defined as follows.

▶ **A subgroup N in a group G is *normal* if for all $a \in G$ we have $aN = Na$.**

The first example $H = \{e, (12)\}$ in S_3 from above is not normal, while the second example $H = \{e, (123), (132)\} \subset S_3$ is normal. Note that if G is abelian, every subgroup is normal. The following gives alternative characterizations for a subgroup to be normal.

▶ **Let G be a group and $N \subset G$ a subgroup. The following are equivalent:**

(i) N **is a normal subgroup.**

(ii) $aNa^{-1} \subset N$ **for all** $a \in G$**.**

(iii) $aNa^{-1} = N$ **for all** $a \in G$**.**

Proof. (i)⇒(ii): Assume N is normal, i.e. $aN = Na$ for all $a \in G$. Let $ana^{-1} \in aNa^{-1}$ where $n \in N$. Since $aN = Na$, we have $an \in Na$, i.e. $an = n'a$ for some $n' \in N$. Then $ana^{-1} = n'aa^{-1} = n' \in N$. Thus $aNa^{-1} \subset N$.

(ii)⇒(iii): Assume $aNa^{-1} \subset N$ for all $a \in G$. Let $n \in N$. Then (ii) with a^{-1} in place of a says $a^{-1}Na \subset N$. (Note (ii) holds for *all* $a \in G$.) Thus $a^{-1}na = n'$ for some $n' \in N$. Then $n = an'a^{-1} \in aNa^{-1}$, so $N \subset aNa^{-1}$. It follows that $N = aNa^{-1}$.

(iii)⇒(i): Assume $aNa^{-1} = N$ for all $a \in G$. Let $an \in aN$ where $n \in N$. Our assumption implies $ana^{-1} = n'$ for some n'. Then $an = n'a$. Thus $aN \subset Na$. Similar reasoning shows $Na \subset aN$, and thus $aN = Na$. □

Now fix a normal subgroup N within any group G. Define G/N to be:

$$G/N = \{\text{right cosets of } N \text{ in } G\} = \{\text{left cosets of } N \text{ in } G\}$$

We define a binary operation on G/N as follows: for $aN, bN \in G/N$ we set

$$(aN)(bN) = abN$$

Of course we could have also written $(Na)(Nb) = Nab$, and the fact that it doesn't matter if we use right or left cosets is thanks to the normality of N.

▶ **Let G be a group and $N \subset G$ a normal subgroup. Then the set of cosets G/N endowed with the above binary operation satisfies the axioms of a group.**

To see this we first check that the binary operation is well-defined. More precisely, suppose $aN = a'N$ and $bN = b'N$. Then $a = a'n_1$ and $b = b'n_2$ for some $n_1, n_2 \in N$. Since $b'N = Nb'$ we have $n_1 b' = b'n_3$ for some $n_3 \in N$. We compute

$$ab = (a'n_1)(b'n_2) = a'(n_1 b')n_2 = a'(b'n_3)n_2 = a'b'(n_3 n_2)$$

Thus $ab \equiv a'b' \pmod{N}$, or in other words $abN = a'b'N$.

Next, it is easy to see that $N = eN = Ne$ is an identity. Associativity follows because $(aNbN)cN = (abN)cN = abcN = aN(bcN) = aN(bNcN)$. Finally, the inverse of aN is given by the coset $a^{-1}N$. Thus G/N is a group.

We call the group G/N from the above construction the *quotient group* of G by N. The construction crucially depends on N being a normal subgroup, and does not work for subgroups that are not normal.

▶ **The order of the group G/N is equal to $[G : N]$.**

This is just from the definitions. Note that if G is finite, Lagrange's Theorem implies

$$|G/N| = [G : N] = |G|/|N|$$

Examples

1. For the normal subgroup $N = \{e, (123), (132)\} \subset S_3$ we have $|S_3/N| = |S_3|/|N| = 6/3 = 2$. Thus the quotient group is of order 2. Note $S_3/N = \{N, (12)N\}$.

2. Let $G = (\mathbb{Z}, +)$ and consider the subgroup $N = n\mathbb{Z} \subset G$ for any positive integer n. Then N is a normal subgroup, as G is abelian, and G/N is exactly the group $(\mathbb{Z}_n, +)$.

3. Consider the alternating group A_n as a subgroup of S_n. Using the properties of even and odd permutations, we have that A_n is a normal subgroup. Then $|S_n/A_n| = 2$ and $S_n/A_n = \{E, O\}$ where $E = \{\text{even permutations}\}$, $O = \{\text{odd permutations}\}$.

Homomorphisms and Isomorphisms

When are two groups "the same"? We have implicitly answered this question in our earlier investigations. Namely, we have seen several different incarnations of a certain non-abelian group of order 6. First, we abstractly defined a group

$$G = \{e, r, b, g, y, o\}$$

in Lecture 1, by writing out its Cayley table. We saw later that the group formed by the symmetries of an equilateral triangle, and also the symmetric group

$$S_3 = \{e, (12), (23), (31), (123), (132)\}$$

have the same Cayley table when the elements are matched up appropriately. The Cayley tables of G and S_3 are shown below, and we see that they are essentially the same.

	e	r	b	g	y	o
e	e	r	b	g	y	o
r	r	e	o	y	g	b
b	b	y	e	o	r	g
g	g	o	y	e	b	r
y	y	b	g	r	o	e
o	o	g	r	b	e	y

	e	(12)	(23)	(31)	(132)	(123)
e	e	(12)	(23)	(31)	(132)	(123)
(12)	(12)	e	(123)	(132)	(31)	(23)
(23)	(23)	(132)	e	(123)	(12)	(31)
(31)	(31)	(123)	(132)	e	(23)	(12)
(132)	(132)	(23)	(31)	(12)	(123)	e
(123)	(123)	(31)	(12)	(23)	e	(132)

Let us formalize what we mean by *the same*. What we really have is a map $\phi : G \to S_3$, that is, an assignment to each element of G some element of S_3. This assignment is a 1-to-1 and onto map, which means that every element of S_3 is mapped to by something in G, and that no two elements of G are mapped to the same element of S_3. In our example this map is:

$$\phi(e) = e, \quad \phi(r) = (12), \quad \phi(b) = (23), \quad \phi(g) = (31), \quad \phi(y) = (132), \quad \phi(o) = (123)$$

Furthermore, this map preserves the group structures, or equivalently, it maps the Cayley table of G in the correct way to the Cayley table of S_3. What this really means is that for any $a_1, a_2 \in G$ the product $a_1a_2 \in G$ is mapped to the product $\phi(a_1)\phi(a_2) \in S_3$. For example,

$$(123) = \phi(o) = \phi(rb) = \phi(r)\phi(b) = (12)(23) = (123)$$

We formalize the kind of map that is appearing here into a definition.

▶ **A *homomorphism* from a group G to a group G' is a map $\phi : G \to G'$ such that**

$$\phi(ab) = \phi(a)\phi(b) \qquad \textbf{for all } a, b \in G.$$

A homomorphism ϕ is an *isomorphism* if the map ϕ is 1-to-1 and onto.

Therefore, two groups are the "same" if there is an isomorphism $\phi : G \to G'$.

The only difference between a homomorphism and an isomorphism is that the map ϕ does not need to be 1-to-1 and onto for a homomorphism. Every isomorphism is of course a homomorphism, and so for the moment we will develop some properties of this more general notion.

Examples

1. Let G and G' be groups. Define $\phi : G \to G'$ defined by $\phi(a) = e'$ for all $a \in G$, where e' is the identity of G'. For any $a, b \in G$, we have $\phi(ab) = e' = e'e' = \phi(a)\phi(b)$, and so ϕ is a homomorphism. This is the *trivial* homomorphism from G to G'.

2. Consider the groups $\mathrm{GL}_2(\mathbb{R})$ and $(\mathbb{R}^\times, \times)$. Define a map $\phi : \mathrm{GL}_2(\mathbb{R}) \to \mathbb{R}^\times$ by setting $\phi(A) = \det(A)$ for each matrix $A \in \mathrm{GL}_2(\mathbb{R})$. Then

$$\phi(A)\phi(B) = \det(A)\det(B) = \det(AB) = \phi(AB)$$

Thus ϕ is a homomorphism.

3. The map $\phi : (\mathbb{R}, +) \to (\mathbb{R}^\times, \times)$ defined by $\phi(a) = |a| + 1$ is not a group homomorphism. For example, the following is a failure of the homomorphism property:

$$3 = |2| + 1 = \phi(2) = \phi(1 + 1) \neq \phi(1)\phi(1) = (|1| + 1)(|1| + 1) = 4.$$

4. Define a map $\phi : (\mathbb{R}, +) \to (\mathbb{R}^\times, \times)$ by $\phi(x) = 2^x$ for each $x \in \mathbb{R}$. Then

$$\phi(x + y) = 2^{x+y} = 2^x 2^y = \phi(x)\phi(y)$$

for all $x, y \in \mathbb{R}$. Therefore ϕ is a homomorphism of groups.

5. Let G be a subgroup and $H \subset G$ a subgroup. Define $\phi : H \to G$ by $\phi(a) = a$ for all $a \in H$. In other words, ϕ is the "inclusion map" which simply sends H to itself viewed as a subset of G. Clearly ϕ is a homomorphism. For example, we have a homomorphism $(\mathbb{Z}, +) \to (\mathbb{Q}, +)$, and a homomorphism $(\mathbb{Q}^\times, \times) \to (\mathbb{R}^\times, \times)$.

6. The map $\phi : \mathrm{GL}_2(\mathbb{R}) \to \mathrm{GL}_2(\mathbb{R})$ defined by $\phi(A) = A^2$ is not a homomorphism.

7. Consider the symmetric group S_n, and the additive group $\mathbb{Z}_2$. Define $\phi : S_n \to \mathbb{Z}_2$ by:

$$\phi(\sigma) = \begin{cases} 0 \pmod 2 & \text{if } \sigma \text{ is even} \\ 1 \pmod 2 & \text{if } \sigma \text{ is odd} \end{cases}$$

Then ϕ is a homomorphism, thanks to our understanding of how even and odd permutations compose. For example, if σ is even and σ' is odd, we know $\sigma\sigma'$ is odd, and thus

$$\phi(\sigma\sigma') = 1 = 0 + 1 = \phi(\sigma) + \sigma(\sigma')$$

8. The map $\phi : (\mathbb{Z}, +) \to \mathrm{GL}_2(\mathbb{R})$ defined by $\phi(n) = \begin{pmatrix} 1 & n \\ 0 & 1 \end{pmatrix}$ is a homomorphism.

The following gives another source of many group homomorphisms, and is related to our construction of quotient groups from last lecture.

▶ **Let G be a group and $N \subset G$ a normal subgroup. Define $\phi : G \to G/N$ by $\phi(a) = aN$ for all $a \in G$. Then ϕ is a group homomorphism which is onto.**

To see this, we simply compute $\phi(a)\phi(b) = (aN)(bN) = abN = \phi(ab)$. This shows that ϕ is a homomorphism. The homomorphism ϕ is onto because given a coset $aN \in G/N$ the element $a \in G$ satisfies $\phi(a) = aN$.

As an example, we have onto homomorphisms $\mathbb{Z} \to \mathbb{Z}_n$ for every positive integer n, defined by sending an integer k to its congruence class $k \pmod{n}$.

▶ **Let $\phi : G \to G'$ be a homomorphism of groups. Then ϕ maps the identity $e \in G$ to the identity $e' \in G'$. Further, for all $a \in G$ we have $\phi(a^{-1}) = \phi(a)^{-1}$.**

To prove the first part, we compute

$$\phi(e) = \phi(ee) = \phi(e)\phi(e)$$

Multiply both sides by $\phi(e)^{-1}$ to obtain $e' = \phi(e)$. Next, for $a \in G$ we compute

$$\phi(a)\phi(a^{-1}) = \phi(aa^{-1}) = \phi(e) = e'$$

and similarly $\phi(a^{-1})\phi(a) = e'$. Thus the inverse of $\phi(a) \in G'$ is given by $\phi(a^{-1})$.

Complex numbers

The use of complex numbers historically emerged as a way of dealing with the problem that some simple polynomial equations, such as

$$x^2 + 1 = 0,$$

have no solutions in the real numbers. It then became apparent over time that complex numbers are much more than just an invention of this necessity.

Formally, a complex number is a pair $(a, b) \in \mathbb{R} \times \mathbb{R}$, which we often write as $a + ib$:

$$\mathbb{C} = \{a + ib : \quad a, b \in \mathbb{R}\}$$

The symbol i satisfies $i^2 = -1$, and we write $i = \sqrt{-1}$. The operations $+$ and $\times$ on $\mathbb{C}$ are:

$$(a + bi) + (c + di) = (a + c) + (b + d)i$$

$$(a + bi) \times (c + di) = (ac - bd) + (ad + bc)i$$

As usual we often omit the symbol $\times$. These operations satisfy the usual distributive law. Following earlier notation we write $\mathbb{C}^\times$ for the non-zero complex numbers.

▶ **$(\mathbb{C}, +)$ and $(\mathbb{C}^\times, \times)$ are abelian groups.**

The claim for $(\mathbb{C}, +)$ is straightforward. For $(\mathbb{C}^\times, \times)$, associativity, identity ($e = 1$), and commutativity are clear. To show inverses exist, consider $z = a + bi \in \mathbb{C}^\times$. We write

$$z^{-1} = \frac{1}{z} = \frac{1}{a + bi} = \frac{1}{a + bi} \cdot \frac{a - bi}{a - bi} = \frac{a - bi}{a^2 + b^2} = \left(\frac{a}{a^2 + b^2}\right) + \left(-\frac{b}{a^2 + b^2}\right) i$$

This is an element of $\mathbb{C}^\times$, and is an inverse to z. A priori "$1/z$" (and the symbols in red) does not make sense, because we only defined multiplication above: but the expression on the right of the equation can be taken as definition for $1/z$ and justifies the notation.

▶ **The order of $i \in \mathbb{C}^\times$ is equal to 4.**

This is immediate from the computations $i^2 = -1$, $i^3 = i^2 \cdot i = -i$, $i^4 = (i^2)^2 = (-1)^2 = 1$.

One of the most important functions of complex numbers is the exponential function. This can be defined in several ways. One way is to use power series:

$$e^z = \sum_{n=0}^{\infty} \frac{z^n}{n!} = 1 + z + \frac{z^2}{2!} + \frac{z^3}{3!} + \cdots$$

If we plug $z = i\theta$ into this formula, where θ is a real number, we get:

$$\begin{aligned} e^{i\theta} &= 1 + (i\theta) + \frac{(i\theta)^2}{2!} + \frac{(i\theta)^3}{3!} + \cdots \\ &= \left(1 - \frac{\theta^2}{2!} + \frac{\theta^4}{4!} - \cdots\right) + i\left(\theta - \frac{\theta^3}{3!} + \frac{\theta^5}{5!} - \cdots\right) = \cos(\theta) + i\sin(\theta) \end{aligned}$$

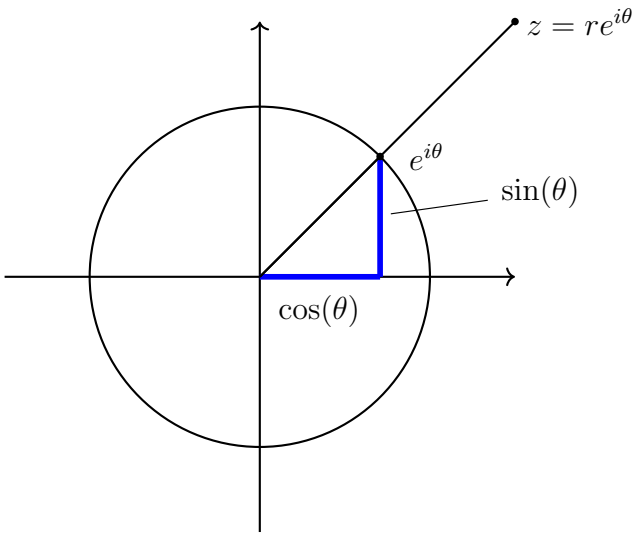
The remarkable formula $e^{i\theta} = \cos(\theta) + i\sin(\theta)$ is usually called *Euler's Formula*.

The *norm* of $z = a + bi \in \mathbb{C}$ is the non-negative real number defined by $|z| = \sqrt{a^2 + b^2}$. The *conjugate* of $z = a + bi$ is the complex number defined by $\overline{z} = a - bi$. We have

$$|z|^2 = a^2 + b^2 = z\overline{z},$$

which gives the expression $1/z = \overline{z}/|z|^2$, and expands to our previous expression for $1/z$. Note that $|e^{i\theta}| = \sqrt{\cos(\theta)^2 + \sin(\theta)^2} = 1$. A general property is that $|zz'| = |z||z'|$ for $z, z' \in \mathbb{C}$.

Every complex number z can be written in *polar form* $z = re^{i\theta}$ where r is a non-negative real number, and θ is some real number. Note $|z| = |re^{i\theta}| = |r||e^{i\theta}| = r$. The number $r = |z|$ is sometimes also called the *amplitude*, and θ the *phase* of z.



The *complex plane* is the usual xy-plane, where the point (x, y) represents the complex number $z = x + yi$. Above we show how in polar form, $z = re^{i\theta}$ is determined by locating the position of $e^{i\theta} = \cos(\theta) + i\sin(\theta)$ on the unit circle (it forms angle θ with the x-axis) and then scales according to its amplitude r.

The exponential law $e^{z+z'} = e^z e^{z'}$ holds for complex numbers $z, z' \in \mathbb{C}$. This can be proved in any number of ways. When $z = i\theta$ and $z' = i\theta'$ for real numbers θ, θ' this law is equivalent to the trigonometric identities for angle addition:

$$\begin{aligned} e^{i\theta}e^{i\theta'} &= (\cos(\theta) + i\sin(\theta))(\cos(\theta') + i\sin(\theta')) \\ &= (\cos(\theta)\cos(\theta') - \sin(\theta)\sin(\theta')) + i(\cos(\theta)\sin(\theta') + \sin(\theta)\cos(\theta')) \\ &= \cos(\theta + \theta') + i\sin(\theta + \theta') \\ &= e^{i(\theta+\theta')} \end{aligned}$$

The exponential law can be phrased in terms of group theory as follows:

▶ **Exponentiation gives an onto homomorphism** $(\mathbb{C}, +) \to (\mathbb{C}^\times, \times)$**.**

The geometric way of multiplying $z = re^{i\theta}$ and $z' = r'e^{i\theta'}$ now emerges: you multiply $e^{i\theta}$ and $e^{i\theta'}$ to get $e^{i\theta+i\theta'}$ (so the phases add) and then scale this by rr', the new amplitude.

We now give a group-theoretic interpretation of polar decomposition multiplication.

▶ **Let** $U(1) = \{z \in \mathbb{C} : |z| = 1\} = \{e^{i\theta} : \theta \in \mathbb{R}\}$**. Then with the operation of multiplication, this is a subgroup of** $(\mathbb{C}^\times, \times)$**, and is in particular an abelian group.**

This is straightforward to prove. The group $U(1)$ is called the *circle group*. Geometrically, it is the group of rotational symmetries of the complex plane which fix the origin. It is one of the most important and fundamental examples of a group. We have:

▶ **Exponentiation gives an onto homomorphism** $(\mathbb{R}, +) \to (U(1), \times)$**.**

For the next statement we let $\mathbb{R}^\times_{>0}$ be the subgroup of $(\mathbb{R}^\times, \times)$ containing the positive reals.

▶ **The map** $\phi : \mathbb{R}^\times_{>0} \times U(1) \to \mathbb{C}^\times$ **given by** $\phi(r, e^{i\theta}) = re^{i\theta}$ **is an isomorphism of groups.**

In other words, polar decomposition of complex numbers helps us understand how the group $\mathbb{C}^\times$ is really a product of two simpler groups we already understand.

To prove the statement, first we see that ϕ is 1-1 and onto: it is 1-1 because if $re^{i\theta} = r'e^{i\theta'}$ then $r = r'$ and $e^{i\theta} = e^{i\theta'}$ (here it is important that $r, r' \neq 0$). It is onto because every non-zero complex number has a polar representation. Finally, ϕ is a homomorphism because

$$\phi(r, e^{i\theta})\phi(r', e^{i\theta'}) = (re^{i\theta})(r'e^{i\theta'}) = rr'e^{i\theta}e^{i\theta'} = \phi(rr', e^{i\theta}e^{i\theta'}) = \phi((r, e^{i\theta})(r', e^{i\theta'}))$$

It follows that ϕ is an isomorphism as claimed.

More isomorphisms, and kernels

In this lecture we continue our study of homomorphisms and isomorphisms and also introduce the important notion of the *kernel* of an homomorphism.

Recall that an isomorphism is a 1-1 and onto homomorphism $\phi : G \to G'$. The homomorphism property says that $\phi(ab) = \phi(a)\phi(b)$ for all $a, b \in G$. Define a relation on groups: write $G \cong G'$ if there is an isomorphism $\phi : G \to G'$.

▶ **The relation $G \cong G'$ is an equivalence relation on the collection of all groups.**

In proving this we will introduce a number of important properties of isomorphisms.

First, given any group G, consider the map $\phi : G \to G$ given by $\phi(a) = a$, the "identity" homomorphism. This is clearly an isomorphism, so $G \cong G$, i.e. the relation is reflexive. The identity homomorphism is sometimes written $\mathrm{id}_G = \phi : G \to G$.

Next, suppose $G \cong G'$, i.e. there is an isomorphism $\phi : G \to G'$. Consider the inverse map of ϕ, denoted $\phi' : G' \to G$. This is defined as follows. Let $a' \in G'$. We would like to desribe $\phi'(a')$. Because ϕ is onto, there is an $a \in G$ with $\phi(a) = a'$, and this a is unique as ϕ is 1-1. Then we define $\phi'(a')$ by $\phi'(a') = a$. We see that ϕ' is characterized by the relation $\phi'(\phi(a)) = a$ for all $a \in G$. You can check that ϕ' is 1-1 and onto, and also $\phi(\phi'(a')) = a'$ for all $a' \in G'$. Next, given $a', b' \in G'$, let $a, b \in G$ such that $\phi(a) = a'$, $\phi(b) = b'$. Then

$$\phi'(a'b') = \phi'(\phi(a)\phi(b)) = \phi'(\phi(ab)) = ab = \phi'(a')\phi'(b')$$

Therefore $\phi' : G' \to G$ is an isomorphism, and so $G' \cong G$. The map ϕ' is called the *inverse* isomorphism of ϕ. The inverse isomorphism is sometimes written $\phi^{-1} : G' \to G$. This establishes symmetry of the relation $\cong$.

In general, if $\phi : G \to G'$ and $\phi' : G' \to G''$ are homomorphisms, the composition map $\phi' \circ \phi : G \to G''$ is also a homomorphism. Indeed, if $a, b \in G$ then we compute

$$(\phi' \circ \phi)(ab) = \phi'(\phi(ab)) = \phi'(\phi(a)\phi(b)) = \phi'(\phi(a))\phi'(\phi(b)) = (\phi' \circ \phi)(a)(\phi' \circ \phi)(b)$$

If $G \cong G'$ and $G' \cong G''$ by isomorphisms $\phi : G \to G'$ and $\phi' : G' \to G''$ then $\phi' \circ \phi$ is an isomorphism, as you may check. This establishes transitivity. Thus $\cong$ is an equivalence relation.

We say two groups G and G' are *isomorphic* if $G \cong G'$. A useful criterion for $\phi : G \to G'$ to be an isomorphism is as follows.

▶ **Let $\phi : G \to G'$ be a homomorphism. Then ϕ is an isomorphism if and only if there exists a homomorphism $\phi' : G' \to G$ such that**

$$\phi' \circ \phi = \mathrm{id}_G \qquad \text{and} \qquad \phi \circ \phi' = \mathrm{id}_{G'}$$

Indeed if ϕ is an isomorphism, then the inverse isomorphism considered above provides such a map ϕ'. Conversely, if ϕ is a homomorphism and such a ϕ' in the statement exists, let us check that ϕ is 1-1 and onto. First, let $a' \in G'$. Then $a' = (\phi \circ \phi')(a') = \phi(\phi'(a'))$. Thus ϕ is onto. Second, suppose $a, b \in G$ and $\phi(a) = \phi(b)$. Then applying ϕ' we get $a = \phi'(\phi(a)) = \phi'(\phi(b)) = b$ and thus ϕ is 1-1. We conclude that ϕ is an isomorphism.

In fact in the above statement, ϕ' is always uniquely determined as being the inverse isomorphism of ϕ that we considered earlier.

Most of the properties of groups that we are interested in are preserved under isomorphisms.

▶ **Let G and G' be isomorphic groups. Then:**

(i) If G is abelian, then G' is abelian.

(ii) If G is cyclic, then G' is cyclic.

(iii) If G is finite, then so is G' and $|G| = |G'|$.

The verification is left as an exercise.

Let $\phi : G \to G'$ be a homomorphism. We define the *kernel* of ϕ to be

$$\ker(\phi) = \{a \in G : \phi(a) = e'\} \subset G$$

▶ **Let $\phi : G \to G'$ be a homomorphism. Then $\ker(\phi)$ is a normal subgroup of G.**

To see this, first note $\phi(e) = e'$, so $e \in \ker(\phi)$. Next, let $a, b \in \ker(\phi)$. This means $\phi(a) = e'$ and $\phi(b) = e'$ where $e' \in G'$ is the identity. Then

$$\phi(ab) = \phi(a)\phi(b) = e'e' = e'$$

and so $ab \in \ker(\phi)$. Similarly, $\phi(a^{-1}) = \phi(a)^{-1} = (e')^{-1} = e'$ and so $a^{-1} \in \ker(\phi)$. Therefore $\ker(\phi)$ is a subgroup of G. Next, let $a \in \ker(\phi)$, and let $g \in G$ be any element. Then

$$\phi(gag^{-1}) = \phi(g)\phi(a)\phi(g^{-1}) = \phi(g)e'\phi(g)^{-1} = e'$$

This shows $g \cdot \ker(\phi) \cdot g^{-1} \subset \ker(\phi)$. Therefore $\ker(\phi)$ is normal.

The following is a sort of converse to the above result:

▶ **Let G be a group and $N \subset G$ a normal subgroup. Then there is a homomorphism ϕ from G to some other group such that $\ker(\phi) = N$.**

In fact we can take the homomorphism ϕ to be one that we have seen before:

$$\phi : G \to G/N$$

defined by $\phi(a) = aN$ for each $a \in G$. The kernel of this homomorphism is the set of $a \in G$ such that $\phi(a) = N$, i.e. $aN = N$. But these are exactly the a that are in the coset N. Thus $\ker(\phi) = N$. This proves what we claimed.

Finally, the following shows that kernels are useful for understanding isomorphisms.

▶ **An onto homomorphism $\phi : G \to G'$ is an isomorphism if and only if $\ker(\phi) = \{e\}$.**

First, suppose ϕ is an isomorphism. If $a \in \ker(\phi)$ then $\phi(a) = e'$. But $\phi(e) = e'$ also, so $\phi(e) = \phi(a)$ implies $e = a$, as ϕ is 1-1. This shows $\ker(\phi) = \{e\}$.

Conversely, suppose $\ker(\phi) = \{e\}$. We aim to show ϕ is 1-1. Suppose $\phi(a) = \phi(b)$. Then

$$e' = \phi(a)\phi(b)^{-1} = \phi(a)\phi(b^{-1}) = \phi(ab^{-1})$$

and in particular $ab^{-1} \in \ker(\phi)$. Since $\ker(\phi) = \{e\}$ we must have $ab^{-1} = e$, or equivalently $a = b$. Thus ϕ is 1-1. By assumption ϕ is onto, and thus it is an isomorphism.

First isomorphism theorem, and symmetries of a cube

In this lecture we prove the "1st Isomorphism Theorem" and discuss some consequences. We also explore the rotational symmetries of a cube in 3-dimensional Euclidean space.

▶ **1st Isomorphism Theorem**: **Let $\phi : G \to G'$ be an *onto* homomorphism. Then there is a naturally induced map which we write as**

$$\psi : G/\mathbf{ker}(\phi) \longrightarrow G'$$

and this map ψ is an isomorphism of groups.

Proof. Write $N = \ker(\phi)$. Define the map ψ as follows: for any coset $aN \in G/N$ we let $\psi(aN) = \phi(a)$. Let us check this is well-defined. Suppose $aN = bN$. This means $ab^{-1} \in N$, i.e. $e' = \phi(ab^{-1}) = \phi(a)\phi(b)^{-1}$. Thus $\phi(a) = \phi(b)$. In particular,

$$\phi(a) = \psi(aN) = \psi(bN) = \phi(b)$$

This tells us ψ is well-defined map, independent of how the coset aN is written.

Next, we check ψ is a homomorphism. For $aN, bN \in G/N$ we simply compute

$$\psi(aN)\psi(bN) = \phi(a)\phi(b) = \phi(ab) = \psi(abN) = \psi(aNbN)$$

and thus ψ is a homomorphism.

Finally, we check that ψ is 1-1 and onto. Let $a' \in G'$. Because ϕ is onto, there is some $a \in G$ such that $\phi(a) = a'$. Then also $\psi(aN) = \phi(a) = a'$. Therefore ψ is onto. Finally, suppose $aN, bN \in G/N$ are such that $\psi(aN) = \psi(bN)$. This implies $\phi(a) = \phi(b)$, or $\phi(ab^{-1}) = e'$, implying $ab^{-1} \in N$. In particular, $aN = bN$. Thus ψ is 1-1, and ψ is an isomorphism. □

Examples

1. Let $\phi : \mathbb{Z}_{20} \to \mathbb{Z}_5$ be the homomorphism given by $\phi(k \pmod{20}) = k \pmod 5$. Then $\ker(\phi) = \langle 5 \rangle \subset \mathbb{Z}_{20}$. The 1st Isomorphism Theorem gives

$$\mathbb{Z}_{20}/\langle 5 \rangle \cong \mathbb{Z}_5$$

2. Consider the exponential map $\phi : (\mathbb{R}, +) \to (U(1), \times)$, where $U(1) \subset \mathbb{C}^\times$ is the circle group, defined by $\phi(\theta) = e^{2\pi i\theta}$. This is an onto homomorphism. The kernel is

$$\ker(\phi) = \{\theta \in \mathbb{R} : e^{2\pi i\theta} = 1\} = \mathbb{Z} \subset \mathbb{R}$$

By the 1st isomorphism theorem we conclude we have an isomorphism

$$\mathbb{R}/\mathbb{Z} \cong U(1)$$

3. Consider the group of upper triangular matrices in $\mathrm{GL}_2(\mathbb{R})$ given by

$$G = \left\{ A = \begin{pmatrix} a & b \\ 0 & c \end{pmatrix} : \quad a, b, c \in \mathbb{R}, \quad ac \neq 0 \right\}$$

Define a map $\phi : G \to \mathbb{R}^\times \times \mathbb{R}^\times$ by $\phi(A) = (a, c)$. This is easily checked to be an onto homomorphism. The kernel of ϕ is given by the subgroup $H \subset G$ of upper triangular matrices with $a = c = 1$. Thus H is normal. The 1st Isomorphism Theorem gives

$$G/H \cong \mathbb{R}^\times \times \mathbb{R}^\times$$

In particular, G/H is abelian.

▶ **Let G be cyclic. If $|G| = \infty$ then $G \cong \mathbb{Z}$. If $|G|$ is finite then $G \cong \mathbb{Z}_n$ where $n = |G|$.**

Proof. Since G is cyclic, $G = \langle a \rangle = \{a^k : k \in \mathbb{Z}\}$ for some $a \in G$. Define a map

$$\phi : \mathbb{Z} \longrightarrow G$$

by setting $\phi(k) = a^k$. Then this is onto, since a generates G. The kernel of ϕ consists of $m \in \mathbb{Z}$ such that $a^m = e$ in G. There are two cases. If G is finite, then $|G| = n$ is the order of a and $\ker(\phi) = n\mathbb{Z} \subset \mathbb{Z}$. The 1st Isomorphism Theorem gives

$$\mathbb{Z}_n = \mathbb{Z}/n\mathbb{Z} = \mathbb{Z}/\ker(\phi) \cong G$$

In the other case, G is infinite, and there are no m aside from $m = 0$ such that $a^m = e$, and thus the kernel of ϕ is trivial. Then $G \cong \mathbb{Z}$. □

Symmetries of a cube

Now let us turn to the group G which consists of the rotational symmetries of a cube situated in 3-dimensional Euclidean space. This group has 24 elements, as shown on the next page. We first follow a familiar strategy of describing this group: label the vertices 1 to 8, and associate to each rotation $a \in G$ a corresponding permutation $\phi(a) \in S_8$ based on how the vertices are moved around. This gives a homomorphism

$$\phi : G \longrightarrow S_8$$

This homomorphism ϕ is 1-1, but it is not onto: indeed, $|G| = 24$ but the target group S_8 has $8! = 40320$ elements.

Mapping G into S_8 is not the most ideal scenario. After all, $|S_8| = 40320$ whereas G has order only 24. A better way of representing G as a group of permutations is as follows.

There are 4 diagonal axes that pass through the cube; each one goes through two vertices that are as far apart as possible. Label these 4 diagonal axes 1, 2, 3, 4. Then for a rotation $a \in G$ we define $\psi(a) \in S_4$ to be the permutation determined by how the diagonal axes are moved around by the rotation a. This gives a homomomorphism

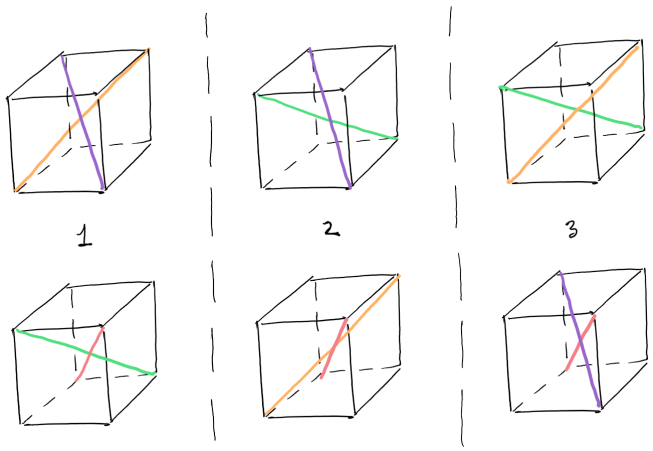
$$\psi : G \longrightarrow S_4$$

It is straightforward to verify that ψ is 1-1, i.e. that a rotation of the cube is entirely determined by how these 4 diagonal axes are permuted. Then, since $|G| = 24 = |S_4|$, we know ψ must also be onto. Therefore ψ is an isomorphism! The construction of the map ψ is illustrated on the next page.

Finally, let us return to the 1st Isomorphism Theorem. To this end, we define a map

$$\mu : S_4 \longrightarrow S_3$$

as follows. Let $\sigma \in S_4$. Then $a = \psi^{-1}(\sigma) \in G$ is a rotational symmetry of the cube. Consider the six pictures of the cube below, each with a distinguished pair of diagonal axes chosen. These six pictures of the cube are divided into 3 columns labelled $1, 2, 3$.



You may verify that the rotation a takes the two pictures in column 1 to either the two pictures in column 2, or the two pictures in column 3, and so on. Thus the columns $1, 2, 3$ above are permuted. We obtain a permutation of $\{1, 2, 3\}$ which we call $\mu(\sigma) \in S_3$. Further,

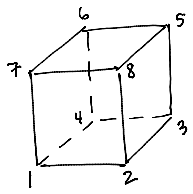
$$\ker(\mu) = H = \{e, (12)(34), (13)(24), (14)(23)\},$$

is a normal subgroup, and μ is onto. The 1st Isomorphism Theorem then tells us that

$$S_4/H \cong S_3$$

This example is actually rare: there are no homomorphisms $S_n \to S_{n-1}$ when $n > 4$!

Rotational symmetries of a cube



Original position with vertices labelled 1-8

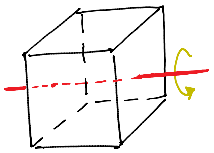
There are 24 symmetries.

By keeping track of where vertices go, we get a 1-1 homomorphism

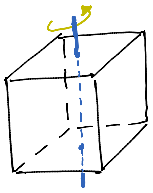
{rotational symmetries of the cube} $\longrightarrow S_8$.

identity: corresponds to identity permutation $e \in S_8$

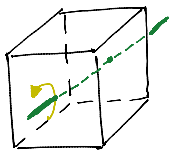
For each axis as below, get 3 nontrivial rotations:



90° (2358)(1467)
180° (25)(38)(16)(47)
270° (2853)(1764)

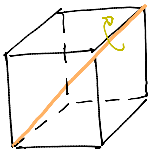


90° (1234)(5678)
180° (13)(24)(57)(68)
270° (1432)(5876)

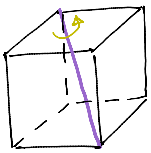


90° (1287)(5643)
180° (18)(27)(54)(63)
270° (1782)(5346)

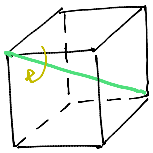
2 nontrivial rotations around each of the 4 diagonal axes:



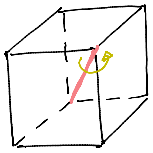
120° (386)(274)
240° (368)(247)



120° (183)(475)
240° (138)(457)

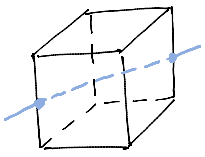


120° (168)(245)
240° (186)(254)

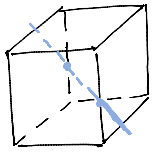


120° (257)(136)
240° (275)(163)

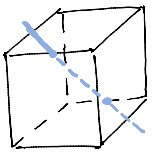
180° rotation around each edge-midpoint diagonal:



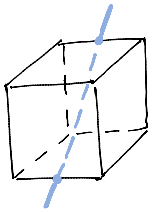
(17)(26)(35)(48)



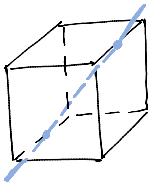
(15)(28)(37)(46)



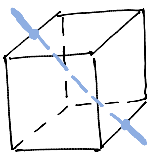
(15)(26)(34)(78)



(12)(37)(48)(56)



(14)(26)(37)(58)



(15)(23)(48)(67)

More isomorphism theorems

Last lecture we learned about the 1st Isomorphism Theorem. As its name suggests, it has a few sequels, whose names are just as imaginative!

For subsets H and N of a group G we define the subset $HN \subset G$ to be:

$$HN = \{ab : \ a \in H, b \in N\}$$

▶ **2nd Isomorphism Theorem**: **Let H, N be subgroups a group G with N normal. Then HN is a subgroup of G, $H \cap N$ is a normal subgroup of H, and**

$$\frac{H}{H \cap N} \cong \frac{HN}{N}$$

Proof. Let us first show HN is a subgroup of G. Clearly $e \in HN$, since $e = ee$ and $e \in H$, $e \in N$. Let $ab, a'b' \in HN$ where $a, a' \in H$, $b, b' \in N$. Then $Na' = a'N$ since N is normal, and this implies $ba' = a'b''$ for some $b'' \in N$. Thus

$$(ab)(a'b') = a(ba')b' = (aa')(b''b') \in HN.$$

Next, let $ab \in HN$. Then $Na^{-1} = a^{-1}N$ implies $b^{-1}a^{-1} = a^{-1}b'$ for some $b' \in N$. Thus $(ab)^{-1} = b^{-1}a^{-1} = a^{-1}b' \in HN$. Thus HN is a subgroup of G.

The argument that $H \cap N$ is a normal subgroup of H is also straightforward, and is omitted.

To prove the isomorphism, we construct a homomorphism

$$\phi : H \longrightarrow HN/N$$

by setting, for each $a \in H$, $\phi(a) = aN$. Any coset in HN/N is of the form $abN = aN$ for $ab \in HN$, and $\phi(a) = aN$, so ϕ is onto. The map ϕ is a homomorphism:

$$\phi(aa') = aa'N = aNa'N = \phi(a)\phi(a')$$

Now let us compute $\ker(\phi)$. Suppose $a \in H$ and $\phi(a) = aN = N$. This means $a \in N$. Thus $\ker(\phi) = H \cap N$. The 1st Isomorphism Theorem then gives

$$HN/N \cong H/\ker(\phi) = H/H \cap N \qquad \square$$

▶ **Example:** Let m, n be positive integers. Let $G = \mathbb{Z}$ and $H = m\mathbb{Z}$, $N = n\mathbb{Z}$. Recall that $\gcd(m, n)$ is characterized as being the smallest positive integer contained in the set

$$m\mathbb{Z} + n\mathbb{Z} = \{am + bn : \ a, b \in \mathbb{Z}\}$$

This subset of $\mathbb{Z}$ is the subgroup "HN" in this example. We see that $m\mathbb{Z}+n\mathbb{Z} = \gcd(m, n)\mathbb{Z}$. On the other hand, the subset corresponding to $H \cap N$ is given by

$$m\mathbb{Z} \cap n\mathbb{Z} = \{k \in \mathbb{Z} : k = am = an \text{ for some } a, b \in \mathbb{Z}\}$$

Noting that the smallest positive integer in this subset is the lcm of m and n, we obtain $m\mathbb{Z} \cap n\mathbb{Z} = \text{lcm}(m,n)\mathbb{Z}$. The 2nd Isomorphism Theorem gives us

$$\frac{m\mathbb{Z}}{\text{lcm}(m,n)\mathbb{Z}} = \frac{H}{H \cap N} \cong \frac{HN}{N} = \frac{\gcd(m,n)\mathbb{Z}}{n\mathbb{Z}}$$

If k divides l, then $|k\mathbb{Z}/l\mathbb{Z}| = l/k$. (This can be shown explicitly, but is also deduced as a consequence of the 3rd Isomorphism Theorem below.) By comparing orders of groups,

$$\text{lcm}(m,n)/m = n/\gcd(m,n) \quad \Longrightarrow \quad \text{lcm}(m,n)\gcd(m,n) = mn.$$

▶ **Example:** Consider $G = S_4$ and the normal subgroup $N = \{e, (12)(34), (13)(24), (14)(23)\}$ of order 4 inside of S_4. Then we saw from last lecture that $S_4/N \cong S_3$ by constructing an onto homomorphism $\phi : S_4 \to S_3$, recognizing its kernel to be N, and using the 1st Isomorphism Theorem. Let us give a different method for showing $S_4/N \cong S_3$.

Take $H = S_3 \subset S_4$. This is a (non-normal) subgroup. Then we first claim

$$HN = S_4.$$

To see this, we proceed as follows. First, $S_3 = S_3 e \subset HN$, so in particular $(12), (23), (31) \in HN$. Next, as $(12), (23), (31) \in S_3$ and $(12)(34), (23)(14), (31)(24) \in N$ we get the products

$$(12)\,((12)(34)) = (34), \qquad (23)\,((23)(14)) = (14), \qquad (31)\,((31)(24)) = (24)$$

also as elements of HN. Thus every transposition $(12), (13), (14), (23), (24), (34)$ in S_4 is in HN. Every permutation in S_4 is a product of transpositions, and therefore $S_4 = HN$. Next note that $H \cap N = \{e\}$ by direct inspection. Finally, the 2nd Isomorphism Theorem gives

$$S_4/N = HN/N \cong S_3/S_3 \cap N = S_3/\{e\} \cong S_3$$

▶ **3rd Isomorphism Theorem**: **Let G be a group and suppose $H, N \subset G$ are normal subgroups with $N \subset H$. Then we have an isomorphism**

$$\frac{G}{H} \cong \frac{G/N}{H/N}$$

Proof. We define a map $\phi : G/N \to G/H$ by setting $\phi(aN) = aH$. Note that if $aN = bN$, then $ab^{-1} \in N \subset H$, and thus $aH = bH$. Thus $\phi(aN) = aH = bH = \phi(bN)$, and ϕ is well-defined. Furthermore, ϕ is an onto homomorphism, by repeating earlier arguments.

The kernel of ϕ is given by $aN \in G/N$ such that $\phi(aN) = aH = H$, which is true if and only if $a \in H$. Thus $\ker(\phi) = \{aN : a \in H\} = H/N$. The 1st Isomorphism Theorem gives

$$G/H \cong \frac{G/N}{\ker(\phi)} = \frac{G/N}{H/N}$$

□

▶ **Example:** Consider $\mathbb{Z}_n = \mathbb{Z}/n\mathbb{Z}$. Note that $mn\mathbb{Z} \subset n\mathbb{Z}$ for any integer m. With $G = \mathbb{Z}$, $H = n\mathbb{Z}$ and $N = mn\mathbb{Z}$ we obtain from the 3rd Isomorphism Theorem:

$$\mathbb{Z}_n = \frac{\mathbb{Z}}{n\mathbb{Z}} \cong \frac{\mathbb{Z}/mn\mathbb{Z}}{n\mathbb{Z}/mn\mathbb{Z}}$$

Taking orders, we get $n = |\mathbb{Z}_n| = |\mathbb{Z}/mn\mathbb{Z}|/|n\mathbb{Z}/mn\mathbb{Z}| = mn/|n\mathbb{Z}/mn\mathbb{Z}|$. In particular,

$$|n\mathbb{Z}/mn\mathbb{Z}| = m.$$

Setting $l = mn$ and $k = n$ we deduce in general that if k divides l then $|k\mathbb{Z}/l\mathbb{Z}| = l/k$.

Cayley's Theorem

How can we understand all finite groups? One way, which we return to later, is to try and classify, i.e. list, all groups up to isomorphism. Here is another way to "understand all finite groups": show that all such groups are isomorphic to subgroups of groups we understand. This can be done using the following theorem.

▶ **Cayley's Theorem: Let G be a finite group, and $|G| = n$. Then there is a 1-1 homomorphism $\phi : G \to S_n$. In particular, G is isomorphic to a subgroup of S_n.**

Before proving this theorem, let's see why it does what we want. Given any homomorphism $\phi : G \to G'$ we can define a new homomorphism $\psi : G \to \text{im}(\phi)$ by setting $\psi(a) = \phi(a)$ for all $a \in G$; we have only changed the definition of the target group. The homomorphism $\psi : G \to \text{im}(\phi)$ has the advantage that it is *onto*. The 1st Isomorphism Theorem for ψ yields

$$G/\ker(\phi) \cong \text{im}(\phi)$$

In particular, when ϕ is 1-1, we have $G \cong \text{im}(\phi)$. Now return to the statement of Cayley's Theorem, which gives a 1-1 homomorphism $\phi : G \to S_n$. Then $G \cong \text{im}(\phi)$, and $\text{im}(\phi) \subset S_n$ is a subgroup of S_n. Thus G is isomorphic to a subgroup of S_n.

Proof. Label elements of G as $a_1, a_2, \ldots, a_n$. Let $a \in G$. (This a is among the a_i, but we will just write a for it.) Define a function $\sigma_a : \{1, \ldots, n\} \to \{1, \ldots, n\}$ as follows:

$$aa_i = a_{\sigma_a(i)}$$

In other words, given $i \in \{1, \ldots, n\}$, we take the product $aa_i \in G$, and it is equal to some a_k where $k \in \{1, \ldots, n\}$; we let $\sigma_a(i) = k$. The map σ_a is 1-1: suppose $\sigma_a(i) = \sigma_a(j)$. Then

$$aa_i = aa_j \quad \Longrightarrow \quad a_i = a_j \quad \Longrightarrow \quad i = j$$

The map σ_a is onto: suppose $k \in \{1, \ldots, n\}$. Let a_i be such that $a = a_k a_i^{-1}$. Then

$$aa_i = a_k a_i^{-1} a_i = a_k \quad \Longrightarrow \quad \sigma_a(i) = k$$

Thus σ_a is 1-1 and onto, and therefore a permutation, i.e. $\sigma_a \in S_n$.

Now we are in a position to define the sought after homomorphism:

$$\phi : G \longrightarrow S_n$$

For $a \in G$ we declare $\phi(a) = \sigma_a$. We check this is a homomorphism. Let $a, b \in G$. We must compare $\phi(a)\phi(b) = \sigma_a \circ \sigma_b$ with $\phi(ab) = \sigma_{ab}$. We compute

$$a_{\sigma_{ab}(i)} = (ab)a_i = a(ba_i) = aa_{\sigma_b(i)} = a_{\sigma_a(\sigma_b(i))}$$

for any $i \in \{1, \ldots, n\}$, which shows that $\sigma_{ab} = \sigma_a \circ \sigma_b$. To show ϕ is 1-1, we compute its kernel. Suppose $\phi(a) = e$, i.e. $\sigma_a(i) = i$ for all $i \in \{1, \ldots, n\}$. This means

$$aa_i = a_{\sigma_a(i)} = a_i$$

which implies $a = e$. Thus $\ker(\phi) = \{e\}$ and therefore ϕ is 1-1. □

Let us see Cayley's Theorem in action. To clarify the main point, take G to be an interesting group which is not defined using permutations. Let us introduce the *quaternion group*

$$Q_8 = \{\, 1, -1,\ \mathbf{i}, -\mathbf{i},\ \mathbf{j}, -\mathbf{j},\ \mathbf{k}, -\mathbf{k}\}$$

of order 8, with the relations that -1 commutes with everything, $(-1)^2 = 1$, and also $\mathbf{i}^2 = \mathbf{j}^2 = \mathbf{k}^2 = -1$, and $\mathbf{ij} = \mathbf{k}$. These relations generate the Cayley table of Q_8 as follows:

	1	-1	$\mathbf{i}$	$-\mathbf{i}$	$\mathbf{j}$	$-\mathbf{j}$	$\mathbf{k}$	$-\mathbf{k}$
1	1	-1	$\mathbf{i}$	$-\mathbf{i}$	$\mathbf{j}$	$-\mathbf{j}$	$\mathbf{k}$	$-\mathbf{k}$
-1	-1	1	$-\mathbf{i}$	$\mathbf{i}$	$-\mathbf{j}$	$\mathbf{j}$	$-\mathbf{k}$	$\mathbf{k}$
$\mathbf{i}$	$\mathbf{i}$	$-\mathbf{i}$	-1	1	$\mathbf{k}$	$-\mathbf{k}$	$-\mathbf{j}$	$\mathbf{j}$
$-\mathbf{i}$	$-\mathbf{i}$	$\mathbf{i}$	1	-1	$-\mathbf{k}$	$\mathbf{k}$	$\mathbf{j}$	$-\mathbf{j}$
$\mathbf{j}$	$\mathbf{j}$	$-\mathbf{j}$	$-\mathbf{k}$	$\mathbf{k}$	-1	1	$\mathbf{i}$	$-\mathbf{i}$
$-\mathbf{j}$	$-\mathbf{j}$	$\mathbf{j}$	$\mathbf{k}$	$-\mathbf{k}$	1	-1	$-\mathbf{i}$	$\mathbf{i}$
$\mathbf{k}$	$\mathbf{k}$	$-\mathbf{k}$	$\mathbf{j}$	$-\mathbf{j}$	$-\mathbf{i}$	$\mathbf{i}$	1	-1
$-\mathbf{k}$	$-\mathbf{k}$	$\mathbf{k}$	$-\mathbf{j}$	$\mathbf{j}$	$\mathbf{i}$	$-\mathbf{i}$	-1	1

Cayley's Theorem gives us a way to realize Q_8 as a subgroup of the symmetric group S_8. The construction is as follows. First we label the elements of Q_8 as $a_1, \ldots, a_8$. For example, we may use the following labelling (although any labelling will do):

$$a_1 = 1, \quad a_2 = -1, \quad a_3 = \mathbf{i}, \quad a_4 = -\mathbf{i}, \quad a_5 = \mathbf{j}, \quad a_6 = -\mathbf{j}, \quad a_7 = \mathbf{k}, \quad a_8 = -\mathbf{k}$$

Next, given an element $a \in Q_8$ we define the associated permuation $\sigma_a \in S_8$ by the relation

$$aa_i = a_j \qquad \Longleftrightarrow \qquad \sigma_a(i) = j$$

Let us spell this out for the element $a = \mathbf{i} \in Q_8$. We have

$$\begin{aligned} aa_1 &= (\mathbf{i})(1) = \mathbf{i} = a_3 & \qquad aa_2 &= (\mathbf{i})(-1) = -\mathbf{i} = a_4 \\ aa_3 &= (\mathbf{i})(\mathbf{i}) = -1 = a_2 & \qquad aa_4 &= (\mathbf{i})(-\mathbf{i}) = 1 = a_1 \\ aa_5 &= (\mathbf{i})(\mathbf{j}) = \mathbf{k} = a_7 & \qquad aa_6 &= (\mathbf{i})(-\mathbf{j}) = -\mathbf{k} = a_8 \\ aa_7 &= (\mathbf{i})(\mathbf{k}) = -\mathbf{j} = a_6 & \qquad aa_8 &= (\mathbf{i})(-\mathbf{k}) = \mathbf{j} = a_5 \end{aligned}$$

These relations then determine the permutation $\sigma_a = \sigma_{\mathbf{i}}$ as follows:

$$\sigma_a(1) = 3, \quad \sigma_a(2) = 4, \quad \sigma_a(3) = 2, \quad \sigma_a(4) = 1,$$

$$\sigma_a(5) = 7, \quad \sigma_a(6) = 8, \quad \sigma_a(7) = 6, \quad \sigma_a(8) = 5$$

All together we may write our permutation $\sigma_a = \sigma_{\mathbf{i}}$ in cycle notation as

$$\sigma_{\mathbf{i}} = (1324)(5768) \in S_8$$

We may proceed to do this for each element $a \in Q_8$. We get the following subgroup of S_8:

$$\{e,\ (12)(34)(56)(78),\ (1324)(5768),\ (1423)(5867),$$
$$(1526)(3847),\ (1625)(3748),\ (1728)(3546),\ (1827)(3645)\}$$

The elements are listed in the order corresponding to the above ordering of the elements of Q_8. In summary, Q_8 is isomorphic to the subgroup of S_8 displayed above.

Cayley's Theorem is conceptually very important: it says that every finite group, however abstractly defined, is isomorphic to a subgroup of some symmetric group, a very concrete type of group that we understand how to work with. On the other hand, Cayley's Theorem is often not practical: it realizes our group as a subgroup of a generally very large group. For example, in the example above, Q_8 has order 8 and we realized it as a subgroup of S_8 which is of order $8! = 40320$.

The following is a generalization of Cayley's Theorem.

▶ **Let G be a group, $H \subset G$ a subgroup, and suppose $[G : H] = n$ is finite. Then there is a homomorphism $\phi : G \to S_n$. Furthermore, $\ker(\phi)$ is equal to the largest normal subgroup of G which is contained in H.**

To obtain Cayley's Theorem as a special case, let $H = \{e\}$. Then $n = [G : \{e\}] = |G|$. The largest normal subgroup contained in $\{e\}$ is of course $\{e\}$, and so $\ker(\phi) = \{e\}$, i.e. ϕ is 1-1.

The proof of this generalization is similar to that of Cayley's Theorem. We list the distinct cosets in G/H as $a_1H, \ldots, a_nH$. Then define $\phi : G \to S_n$ by $\phi(a) = \sigma_a$ where $\sigma_a \in S_n$ is determined by the relation: $\sigma_a(i) = j$ if and only if $aa_iH = a_jH$, where $i, j \in \{1, \ldots, n\}$. The details for the rest of the proof are left as an exercise.

▶ **Let G be a finite group, $H \subset G$ a proper subgroup, and suppose $|G|$ does not divide $[G : H]!$. Then H contains a non-trivial normal subgroup of G.**

This is a corollary of the above theorem. Suppose H does *not* contain a non-trivial normal subgroup. Then the homomorphism $\phi : G \to S_n$ from the above theorem, where $n = [G : H]$, must have $\ker(\phi) = \{e\}$. Thus $G \cong \text{im}(\phi)$ is isomorphic to a subgroup of S_n. By Lagrange's Theorem, we must have that $|G|$ divides $|S_n| = n! = [G : H]!$.

Classification of finite groups

A main goal of group theory is the *classification of groups*. That is, we would like to understand what all the possible groups are. If we can achieve this goal, then we understand all groups! Let us focus today on the case of *finite* groups.

When finding all possible finite groups, we want to avoid duplicates. For example, we know that the symmetric group S_3 is isomorphic to the group of symmetries of an equilateral triangle. As another example, we know $\mathbb{Z}_{10}^\times$ is isomorphic to $\mathbb{Z}_4$. We only want to list one such group for each isomorphism type. So we only write S_3 and $\mathbb{Z}_4$ for the above examples, and forget about the others, because they are isomorphic to one of these groups.

Thus our goal is narrowed down to the classification of finite groups *up to isomorphism*. Below is a table which classifies groups up to isomorphism up to order 14.

Order	Groups
1	$\{e\}$
2	$\mathbb{Z}_2$
3	$\mathbb{Z}_3$
4	$\mathbb{Z}_4$, $\mathbb{Z}_2 \times \mathbb{Z}_2$
5	$\mathbb{Z}_5$
6	$\mathbb{Z}_6$, S_3
7	$\mathbb{Z}_7$

Order	Groups
8	$\mathbb{Z}_8$, $\mathbb{Z}_4 \times \mathbb{Z}_2$, $\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2$, D_4, Q_8
9	$\mathbb{Z}_9$, $\mathbb{Z}_3 \times \mathbb{Z}_3$
10	$\mathbb{Z}_{10}$, D_5
11	$\mathbb{Z}_{11}$
12	$\mathbb{Z}_{12}$, $\mathbb{Z}_6 \times \mathbb{Z}_2$, A_4, D_6, T
13	$\mathbb{Z}_{13}$
14	$\mathbb{Z}_{14}$, D_7

Let us explain some of what is in this table. Recall we proved that every group of prime order is cyclic, and further that every finite cyclic group is isomorphic to some $\mathbb{Z}_n$. This covers the cases $2, 3, 5, 7, 11, 13$ in the table.

Let G be a group of order 4. Suppose G is not cyclic, i.e. $G \ncong \mathbb{Z}_4$. Let $a, b \in G$ be distinct non-identity elements. Then ab must have order dividing $|G| = 4$. We cannot have $\text{ord}(ab) = 4$ for otherwise G is cyclic. Suppose $\text{ord}(ab) = 1$. Then $ab = e$, or $a = b^{-1}$, and $b^2 = e$ implies $a = b^{-1} = b$, contradiction. Thus $\text{ord}(ab) = 2$. Further, ab is distinct from a, b. To see this, note that if $ab = a$ then $b = e$, and if $ab = b$ then $a = e$. We may thus write

$$G = \{e, a, b, ab\}$$

Further, $(ab)^2 = abab = e = a^2b^2$ implies $ba = ab$. This determines the group operation on all elements of G, and in particular the Cayley table for G. Now we define a map

$$\phi : G \longrightarrow \mathbb{Z}_2 \times \mathbb{Z}_2$$

by setting $\phi(e) = 0$, $\phi(a) = (1, 0)$, $\phi(b) = (0, 1)$ and $\phi(ab) = (1, 1)$. It is straightforward to verify that ϕ is an isomorphism. Thus we have shown that a group of order 4 which is not isomorphic to $\mathbb{Z}_4$ is in fact isomorphic to $\mathbb{Z}_2 \times \mathbb{Z}_2$.

You have seen most of the groups in the above table: the exceptions are perhaps D_n and T. The group D_n is the *Dihedral group* of order $2n$, and for $n \geqslant 3$ is non-abelian. It can be defined as the symmetries of a regular n-gon in the plane.

The group T of order 12 may be described in quaternion notation as follows:

$$T = \Big\{1,\ \ -1,\ \ \tfrac{1}{2} + \tfrac{\sqrt{3}}{2}\mathbf{i},\ \ -\tfrac{1}{2} + \tfrac{\sqrt{3}}{2}\mathbf{i},\ \ -\tfrac{1}{2} - \tfrac{\sqrt{3}}{2}\mathbf{i},\ \ \tfrac{1}{2} - \tfrac{\sqrt{3}}{2}\mathbf{i},$$
$$\mathbf{j},\ \ -\mathbf{j},\ \ \tfrac{1}{2}\mathbf{j} + \tfrac{\sqrt{3}}{2}\mathbf{k},\ \ -\tfrac{1}{2}\mathbf{j} + \tfrac{\sqrt{3}}{2}\mathbf{k},\ \ -\tfrac{1}{2}\mathbf{j} - \tfrac{\sqrt{3}}{2}\mathbf{k},\ \ \tfrac{1}{2}\mathbf{j} - \tfrac{\sqrt{3}}{2}\mathbf{k}\Big\}$$

As in the quaterion group Q_8, the quaternions $\mathbf{i}$, $\mathbf{j}$, $\mathbf{k}$ satisfy $\mathbf{i}^2 = \mathbf{j}^2 = \mathbf{k}^2 = -1$, and $\mathbf{ij} = \mathbf{k}$.

You can continue with the brute force reasoning we used above for $|G| = 4$ and show that every non-abelian group of order 6 is isomorphic to S_3. But continuing in this fashion will only get you so far.

The first measure we can take to narrow our scope is to separate the abelian and non-abelian finite groups. The non-abelian entries in the above table are in red. Classifying finite *abelian* groups is a problem that can be solved, and the answer is rather simple:

▶ **Classification of finite abelian groups: Any finite abelian group is isomorphic to a product of cyclic groups of the following form, where the p_i's are primes:**

$$\mathbb{Z}_{p_1^{a1}} \times \mathbb{Z}_{p_2^{a2}} \times \cdots \times \mathbb{Z}_{p_n^{an}}$$

In this classification, note that the primes p_i need not be distinct. For example, $\mathbb{Z}_2 \times \mathbb{Z}_2$ is the case where $n = 2$, $p_1 = p_2 = 2$ and $a_1 = a_2 = 1$.

▶ **Example:** Consider the group $\mathbb{Z}_{24}^\times$. This is a finite abelian group, so it must be isomorphic to a group as described above. We first write out

$$\mathbb{Z}_{24}^\times = \{1, 5, 7, 11, 13, 17, 19, 23\}$$

where each integer is understood $\pmod{24}$. Thus $|\mathbb{Z}_{24}^\times| = 8$ and so by the Classification Theorem of finite abelian groups, $\mathbb{Z}_{24}^\times$ must be isomorphic to one of

$$\mathbb{Z}_8 = \mathbb{Z}_{2^3} \qquad \mathbb{Z}_{2^2} \times \mathbb{Z}_2 = \mathbb{Z}_4 \times \mathbb{Z}_2 \qquad \mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2$$

Note that $5^2 \equiv 25 \equiv 1 \pmod{24}$, $7^2 \equiv 49 \equiv 1 \pmod{24}$, and so on – every element of $\mathbb{Z}_{24}^\times$ apart from the identity has order 2. The only of the above 3 groups that shares this property is $\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2$. We conclude that $\mathbb{Z}_{24}^\times \cong \mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2$.

After classifying abelian finite groups, we may restrict our attention to *non-abelian* finite groups. Another way to narrow our scope is to focus only on *simple* groups.

▶ **A group G is *simple* if $G \neq \{e\}$ and its only normal subgroups are $\{e\}$ and G.**

Finite simple groups are thought of as the "building blocks" for all finite groups. If a finite group G is *not* simple, it has a non-trivial proper normal subgroup $N \subset G$. We obtain an onto homomorphism $G \to G/N$ to a *smaller* group G/N. Then G is to be thought of as decomposed into the smaller parts of N and G/N. If on the other hand G is simple, there is no homomorphism to a smaller non-trivial group, and G cannot be "decomposed".

▶ **Classification of finite simple groups: A finite simple group is isomorphic to:**

(1) A cyclic group $\mathbb{Z}_p$ of prime order.

(2) An alternating group A_n where $n \geqslant 5$.

(3) A finite non-abelian group of "Lie Type".

(4) A "Sporadic" finite non-abelian group, of which there are 26.

This classification theorem is one of the most important results in the history of mathematics. Its proof originally rested on tens of thousands of pages of work due to many mathematicians, although the proof has been simplified over the years. The classification was stated in the 1980's by Gorenstein, although the proof was finished years later.

What are the groups on this list? We know all about the cases (1) $\mathbb{Z}_p$ and (2) A_n. The groups in (3), of "Lie Type", fall into a list of infinite families of systematically constructed groups. To give one family in this type, for each prime p consider:

$$\mathrm{SL}_2(\mathbb{Z}_p) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} : \quad a,b,c,d \in \mathbb{Z}_p, \qquad ad - bc \equiv 1 \pmod{p} \right\}$$

This group is generally not simple, because it has a normal subgroup of order 2 consisting of $\{I, -I\}$ where I is the identity matrix (modulo p). However, the quotient group

$$\mathrm{PSL}_2(\mathbb{Z}_p) = \mathrm{SL}_2(\mathbb{Z}_p)/\{I, -I\}$$

is simple when $p \geqslant 5$. In fact, the first few of these are isomorphic to groups you know:

$$\mathrm{PSL}_2(\mathbb{Z}_2) \cong S_3 \qquad \mathrm{PSL}_2(\mathbb{Z}_3) \cong A_4 \qquad \mathrm{PSL}_2(\mathbb{Z}_5) \cong A_5$$

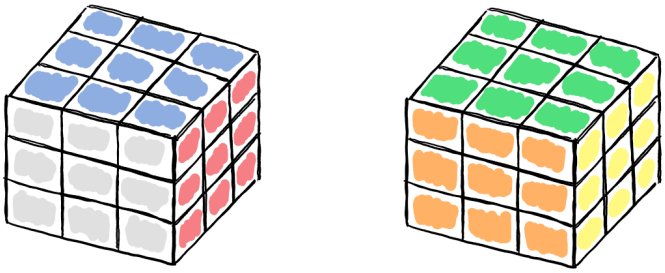
But for higher primes p these are genuinely new groups. Other groups of "Lie type" are of a similar flavor, in a broad sense.

The final category (4) of "Sporadic" finite simple groups are a list of 26 groups that do not follow a systematic pattern. They all have order well over 1000. The largest of these, called the Monster Group, has order about 8×10^{53}, and was constructed by Richard Griess in 1982 as a group of rotational symmetries in 196883-dimensional space.

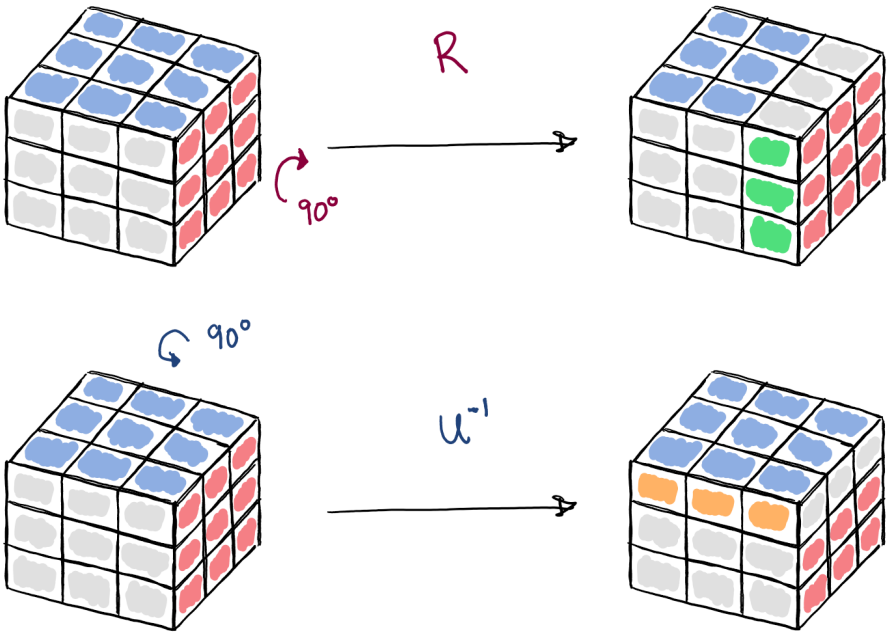
The Rubik's cube group

Today we will study the Rubik's cube using some of the group theory we have developed. During lecture, I'll hand out Rubik's cubes and you should play along. If you don't know how to solve the cube, however, be careful not to scramble it!

Let's begin with a top (on the left) and bottom (on the right) view of the Rubik's cube:



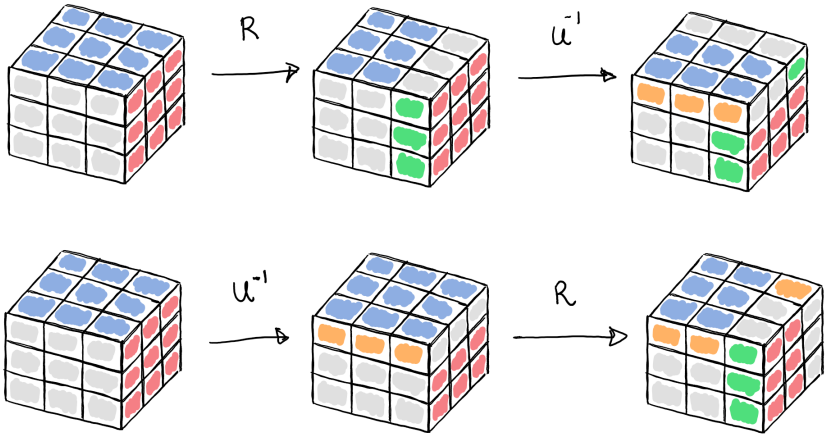
We label the 6 faces of the cube F (front), B (back), U (up), D (down), L (left), R (right). To each face we associate a symmetry of the cube which is a 90° clockwise rotation of that face, if you are facing the face. For example, we have the moves R and U^{-1}:



And R^2, for example, is 180° rotation of the right face. We write G for the *Rubik's cube group*, which is the group of symmetries generated by the moves F, B, U, D, L, R.

Our convention is that the expression $LU^{-1}RD$ means: first do D, then do R, then U^{-1}, and then L. Warning: This is opposite the order in which Rubik's cube enthusiasts write moves, but it goes better with our conventions in group theory.

Let us try to understand the group G. First, compare the following two sequences of moves:



This shows that $U^{-1}R \neq RU^{-1}$. In particular G is non-abelian.

However we can find non-trivial abelian subgroups. Here's an example. Consider the subgroup generated by R. This is clearly $\langle R \rangle = \{e, R, R^2, R^3\}$ as R is a 90° rotation of a face. So we get a cyclic subgroup of order 4, which is isomorphic to $\mathbb{Z}_4$. Similarly, $\langle L \rangle \cong \mathbb{Z}_4$. Note also that L and R *commute* as these two faces are not sharing any part of the cube. From this we get a subgroup $\langle R, L \rangle \subset G$ isomorphic to $\mathbb{Z}_4 \times \mathbb{Z}_4$.

Let's look at the orders of some elements in G. Check with your cube the following:

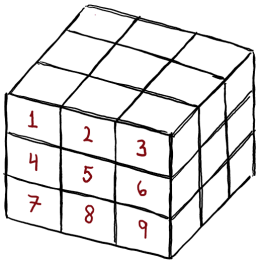
$$\text{ord}(R^{-1}UR) = 4, \qquad \text{ord}(UR^{-1}UR) = 5$$

As another example, you can check (after a few minutes!) that $\text{ord}(UR) = 105$. For a more extravagant example, one can check that

$$\text{ord}(B^{-1}UB^{-1}F^2R) = 1260,$$

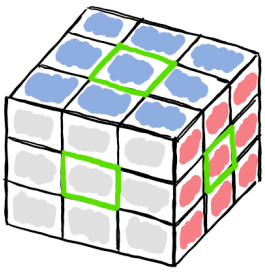
although I suggest you don't waste your time doing it! This order, 1260, is in fact the highest possible order of any element in G.

How should we better understand G? One way to is map it into a symmetric group, just as we've done for the triangle and the tetrahedron. Label the "sticker positions" of the cube:



In the picture we labelled one face, but continuing in this fashion we will get $(9)(6) = 54$ labels. Each move $a \in G$ is determined by how these sticker positions are permuted, and to each such a we get permutation in S_{54}. This gives a 1-1 homomorphism $G \to S_{54}$.

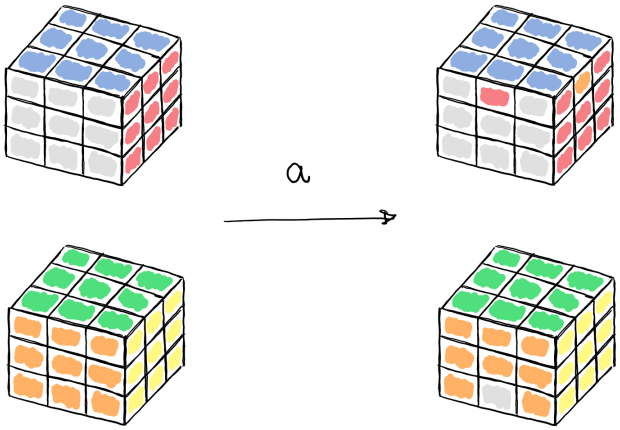
But we can do better! Note each $a \in G$ fixes the middle sticker position on each face.



So forget about these. We only label the other sticker positions, of which there are 48. This gives a homomorphism $G \to S_{48}$. Note that $48! \approx 10^{61}$. However it turns out $|G| = 8!12!2^{10}3^7 \approx 4 \times 10^{19}$, which is very large but much smaller than $|S_{48}| = 48!$.

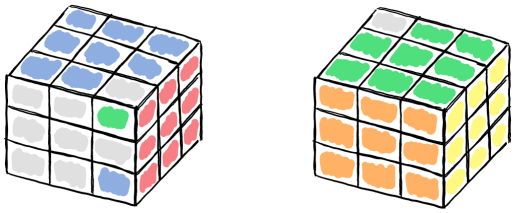
How can you solve the Rubik's cube? The basic strategy is as follows: (1) first, come up with some basic moves that permute only a few parts of the cube; (2) use these basic moves to progressively solve parts of the cube. (In reality it is best to just learn one of the algorithms that solves it.) We won't solve the cube but rather aim to understand G better.

But let's focus on (1) for a moment. Here is a "basic move" called an *edge 3-cycle*:



Both top and bottom views are shown. It is called an edge 3-cycle because exactly 3 pieces (subcubes) are permuted. A representation of $a \in G$ is given by $a = F^2UL^{-1}RF^2LR^{-1}UF^2$.

Having shown that we can permute 3 edge cubes as above, we might suspect we can permute any given set of subcubes as we like. Not true! For example, consider the configuration:



This configuration has exactly 2 corner cubes swapped. To show this is impossible, we introduce another homomorphism. Consider subcube positions which are not in the middles of faces. You can count there are 20 such non-middle subcube positions in the cube.

Next, label these subcube positions $1, \ldots, 20$ in any fashion. By seeing how a symmetry $a \in G$ permutes these 20 subcube positions, we obtain a homomorphism

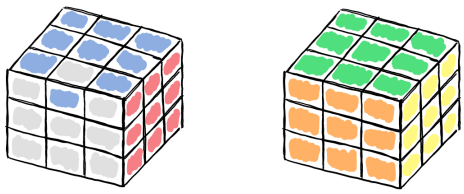
$$\phi : G \longrightarrow S_{20}$$

For example, take the subcubes starting on the right face at a corner and label them $1, \ldots, 8$ in a clockwise fashion (and label the others arbitrarily) then $\phi(R) = (1357)(2468)$. Note this permutation is *even*. A similar computation shows $\phi(a)$ is even for F, B, U, D, L. But every element in G can be written as a composition of these moves, so we have shown:

▶ **The homomorphism $\phi : G \to S_{20}$ has $\mathbf{im}(\phi) \subset A_{20}$.**

In other words, for any symmetry $a \in G$, the permutation $\phi(a)$ is even. Now the proof of why the configuration drawn previously is impossible is easy! Just note that it swaps exactly 2 subcubes, so the picture corresponds to a transposition in S_{20}, which is odd, and this contradicts what we have just found.

We can also prove that the following configuration is impossible:



Note that this configuration does not swap any subcubes; it only changes the "orientation" of exactly one subcube. It is in fact in the kernel of the homomorphism ϕ, so that homomorphism will not help us here.

We can define another homomorphism as follows. Notice that each edge piece (not a corner) has 2 stickers. Collecting all such pairs of stickers, there are 24. Labelling these $1, \ldots, 24$ and seeing how each $a \in G$ permutes these, we obtain a homomorphism $\psi : G \to S_{24}$. Again, it is not hard to show that $\psi(R)$ is even, and the same is true for F, B, U, D, L, and thus for all $a \in G$. On the other hand, the above drawn configuration corresponds to a transposition which swaps exactly 2 of these edge stickers, which is impossible!

Definition of a ring

A *ring* R is a set, with elements $0, 1 \in R$, together with two binary operations $+$ (addition) and $\times$ (multiplication, often written $\cdot$ or omitted) such that the following properties hold:

(Associativity for $+$)	$a + (b + c) = (a + b) + c$
(Associativity for $\times$)	$a(bc) = (ab)c$
(Identity for $+$)	$a + 0 = 0 + a = a$
(Identity for $\times$)	$1 \cdot a = a \cdot 1 = a$
(Inverses for $+$)	$a + (-a) = (-a) + a = 0$
(Distributivity)	$a(b + c) = ab + ac, \quad (a + b)c = ac + bc$
(Commutativity for $+$)	$a + b = b + a$

These properties are for all $a, b, c \in R$. The "Inverses for $+$" property should be understood as: for each $a \in R$, there is an element called $-a \in R$ such that $a + (-a) = (-a) + a = 0$. We abbreviate $a + (-a)$ as $a - a$. A ring is *commutative* if $ab = ba$ for all $a, b \in R$.

Examples of rings include $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$ and $\mathbb{Z}_n$, all with the usual operations of addition and multiplication. Note these are all *commutative* rings.

A "shorter" definition of a ring is as follows: a ring R is a set with binary operations $+$ and $\times$ such that $(R, +)$ is an abelian group with identity $0 \in R$; and $\times$ is associative and has an identity $1 \in R$; and the distributivity properties hold.

We remark that the axiom that "$+$" is commutative is implied by the other axioms. To see this we compute $(1 + 1)(a + b)$ in two different ways:

$$(1 + 1)(a + b) = 1(a + b) + 1(a + b) = a + b + a + b$$

$$(1 + 1)(a + b) = (1 + 1)a + (1 + 1)b = a + a + b + b$$

We have used distributivity and also that 1 is a multiplicative identity. Then we can add $-a$ to the left sides and $-b$ to the right sides of these equations to get $b + a = a + b$.

▶ **Let R be a ring and $a, b \in R$. Then:**

(i) $a0 = 0a = 0$

(ii) $a(-b) = -ab = (-a)b$

(iii) $(-a)(-b) = ab$

Proof. To prove (i), using $0 = -0$, we compute $a0 = a(0 - 0) = a0 - a0 = 0$, and similarly $0a = 0$. For (ii), use distributivity: $ab + a(-b) = a(b + (-b)) = a0 = 0$; this implies $a(-b) = -ab$. Similarly $(-a)b = -ab$. For (iii): from (ii), $(-a)(-b) = -(a(-b)) = -(-(ab)) = ab$. □

▶ **Suppose $0 = 1$ in a ring R. Then $R = \{0\}$.**

To see this, we compute for any $a \in R$: $a = a1 = a0 = 0$. Thus every element in R is equal to 0, and this proves the claim. We call $R = \{0\}$, with the operations $+$ and $\times$ defined in the only possible way, the *zero ring*.

For a ring $R \neq \{0\}$, $(R, \times)$ is *not* a group. To see this, note that $R \neq \{0\}$ implies $0 \neq 1$. Then $a0 = 0$ for all $a \in R$. Thus there is no $a \in R$ such that $a0 = 1$. This means 0 does not have a multiplicative inverse. So $(R, \times)$ is not a group.

▶ **Let R be a ring. We make the following definitions:**

(i) **$a \in R$ is a *unit* if there is a $b \in R$ such that $ab = 1$. We write $b = a^{-1}$.**

(ii) **$a \in R$ is a *zero divisor* if there is a non-zero $b \in R$ such that $ab = 0$.**

The following is a straightforward verification from the definitions.

▶ **Let R be a ring and define $R^\times = \{a \in R : \ a$ is a unit$\}$. Then $(R^\times, \times)$ is a group.**

Note this notation agrees with our earlier notations for $\mathbb{C}^\times$, $\mathbb{Q}^\times$, $\mathbb{R}^\times$ and $\mathbb{Z}_n^\times$.

Here is an example of a non-commutative ring. Consider the set of 2×2 real matrices:

$$\mathrm{M}_2(\mathbb{R}) = \left\{ A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} : \ \ a, b, c, d \in \mathbb{R} \right\}$$

Then we define $+$ to be addition of matrices and $\times$ to be multiplication of matrices. Then, not surprisingly, $\mathrm{M}_2(\mathbb{R})$ is a ring with additive and multiplicative identities given by:

$$\text{“0”} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \qquad\qquad \text{“1”} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

An example of a zero divisor in $\mathrm{M}_2(\mathbb{R})$ (which is not 0) is the following matrix:

$$A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$$

In fact $AA = A^2 = 0$. Finally, we note that $A \in \mathrm{M}_2(\mathbb{R})$ is a unit if and only if $A \in \mathrm{GL}_2(\mathbb{R})$. We conclude $\mathrm{M}_2(\mathbb{R})^\times = \mathrm{GL}_2(\mathbb{R})$.

▶ **Let R be a ring. We make the following definitions:**

(i) **R is an *integral domain* if it is commutative and $ab = 0$ implies $a = 0$ or $b = 0$.**

(ii) **R is a *division ring* if every non-zero $a \in R$ is a unit.**

(iii) R is a *field* if it is a commutative division ring.

Note a commutative ring R is an integral domain if and only if the only zero divisor in R is 0. Also, every field is an integral domain. Note also that R is a division ring if and only if the group of units $R^\times$ is exactly $R \setminus \{0\}$.

Examples of fields are $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$. The ring $\mathbb{Z}$ is an integral domain but not a division ring or a field. The ring $\mathrm{M}_2(\mathbb{R})$ is not commutative (hence not an integral domain or a field) and also not a division ring, because it has non-zero zero divisors.

Let's look at some other examples. Consider $\mathbb{Z}_3 = \{0, 1, 2\}$. Note $1 \cdot 1 = 1$ and $2 \cdot 2 = 1$. Thus $\mathbb{Z}_3$ is an integral domain, and even a field. For another example, consider $\mathbb{Z}_4$. Note that $2 \cdot 2 = 0$, so this ring has a non-zero zero divisor, and is not an integral domain.

Now consider $\mathbb{Z}_n$ for n a general positive integer $n > 1$. We know that

$$\mathbb{Z}_n^\times = \{\text{units in } \mathbb{Z}_n\} = \{a (\mathrm{mod}\ n) : \ \gcd(a, n) = 1\}$$

The only case in which $\mathbb{Z}_n^\times = \mathbb{Z}_n \setminus \{0\}$ is when n is a prime. Thus the commutative ring $\mathbb{Z}_n$ is a field if and only if n is prime.

Furthermore, suppose n is not prime, and write $n = ab$ for some positive integers a, b less than n. Then $a, b \pmod n$ are non-zero, but $ab \equiv n \equiv 0 \pmod n$. Thus $a, b \pmod n$ are zero divisors in $\mathbb{Z}_n$. Thus $\mathbb{Z}_n$ is not even an integral domain when n is not prime. Thus:

▶ **If $n > 1$ is prime, the ring $\mathbb{Z}_n$ is a field. If n is not prime, $\mathbb{Z}_n$ has (non-zero) zero divisors and so is not an integral domain.**

More ring basics

In this lecture we continue our study of the basic properties of rings.

Let us begin with an example. The ring of quaternions $\mathbb{H}$ is the set of expressions

$$x = a + b\mathbf{i} + c\mathbf{j} + d\mathbf{k}$$

where $a, b, c, d \in \mathbb{R}$, and where addition and multiplication are done in a similar fashion to the complex numbers, but where we have the relations $\mathbf{i}^2 = \mathbf{j}^2 = \mathbf{k}^2 = -1$ and $\mathbf{ij} = \mathbf{k}$. Note a consequence is that $\mathbf{ij} = -\mathbf{ji}$, $\mathbf{jk} = -\mathbf{kj}$, $\mathbf{ki} = -\mathbf{ik}$. In particular, $\mathbb{H}$ is not a commutative ring.

For another example, consider the quaternions $x = 1 + 2\mathbf{j}$ and $y = \mathbf{i} - \mathbf{k}$. Then

$$xy = (1 + 2\mathbf{j})(\mathbf{i} - \mathbf{k}) = (\mathbf{i} - \mathbf{k}) + 2\mathbf{j}(\mathbf{i} - \mathbf{k}) = \mathbf{i} - \mathbf{k} - 2\mathbf{k} - 2\mathbf{i} = -\mathbf{i} - 3\mathbf{k}$$

$$yx = (\mathbf{i} - \mathbf{k})(1 + 2\mathbf{j}) = \mathbf{i}(1 + 2\mathbf{j}) - \mathbf{k}(1 + 2\mathbf{j}) = \mathbf{i} + 2\mathbf{k} - \mathbf{k} + 2\mathbf{i} = 3\mathbf{i} + \mathbf{k}$$

The *norm* of a quaternion $x = a + b\mathbf{i} + c\mathbf{j} + d\mathbf{k}$ is given by the non-negative real number

$$|x| = \sqrt{a^2 + b^2 + c^2 + d^2}$$

The *conjugate* of $x \in \mathbb{H}$ is defined by $\overline{x} = a - b\mathbf{i} - c\mathbf{j} - d\mathbf{k}$. We compute

$$x\overline{x} = (a + b\mathbf{i} + c\mathbf{j} + d\mathbf{k})(a - b\mathbf{i} - c\mathbf{j} - d\mathbf{k}) = a^2 + b^2 + c^2 + d^2 = |x|^2$$

As a consequence, if $x \neq 0$ (so that $|x| \neq 0$), we have $xy = 1$ where $y = (\overline{x}/|x|^2)$. Clearly y is a quaternion, and we have found a multiplicative inverse for every nonzero $x \in \mathbb{H}$. Thus:

▶ **The ring of quaternions $\mathbb{H}$ is a non-commutative division ring.**

Another example of a non-commutative ring that we saw earlier was the ring of 2×2 matrices. In fact the quaternions fit into this framework as we will see shortly. For now we continue to introduce fundamental notions in ring theory.

A *subring* of a ring R is a subset $S \subset R$ which contains $0, 1$ and with the operations $+$ and $\times$ inherited from R is a ring in its own right. Note that a subring must be closed under the operations $+$ and $\times$. The following is a simple test of whether a subset is a subring.

▶ **$S \subset R$ is a subring if and only if the multiplicative identity 1 is in S and for all $a, b \in S$ we have $ab \in S$ and $a - b \in S$.**

The proof is straightforward and omitted. For example, the inclusions $\mathbb{Z} \subset \mathbb{Q} \subset \mathbb{R} \subset \mathbb{C} \subset \mathbb{H}$ exhibit a chain of subrings, each one contained in the next.

A *homomorphism* $\phi : R \to R'$ is a map of sets which satisfies $\phi(1) = 1$[1] and for all $a, b \in R$:

$$\phi(ab) = \phi(a)\phi(b), \qquad \phi(a + b) = \phi(a) + \phi(b)$$

[1]Some references do not require this condition.

A homomorphism of rings is an *isomorphism* if it is 1-1 and onto.

The map $\phi : R \to R$ given by $\phi(a) = a$ for all $a \in R$ is a homomorphism, called the *identity homomorphism*. The map $\phi : \mathbb{Z} \to \mathbb{Z}_n$ which is reduction mod n is an onto ring homomorphism.

For a more interesting example, let us define a map

$$\phi : \mathbb{H} \longrightarrow \mathrm{M}_2(\mathbb{C})$$

where the ring on the right consists of 2×2 complex matrices. We define ϕ by

$$\phi(\mathbf{i}) = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}, \qquad \phi(\mathbf{j}) = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \qquad \phi(\mathbf{k}) = \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix},$$

and of course $\phi(1)$ is the identity matrix. These relations determine ϕ completely if extend linearly over the real numbers. Explicitly, this means that for a general quaternion,

$$\phi(a + b\mathbf{i} + c\mathbf{j} + d\mathbf{k}) = a\phi(1) + b\phi(\mathbf{i}) + c\phi(\mathbf{j}) + d\phi(\mathbf{k}) = \begin{pmatrix} a + bi & c + di \\ -c + di & a - bi \end{pmatrix}$$

From this we have $\phi(x + y) = \phi(x) + \phi(y)$ for all $x, y \in \mathbb{H}$. To show $\phi(xy) = \phi(x)\phi(y)$ is a straightforward computation. Note that special cases of this include the fact that the matrices $\phi(\mathbf{i})^2$, $\phi(\mathbf{j})^2$, $\phi(\mathbf{k})^2$ are minus the identity matrix, and also

$$\phi(\mathbf{ij}) = \phi(\mathbf{k}) = \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix} = \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix} \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} = \phi(\mathbf{i})\phi(\mathbf{j})$$

You can check that ϕ is 1-1, but it is not onto.

▶ **Let $\phi : R \to S$ be a ring homomorphism. Then**

1. **The map ϕ restricts to a group homomorphism $\phi^\times : (R^\times, \times) \to (S^\times, \times)$.**
2. **If ϕ is an isomorphism then $\phi^\times$ is an isomorphism of groups.**
3. **If ϕ is an isomorphism, and R is commutative, then so is S.**

It also easily follows that if ϕ is an isomorphism and R is an integral domain (resp. division ring, field) then S is an integral domain (resp. division ring, field).

Consider the ring homomorphism $\phi : \mathbb{H} \to \mathrm{M}_2(\mathbb{C})$ from above. We obtain a homomorphism of groups $\phi^\times : \mathbb{H}^\times \to \mathrm{GL}_2(\mathbb{C})$. It is instructive to consider the subgroup

$$G = \{x \in \mathbb{H}^\times : \ |x| = 1\}$$

of unit quaternions, with quaternion multiplication. Note that this group $G \subset \mathbb{H}^\times$ is in 1-1 correspondence with the 3-dimensional sphere

$$S^3 = \{(a, b, c, d) \in \mathbb{R}^4 : a^2 + b^2 + c^2 + d^2 = 1\} \subset \mathbb{R}^4$$

Indeed, the quaternion $x \in G$ where $x = a + b\mathbf{i} + c\mathbf{j} + d\mathbf{k}$ corresponds to the point $(a, b, c, d) \in S^3$. Thus we have defined a group structure on the 3-sphere!

This is analogous to the unit complex numbers $U(1) \subset \mathbb{C}^\times$ being in 1-1 correspondence with the unit circle S^1 in $\mathbb{R}^2$, which is defined to be

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

In this case the unit complex number $z = x + yi$ corresponds to $(x, y) \in S^1$. Thus complex multiplication defines a group structure on the "1-sphere" S^1 in this way.

More generally we can define the n-sphere to be the following subset of points in $\mathbb{R}^{n+1}$:

$$S^n = \{(x_1, \ldots, x_{n+1}) \in \mathbb{R}^{n+1} : \ x_1^2 + \cdots + x_{n+1}^2 = 1\}$$

A natural question arises: which spheres S^n have group structures? To put this question on more firm footing we require that the group operation $S^n \times S^n \to S^n$ is the restriction of a differentiable function $\mathbb{R}^{n+1} \times \mathbb{R}^{n+1} \to \mathbb{R}^{n+1}$.

We have seen such group structures on S^1 and S^3 above, using complex and quaternion multiplication respectively. A somewhat uninteresting example is that of the 0-sphere: $S^0 = \{1, -1\} \subset \mathbb{R}^1$ is a group, in fact a subgroup of $\mathbb{R}^\times$. Remarkably:

▶ **The only spheres that admit (differentiable) group structures are S^0, S^1, S^3.**

The proof is outside the scope of this class and is a result in the theory of Lie groups.

Kernels, ideals and quotient rings

In this lecture we continue our study of rings and homomorphisms, with an emphasis on the notions of kernel, ideal and quotient ring.

Let $\phi : R \to R'$ be a ring homomorphism. We define the *kernel* of ϕ as follows:

$$\ker(\phi) = \{a \in R : \ \phi(a) = 0\}$$

Note that $\ker(\phi)$ is a subset of the ring R. However, $\phi(1) = 1$, so the only way $1 \in \ker(\phi)$ is if $1 = 0$ in R', i.e. $R' = \{0\}$. Thus in general $\ker(\phi)$ is not a subring of R. However, the kernel of a homomorphism does have the following kind of structure.

▶ **An *ideal* in a ring R is a subset $I \subset R$ satisfying the following: I is a subgroup of R with respect to addition, and for all $a \in I$, $r \in R$ we have $ra \in I$ and $ar \in I$.**

Note that an ideal is also closed under multiplication. However, it is important to note that the identity $1 \in R$ may not be in an ideal. In fact, if $1 \in I$ then for every $r \in R$ we have $r1 = r \in I$, so $R \subset I$. In conclusion, $I = R$ if and only if $1 \in I$.

In general, to check that a non-empty subset $I \subset R$ is an ideal, it suffices to show that (i) for all $a, b \in I$ we have $a + b \in I$, and (ii) for all $a \in I$ and $r \in R$ we have $ra \in I$ and $ar \in I$. Note in (ii) that if R is commutative, you only need to check $ra \in I$ since $ar = ra$.

▶ **Let $\phi : R \to R'$ be a ring homomorphism. Then $\ker(\phi) \subset R$ is an ideal.**

Proof. As $\phi(0) = 0$ we have $\ker(\phi) \neq \emptyset$. Let $a, b \in \ker(\phi)$, i.e. $\phi(a) = \phi(b) = 0$. Then

$$\phi(a + b) = \phi(a) + \phi(b) = 0 + 0 = 0$$

and thus $a + b \in \ker(\phi)$. Next, let $a \in \ker(\phi)$ and $r \in R$. Then

$$\phi(ra) = \phi(r)\phi(a) = \phi(r)0 = 0$$

and it follows that $ra \in \ker(\phi)$. Similarly $ar \in \ker(\phi)$. Thus $\ker(\phi)$ is an ideal in R. □

Examples

1. Consider the homomorphism $\phi_n : \mathbb{Z} \to \mathbb{Z}_n$ which is reduction mod n. Then $\phi_n(k) = k \pmod n$ is zero in $\mathbb{Z}_n$ if and only if k is a multiple of n. Thus

$$\ker(\phi_n) = \{an : \ a \in \mathbb{Z}\} = n\mathbb{Z} \subset \mathbb{Z}$$

In fact every ideal in $\mathbb{Z}$ is of the form $n\mathbb{Z}$, which is a good exercise for you to check.

2. For any ring R, consider the "zero" homomorphism $\phi : R \to \{0\}$ which sends everything to 0. Then we have $\ker(\phi) = R$.

3. At the other extreme, we remark that a homomorphism $\phi : R \to R'$ is 1-1 if and only if $\ker(\phi) = \{0\}$. Indeed, if ϕ is 1-1, then $\phi(a) = 0 = \phi(0)$ implies $a = 0$, so $\ker(\phi)$. Conversely, suppose $\ker(\phi) = \{0\}$. Then $\phi(a) = \phi(b)$ implies $0 = \phi(a) - \phi(b) = \phi(a - b)$, so that $a - b \in \ker(\phi) = \{0\}$. We obtain $a - b = 0$, i.e. $a = b$, and thus ϕ is 1-1.

4. Let us consider a more interesting example. Consider the following set:

$$R = \mathbb{Z}[\sqrt{-3}] = \left\{a + b\sqrt{-3} : \ a, b \in \mathbb{Z}\right\} \subset \mathbb{C}$$

This is a subring of the complex numbers. Indeed, we have $1 \in R$, and if $a + b\sqrt{-3}$ and $c + d\sqrt{-3} \in R$, then we have $(a + b\sqrt{-3}) - (c + d\sqrt{-3}) = (a - c) + (b - d)\sqrt{-3} \in R$, and also

$$(a + b\sqrt{-3})(c + d\sqrt{-3}) = (ac - 3bd) + (ad + bc)\sqrt{-3} \in R.$$

Having verified that R is a ring, we now define a map

$$\phi : \mathbb{Z}[\sqrt{-3}] \longrightarrow \mathbb{Z}_4$$

as follows: $\phi(a+b\sqrt{-3}) = a+b \pmod 4$. We claim this is a homomorphism. First, we note $\phi(1) = 1 \pmod 4$. Next, we show $\phi(x + y) = \phi(x) + \phi(y)$ for all $x, y \in R$:

$$\begin{aligned}\phi\left((a + b\sqrt{-3}) + (c + d\sqrt{-3})\right) &= \phi\left((a + c) + (b + d)\sqrt{-3})\right) = (a + c) + (b + d) \\ &= (a + b) + (c + d) = \phi(a + b\sqrt{-3}) + \phi(c + d\sqrt{-3}) \pmod 4\end{aligned}$$

Finally, to verify the property $\phi(x)\phi(y) = \phi(xy)$ for all $x, y \in R$ we compute:

$$\begin{aligned}\phi\left((a + b\sqrt{-3})(c + d\sqrt{-3})\right) &= \phi\left((ac - 3bd) + (ad + bc)\sqrt{-3}\right) = ac - 3bd + ad + bc \pmod 4 \\ \phi(a + b\sqrt{-3})\phi(c + d\sqrt{-3}) &= (a + b)(c + d) = ac + bd + ad + bc \pmod 4\end{aligned}$$

and they agree modulo 4. Thus ϕ is a ring homomorphism, and it is onto. The kernel is:

$$\ker(\phi) = \{a + b\sqrt{-3} : \ a + b \equiv 0 \pmod 4\} \subset R$$

Quotient rings

Let $I \subset R$ be an ideal in a ring R. Consider the additive cosets $R/I = a + I$ for $a \in R$. In other words, viewing I as a subgroup of the abelian group $(R, +)$ we are taking the quotient group R/I. We define a multiplication on these cosets: for $a + I, b + I \in R/I$ we define

$$(a + I)(b + I) = ab + I$$

This is well-defined because I is an ideal: if $a' + I = a + I$ and $b' + I = b + I$ then we have $a' - a \in I$, $b' - b \in I$. So $a'b' - ab = (a' - a)b + a'(b' - b)$ is in I. Note we are using that $(a' - a)b \in I$ because $a' - a \in I$ and $b \in R$, and similarly for the other term. Thus

$$(a' + I)(b' + I) = a'b' + I = ab + I = (a + I)(b + I)$$

The additive identity in R/I is the coset I, while the multiplicative identity is $1 + I$. The set R/I with the above described structure satisfies the axioms of a ring, and is called the *quotient ring* of R by the ideal I.

Given any ideal $I \subset R$ in a ring there is a canonical ring homomorphism

$$\phi : R \longrightarrow R/I$$

This homomorphism is onto, and $\ker(\phi) = I$. Conversely, every onto homomorphism can be viewed as such a homomorphism to a quotient ring. Explicitly:

▶ **1st Isomorphism Theorem**: **Let $\phi : R \to R'$ be an onto ring homomorphism. Then we have an isomorphism of rings $R/\ker(\phi) \cong R'$.**

The proof is similar to the proof of the 1st isomorphism theorem for groups.

As an example, consider the homomorphism $\phi : \mathbb{Z}[\sqrt{-3}] \to \mathbb{Z}_4$ we defined earlier. This is onto and its kernel $I \subset \mathbb{Z}[\sqrt{-3}]$ is the set of $a + b\sqrt{-3}$ such that $a + b \equiv 0 \pmod 4$. We obtain $\mathbb{Z}[\sqrt{-3}]/I \cong \mathbb{Z}_4$ as rings.

We also have analogues of the 2nd and 3rd isomorphism Theorems, now for rings:

▶ **2nd Isomorphism Theorem**: **Let S be a subring of R and $I \subset R$ an ideal of R. Then $S \cap I$ is an ideal in S and we have an isomorphism of rings:**

$$\frac{S}{S \cap I} \cong \frac{S + I}{I}$$

In this statement, $S + I = \{a + b : a \in S, b \in I\}$ is a subring of R.

▶ **3rd Isomorphism Theorem**: **Let R be a ring and I, J ideals in R with $J \subset I$. Then we have an isomorphism of rings:**

$$\frac{R}{I} \cong \frac{R/J}{I/J}$$

The proofs are similar in spirit to the ones for groups, and we omit them.

Principal ideals

In this lecture we continue studying ideals, focusing on the new concept of a principal ideal.

First we introduce some notation. Let R be a commutative ring and $a \in R$. We write

$$aR = \{ar : \ r \in R\} \subset R$$

Then aR is an ideal in R. To check this, note $0 = a0 \in aR$; if $ar, ar' \in (a)$ then $ar - ar' = a(r - r') \in aR$; and if $ar \in aR$ and $r' \in R$ then $r'(ar) = a(rr') \in aR$. The ideal $aR \subset R$ is called the *principal ideal* generated by $a \in R$. Another common notation for aR is (a).

Example: The ideals $(n) = n\mathbb{Z} \subset \mathbb{Z}$ are principal ideals.

Example: Let R be any commutative ring. Then $0R = (0) = \{0\}$ is the zero ideal, and $1R = (1) = R$ is the ideal which is the whole ring. These are both principal ideals.

▶ **Let R be a commutative ring.**

(i) If $I \subset R$ is an ideal and $a \in I$ then $(a) \subset I$.

(ii) If $a = ub$ for u a unit, then $(a) = (b)$.

These are straightforward from the definitions. For example, let us prove (ii). If $a = ub$ where u is a unit, then if $ar \in aR$ we have $ar = (ub)r = b(ur) \in bR$, so $(a) \subset (b)$. And if $br \in (b)$ then $br = (u^{-1}a)r = a(u^{-1}r) \in (a)$, so $(b) \subset (a)$. We conclude $(a) = (b)$.

Example: Consider the ring $R = \mathbb{Z}[\sqrt{-3}]$. Last lecture we defined a homomomorphism

$$\phi : R \longrightarrow \mathbb{Z}_4$$

$$\phi(a + b\sqrt{-3}) = a + b \pmod 4.$$

Consider the ideal $\ker(\phi)$. Note that $\phi(1 - \sqrt{-3}) = 1 - 1 = 0 \pmod 4$, so $1 - \sqrt{-3} \in \ker(\phi)$. We obtain an inclusion of ideals $(1 - \sqrt{-3}) \subset \ker(\phi)$.

Is it also true that $\ker(\phi) \subset (1 - \sqrt{-3})$? To rephrase this question, let $a + b\sqrt{-3} \in \ker(\phi)$, i.e. $a + b \equiv 0 \pmod 4$. Then, can we write $a + b\sqrt{-3}$ as a multiple of $1 - \sqrt{-3}$, i.e

$$a + b\sqrt{-3} = (1 - \sqrt{-3})(c + d\sqrt{-3})$$

for some $c, d \in \mathbb{Z}$? Note that the right hand side of this last equation becomes

$$(c + 3d) + (d - c)\sqrt{-3}$$

and so we must have $a = c + 3d$ and $b = d - c$. Solving for c, d we get $c = (a - 3b)/4$, $d = (a + b)/4$ which are integers because $a + b$ is a multiple of 4. Thus the answer is "yes", so that $\ker(\phi) \subset (1 - \sqrt{-3})$. Consequently we have

$$\ker(\phi) = (1 - \sqrt{-3})$$

Thus the kernel of this homomorphism is a principal ideal, generated by $1 - \sqrt{-3}$.

The following gives an example of an ideal which is not principal.

Example (non-principal ideal): We modify our homomorphism from above: define

$$\psi : R = \mathbb{Z}[\sqrt{-3}] \longrightarrow \mathbb{Z}_2$$

$$\psi(a + b\sqrt{-3}) = a + b \pmod 2.$$

Note again $1 - \sqrt{-3} \in \ker(\psi)$. We will argue that $\ker(\psi)$ is *not* principal. Suppose for a contradiction that $\ker(\psi)$ is principal, i.e. there is some for some $a + b\sqrt{-3} \in R$ such that

$$\ker(\psi) = (a + b\sqrt{-3})$$

In other words, $\ker(\psi)$ is the principal ideal generated by $a + b\sqrt{-3}$. Necessarily $a + b \equiv 0 \pmod 2$. Now since $1 - \sqrt{-3} \in \ker(\psi)$, we must have

$$1 - \sqrt{-3} = (a + b\sqrt{-3})(c + d\sqrt{-3})$$

for some $c, d \in \mathbb{Z}$. Note that since $\sqrt{-3}$ appears on the left side of this equation we can assume b and d are not both zero, for if they are both zero then the right hand side is a real number. Take the squared norms of both sides to obtain

$$4 = |1 - \sqrt{-3}|^2 = |a + b\sqrt{-3}|^2 |c + d\sqrt{-3}|^2 = (a^2 + 3b^2)(c^2 + 3d^2)$$

Noting that $a, b, c, d \in \mathbb{Z}$, $a + b \equiv 0 \pmod 2$, and one of b, d is non-zero, we must have $a, b, c \in \{1, -1\}$ and $d = 0$; or $a \in \{2, -2\}$, $c \in \{1, -1\}$ and $b = d = 0$. The only choices among these possibilities which solve the equation $1 - \sqrt{-3} = (a + b\sqrt{-3})(c + d\sqrt{-3})$ are $a = c = 1, b = -1, d = 0$ and $a = c = -1, b = 1, d = 0$. We conclude

$$a + b\sqrt{-3} = \pm(1 - \sqrt{-3})$$

Thus under our current assumption that $\ker(\psi)$ is principal, we obtain

$$\ker(\psi) = (1 - \sqrt{-3}) = \ker(\phi)$$

However, note that $\psi(2) = 0 \pmod 2$ so that $2 \in \ker(\psi)$, while $\phi(2) = 2 \not\equiv 0 \pmod 4$, so $2 \notin \ker(\phi)$. We conclude that $\ker(\psi) \neq \ker(\phi)$. We have a contradiction. Thus $\ker(\psi)$ cannot be a not principal ideal.

Principal ideal domains

Last lecture we introduced the notion of a *principal ideal* in a commutative ring. These are ideals in a commutative ring R that take the form

$$(a) = aR = \{ar : \ r \in R\}$$

A *principal ideal domain* (PID) is an integral domain R (a commutative ring such that $ab = 0$ implies $a = 0$ or $b = 0$) such that every ideal in R is principal.

The ring of integers $\mathbb{Z}$ is the most basic example of a PID. The ideals in $\mathbb{Z}$ are

$$(0), \quad (1), \quad (2), \quad (3), \quad (4), \quad \ldots$$

PIDS are an important class of rings because they occur frequently and the theory of PIDS is considerably simpler than that of general (commutative) rings.

▶ **Let R be a non-zero commutative ring. The following are equivalent:**

(i) **R is a field.**

(ii) **The only ideals in R are $(0) = \{0\}$ and $(1) = R$.**

(iii) **Every homomorphism $\phi : R \to R'$ where $R' \neq \{0\}$ is 1-1.**

Proof. We show (i) implies (ii). Assume (i), i.e. R is a division ring. Consider an ideal $I \subset R$ with $I \neq (0)$. Choose $a \in I$, $a \neq 0$. Since R is a division ring, $a^{-1} \in R$. Then $1 = a^{-1}a \in I$ as $a^{-1} \in R$ and $a \in I$. Now for any $b \in R$, we have $b = b1 \in I$ as $b \in R$ and $1 \in I$. Thus (ii) holds.

Next, we show (ii) implies (iii). Assume (ii): the only ideals in R are $\{0\}$ and R. Consider a homomorphism $\phi : R \to R'$ where $R' \neq \{0\}$. Then $\ker(\phi) \subset R$ is a proper ideal as $1 \notin \ker(\phi)$. By our assumption it must be $\{0\}$. This is equivalent to ϕ being 1-1. Thus (iii) holds.

Finally, we show (iii) implies (i). Assume (iii), i.e. every homomorphism $\phi : R \to R'$ where $R' \neq \{0\}$ is 1-1. Now take $a \in R$ with $a \neq 0$. Consider the natural homomorphism $\phi : R \to R/(a)$. Suppose $R/(a) \neq \{0\}$, so that ϕ is 1-1. Then $\ker(\phi) = (0)$. On the other hand, $\ker(\phi) = (a)$. Then $(a) = (0)$ implies $a = 0$, a contradiction. Thus $R/(a) = \{0\}$, impying $(a) = R = (1)$. In particular, $1 = ar$ for some $r \in R$, so a has a multiplicative inverse. We have shown that every non-zero $a \in R$ is invertible, and thus R is a field. □

A corollary of this result is the following:

▶ **If R is a field, then R is a PID.**

This holds simply because the only ideals in a field R are the ideals $(0) = \{0\}$ and $(1) = R$ which are principal ideals.

Thus the fields $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$ are all examples of PIDs.

A more exotic example of a PID is the ring of Gaussian integers

$$\mathbb{Z}[i] = \{a + bi : \ a, b \in \mathbb{Z}\} \subset \mathbb{C}$$

We will show this is a PID later.

We saw last lecture that the ring $\mathbb{Z}[\sqrt{-3}]$ is *not* a PID, because we showed that the ideal consisting of $a + b\sqrt{-3}$ with $a \equiv b \pmod 2$ is not principal.

Polynomial rings

Another important example involves the following construction. Let R be any ring. Define

$$R[x] = \{\text{polynomials in } x \text{ with coefficients in } R\}$$

That is, a typical element in $R[x]$ is an expression of the form

$$f(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$$

where n is a non-negative integer and $a_0, \ldots, a_n$ are elements of the ring R. The sum $f(x) + g(x)$ and product $f(x)g(x)$ of two such polynomials are defined in the usual fashion, and this makes $R[x]$ into a ring. If R is commutative, so is $R[x]$. Also, if R is an integral domain, so too is the polynomial ring $R[x]$.

▶ **If R is a field, then $R[x]$ is a PID.**

Before explaining the proof, let's explore the consequences of this result.

The rings $\mathbb{Q}[x], \mathbb{R}[x], \mathbb{C}[x]$ are all PIDS. This means, for example, that for any ideal $I \subset \mathbb{Q}[x]$ there exists a polynomial $f(x)$ with rational coefficients such that $I = (f(x))$. The polynomial $f(x)$ is unique up to multiplication by an element in $\mathbb{Q}^\times$.

Let p be a prime. Then $\mathbb{Z}_p$ is a field, so $\mathbb{Z}_p[x]$ is a PID. For example, consider $\mathbb{Z}_2[x]$, polynomials with coefficients in $\mathbb{Z}_2$. Some elements in this ring are

$$0, \quad 1, \quad x, \quad 1+x, \quad x^2, \quad 1+x^2, \quad 1+x+x^2, \quad \ldots$$

For any ideal $I \subset \mathbb{Z}_2[x]$ we can find some such element $f(x)$ such that $I = (f(x))$.

The condition that R be a field in the result is necessary. For example, consider $\mathbb{Z}[x]$. Of course $\mathbb{Z}$ is not a field, so the result does not apply here. In fact $\mathbb{Z}[x]$ is not a PID: the ideal

$$I = \{2f(x) + xg(x) : \ f(x), g(x) \in \mathbb{Z}[x]\} \subset \mathbb{Z}[x]$$

is not a principal ideal, as you can verify.

Now let us prove the statement. We will need the following:

▶ **(Division algorithm for polynomials) Let R be a field. Consider polynomials $f(x)$ and $g(x) \neq 0$ in $R[x]$. Then there exist unique $q(x), r(x) \in R[x]$ such that**

$$f(x) = g(x)q(x) + r(x)$$

and where either $\deg(r(x)) < \deg(g(x))$ or $r(x) = 0$.

Here the degree of $f(x) = \sum a_i x^i \in R[x]$ is the largest n such that $a_n \neq 0$. We omit the proof of the division algorithm and now show that if R is a field then $R[x]$ is a PID.

Let $I \subset R[x]$ be an ideal. We must show I is principal. If $I = \{0\}$ then it is the principal ideal (0), so assume $I \neq (0)$. Let $g(x) \in I$ be non-zero and of minimal possible degree among all non-zero polynomials in I. Note $(g(x)) \subset I$. Now consider any other element $f(x) \in I$. Then the division algorithm gives us $q(x), r(x) \in R[x]$ satisyfing

$$f(x) = g(x)q(x) + r(x)$$

and $\deg(r(x)) < \deg(g(x))$ or $r(x) = 0$. Suppose $r(x) \neq 0$. Then

$$r(x) = f(x) - g(x)q(x)$$

is in I, because $f(x), g(x) \in I$ and $q(x) \in R[x]$. Furthermore, $\deg(r(x)) < \deg(g(x))$. But this contradicts our assumption that $g(x)$ has the minimal possible degree among non-zero polynomials in I. So we must have $r(x) = 0$. Then

$$f(x) = q(x)q(x)$$

This shows $f(x) \in (g(x))$. Thus $I = (g(x))$, and I is a principal ideal. This completes the proof that all ideals in $R[x]$ are principal.

Finally, we remark that if you continue to add variables to your ring, and consider polynomials in several variables, you will not get a PID. For example, consider

$$\mathbb{Q}[x][y] = \mathbb{Q}[x, y] = \{\text{polynomials in } x, y \text{ with coefficients in } \mathbb{Q}\}$$

Then the ideal generated by the polynomials x and y, which is given by $I = \{xf(x, y) + yg(x, y) : \ f(x, y), g(x, y) \in \mathbb{Q}[x]\}$, is not a principal ideal in $\mathbb{Q}[x, y]$.

Rings and geometry

In this lecture we continue studying ideals in rings, and also explain the connection of ring theory to the geometry of algebraic varieties.

Let R be a commutative ring and consider the ring of polynomials $R[x, y]$. In fact for the time being we will take $R = \mathbb{R}$ to be the real numbers, so that we may easily draw pictures.

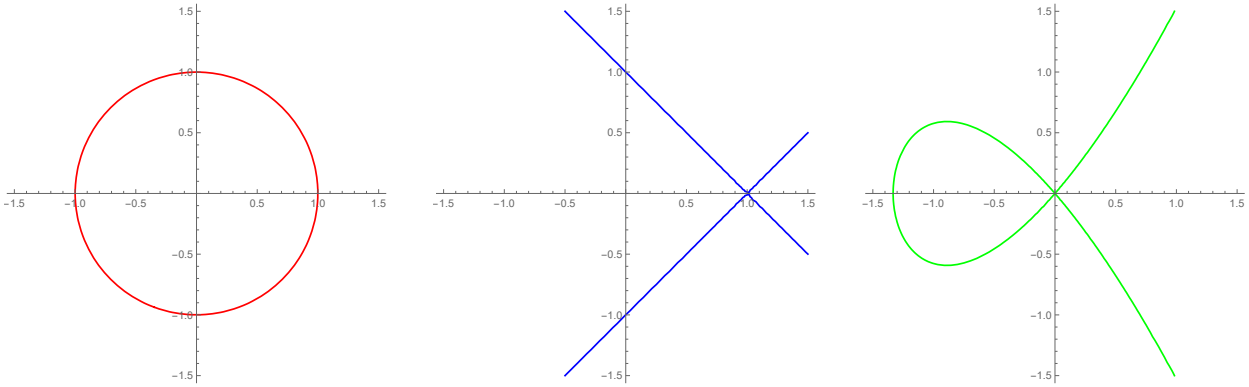
An (algebraic) *variety* in $\mathbb{R}^2$ is a set $V = V(S) \subset \mathbb{R}^2$ which consists of points (x, y) in the plane $\mathbb{R}^2$ that are the common roots of a given collection of polynomials $S \subset R[X]$:

$$V(S) = \{(x_0, y_0) \in \mathbb{R}^2 : f(x_0, y_0) = 0 \text{ for all } f \in S\} \subset \mathbb{R}^2$$

For example, consider the following subsets of $\mathbb{R}[x, y]$:

$$\begin{aligned} S_1 &= \{x^2 + y^2 - 1\} \\ S_2 &= \{(x + y - 1)(x - y - 1)\} \\ S_3 &= \{y^2 - x^2(x + \tfrac{4}{3})\} \\ S_4 &= \{x, y + 1\} \end{aligned}$$

The corresponding varieties $V(S_1)$, $V(S_2)$, $V(S_3)$ are shown below, while $V(S_4) = \{(0, -1)\}$:



Given a subset $S \subset R$ in any ring, we can also associate an ideal to S, called $I(S) \subset R$: it is the smallest ideal containing S. Explicitly, we can define this ideal as follows:

$$I(S) = \{\sum_{i=1}^{n} r_i f_i : \ r_i \in R, f_i \in S\}$$

Each of the examples above yields such an ideal:

$$\begin{aligned} I(S_1) &= (x^2 + y^2 - 1) \\ I(S_2) &= ((x + y - 1)(x - y - 1)) \\ I(S_3) &= (y^2 - x^2(x + \tfrac{4}{3})) \\ I(S_4) &= (x, y + 1) = (x) + (y + 1) \end{aligned}$$

(See below for the sum of ideals.) We have the following (note the reversal of inclusion!):

▶ **For subsets $S \subset T$ in $\mathbb{R}[x, y]$ we have $V(T) \subset V(S)$.**

Proof. Let $(x_0, y_0) \in V(T)$. This means $f(x_0, y_0) = 0$ for all polynomials $f \in T$. Since $S \subset T$ it is then certainly true that $f(x_0, y_0) = 0$ for all $f \in S$. Thus $(x_0, y_0) \in V(S)$. □

The next observation is that the variety corresponding to a set $S \subset \mathbb{R}[x, y]$ of polynomials is the same as the variety associated to the ideal $I(S)$.

▶ **For $S \subset \mathbb{R}[x, y]$ we have $V(S) = V(I(S))$.**

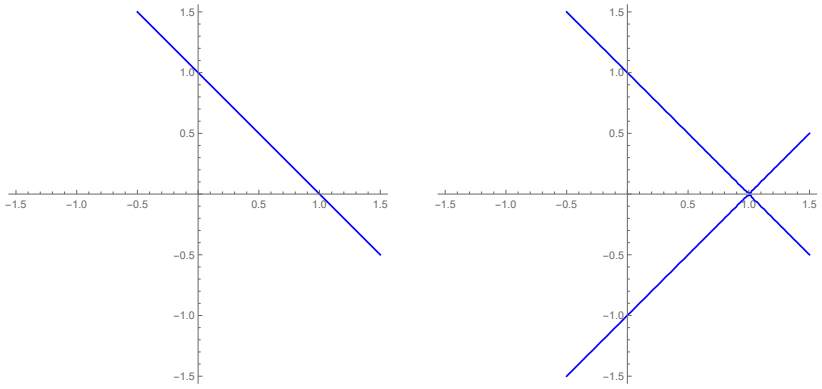
Proof. As $S \subset I(S)$, the previous result gives us $V(I(S)) \subset V(S)$.

For the reverse inclusion, suppose $(x_0, y_0) \in V(S)$. Then $f(x_0, y_0) = 0$ for all $f \in S$. Let $f \in I(S)$. Then $f = \sum_{i=1}^{n} r_i f_i$ for some $r_i \in \mathbb{R}[x, y]$ and $f_i \in S$. Then

$$f(x_0, y_0) = \sum_{i=1}^{n} r_i(x_0, y_0) f_i(x_0, y_0) = \sum_{i=1}^{n} r_i(x_0, y_0) 0 = 0$$

Thus $f(x_0, y_0) = 0$ for all $f \in I(S)$, and $(x_0, y_0) \in V(I(S))$. Thus $V(S) \subset V(I(S))$. □

For example, for the ideals $I = (x + y - 1)$ and $I_2 = I(S_2) = ((x + y - 1)(x - y - 1))$ we have $I_2 \subset I$ and so $V(I) \subset V(I_2)$. On the left below is the variety $V(I)$, on the right is $V(I_2)$.



Given ideals $I, J \subset R$ in a commutative ring, we define the sum, intersection, and product:

$$I + J = \{a + b : \ a \in I, b \in J\}$$
$$I \cap J = \{a : \ a \in I, a \in J\}$$
$$IJ = \{\sum a_i b_i : \ a_i \in I, b_j \in J\}$$

These are again ideals in R, as you can check. If $I = (a)$ and $J = (b)$ are principal, then in fact $IJ = (a)(b) = (ab)$. We now explore the geometric interpretations of these operations.

▶ **For ideals $I, J \subset \mathbb{R}[x, y]$ we have $V(I + J) = V(I) \cap V(J)$.**

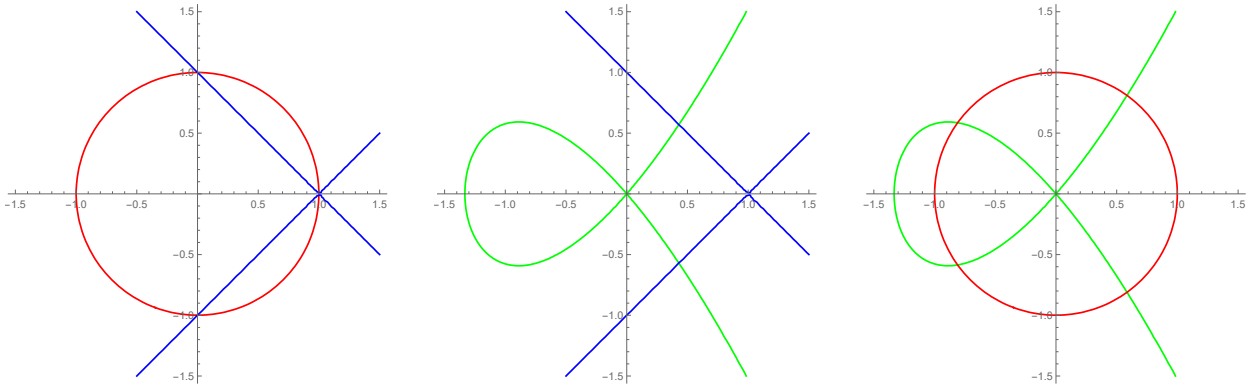
Proof. Let $(x_0, y_0) \in V(I) \cap V(J)$. Then $f(x_0, y_0) = 0$ for all $f \in I$ and $f \in J$. Let $f \in I + J$. Then $f = f' + f''$ where $f' \in I$, $f'' \in J$. We compute

$$f(x_0, y_0) = f'(x_0, y_0) + f''(x_0, y_0) = 0 + 0 = 0$$

and therefore $(x_0, y_0) \in V(I + J)$, and we conclude $V(I) \cap V(J) \subset V(I + J)$.

For the reverse inclusion, since $I \subset I + J$ we have $V(I + J) \subset V(I)$. Similarly, $V(I + J) \subset V(J)$. Thus $V(I + J) \subset V(I) \cap V(J)$. All together we obtain the statement. □

The conclusion we draw from this is that the sum of ideals corresponds to the *intersection* of varieties. For example, consider the ideals $I_1 = I(S_1), I_2 = I(S_2), I_3 = I(S_3)$ from earlier.



Then $I_1 + I_2$ is the ideal which is the sum of the principal ideals $(x^2 + y^2 - 1)$ and $((x + y - 1)(x - y - 1))$. It is not a principal ideal itself. However the variety $V(I_1 + I_2)$ is very simple:

$$V(I_1 + I_2) = V(I_1) \cap V(I_2) = \{(0, 1), (0, -1), (1, 0)\} \subset \mathbb{R}^2$$

It consists of the 3 points of intersection between the circle $V(I_1)$ and the two lines $V(I_2)$. Similarly, we see that $V(I_2 + I_3)$ is 2 points, and $V(I_3 + I_1)$ is 4 points.

▶ **For ideals $I, J \subset \mathbb{R}[x, y]$ we have $V(I \cap J) = V(I) \cup V(J)$.**

The proof is similar to the previous one. Thus the variety associated to the intersection of two ideals is the union of the two varieties of each ideal. As a consequence, $V(I_1 \cap I_2)$, $V(I_2 \cap I_3)$, $V(I_3 \cap I_1)$ are the three pictures above (not just the intersection points!).

In fact, the same property holds for the product of ideals!

▶ **For ideals $I, J \subset \mathbb{R}[x, y]$ we have $V(IJ) = V(I) \cup V(J)$.**

Note in general $IJ \neq I \cap J$. For example, consider $I = (x)$ And $J = (xy)$ in $\mathbb{R}[x, y]$. Then $IJ = (x^2 y)$ while $I \cap J = (xy) = J$. In general, $IJ \subset I \cap J$, but not the other way around.

Thus the above two results give instances of the phenomenon that $V(I) = V(J)$ for two ideals $I \neq J$. In other words, the map given by

$$\{\text{ideals in } \mathbb{R}[x,y]\} \longrightarrow \{\text{varieties in } \mathbb{R}^2\}$$

$$I \longmapsto V(I)$$

is not 1-1. (From our earlier discussion, it is onto.) In fact it is very far from being 1-1! Take any polynomial $f \in \mathbb{R}[x,y]$ which has no real solutions to $f(x,y) = 0$. Examples are $f(x,y) = 1$ and $f(x,y) = x^2 + 1$. Then for any such polynomial,

$$V(f) = \emptyset,$$

since $f(x_0, y_0) = 0$ does not hold for any $(x_0, y_0) \in \mathbb{R}^2$. However the ideals $(1) = \mathbb{R}[x,y]$ and $(1 + x^2)$ are not the same.

To fix this last sort of problem, we can work over a field such as $\mathbb{C}$ which has more solutions. Then for the ideal $(1 + x^2) \subset \mathbb{C}[x,y]$ the variety

$$V(1+x^2) = \{(x_0, y_0) \in \mathbb{C}^2 : 1 + x_0^2 = 0\} = \{(i, y_0) : y_0 \in \mathbb{C}\} \cup \{(-i, y_0) : y_0 \in \mathbb{C}\}$$

is two disjoint copies of $\mathbb{C}$ contained inside $\mathbb{C}^2$. Thus it is distinguished from $V(1) = \emptyset$.

Still, even after replacing $\mathbb{R}$ with $\mathbb{C}$, we do not have a 1-1 correspondence between ideals and varieties. To remedy this, we focus on a particular class of ideals which does make the assignment $I \mapsto V(I)$ into a 1-1 correspondence.

▶ **For any ideal $I \subset R$ in a commutative ring R, the *radical* of I is defined to be**

$$\sqrt{I} = \{a \in R : \ a^n \in I \text{ for some positive } n \in \mathbb{Z}\}$$

It is instructive to verify that $\sqrt{I} \subset R$ is an ideal. A *radical ideal* is an ideal of the form $\sqrt{I}$.

For example, consider $I = (x^2) \subset \mathbb{R}[x,y]$. Then $\sqrt{I} = (x)$. On the other hand, $\sqrt{(x)} = (x)$. A key property is the following: for ideals I, J we have $\sqrt{IJ} = \sqrt{I \cap J}$.

▶ **The following assignment is a 1-1 and onto correspondence:**

$$\{\text{radical ideals in } \mathbb{C}[x,y]\} \longrightarrow \{\text{varieties in } \mathbb{C}^2\}$$

$$I \longmapsto V(I)$$

This sets up a "dictionary" between certain ideals, which are algebraic objects, and varieties, which are geometric objects. The dictionary converts sums of ideals into intersections of varieties, and intersections of ideals into unions of varieties.

This lecture is essentially an introduction to the subject of *Algebraic Geometry*, which is the study of algebraic varieties in a much broader setting.

Prime and maximal ideals

An ideal I in a commutative ring R is *prime* if $I \neq R$ and $ab \in I$ implies that either $a \in I$ or $b \in I$. The ideal $I \subset R$ is *maximal* if $I \neq R$, and $I \subset J$ for some ideal $J \neq I$ implies $J = R$.

▶ **A maximal ideal is a prime ideal.**

Proof. Let I be a maximal ideal, and suppose it is not prime: let $ab \in I$ such that $a \notin I$ and $b \notin I$. Consider the ideals $(a) + I$ and $(b) + I$. These both contain I and are not equal to I. Since I is maximal, it follows that $(a) + I = R$ and $(b) + I = R$. The product ideal of $(a) + I$ and $(b) + I$ is then on the one hand equal to $(1) = R$ and on the other hand contained in I, implying $R = I$, a contradiction. □

The following characterizes prime and maximal ideals in terms of the quotient ring.

▶ **Let R be a commutative ring and $I \subset R$ a proper ideal. Then**

(i) I is prime if and only if the ring R/I is an integral domain.

(ii) I is maximal if and only if the ring R/I is a field.

Proof. (i) Suppose I is prime. Suppose $ab + I = (a + I)(b + I) = I$ in R/I. Then $ab \in I$. Since I is prime, one of a or b is in I. Then one of $a + I$ or $b + I$ is equal to I. This proves R/I is an integral domain. Conversely, suppose R/I is an integral domain. Now let $ab \in I$. Then $ab + I = (a + I)(b + I) = I$, and since R/I is an integral domain, one of $a + I$ or $b + I$ is equal to I. This implies one of a or b is in I. Thus I is prime.

(ii) Suppose I is maximal. We claim R/I is a field. This is equivalent to R/I having only the ideals $\{I\}$, R/I. Suppose R/I has a non-zero proper ideal J'. Then $J = \{a \in R : a + I \in J'\} \subset R$ is an ideal of R. To verify this: if $a, b \in J$, then $a + I, b + I \in J'$, and

$$(a + I) - (b + I) = (a - b) + I$$

is in J' because J' is an ideal; thus $a - b \in J$. Similarly, if $a \in J$ and $b \in R$ then $(a + I)(b + I) = ab + I$ is in J', hence $ab \in J$. Also, $0 \in J$. Thus J is an ideal. Next, since I is maximal and $I \subset J$, we must have either $J = I$ or $J = R$. If $J = I$ then $J' = \{I\}$, a contradiction. If $J = R$ then $J' = R/I$, contradicting our assumption that J' is proper.

Conversely, suppose R/I is a field. Suppose I is not maximal, and let $I \subset J$ with $I \neq J$ and $J \neq R$. Then consider $J' = \{a + I : a \in J\} \subset R/I$. This is an ideal of R/I. Since R/I is a field, it must be either zero or R/I. If J' is zero then $a + I = I$ for all $a \in J$ i.e. $a \in I$ for all $a \in J$; but this implies $J \subset I$, a contradiction. Thus $J' = R/I$. Then for all $a \in R$ we have $a + I = b + I$ for some $b \in J$, i.e. $a + c = b + d$ for some $c, d \in I$ and $b \in J$. Then $a = b + d - c \in J$. We have shown $J = R$, a contradiction. □

Examples

1. Consider an ideal $(n) \subset \mathbb{Z}$ where $n \geq 0$. Suppose it is a prime ideal. This means $ab \in (n)$ implies one of a or b is in (n). Note $ab \in (n)$ if and only if $ab = nk$ for some $k \in \mathbb{Z}$, i.e. n divides ab. So (n) is prime if and only if "n divides ab" implies "n divides a or b". For this to be true n must be a prime number. Thus the prime ideals of $\mathbb{Z}$ are

$$(2), \quad (3), \quad (5), \quad (7), \quad (11), \quad \ldots$$

We recover that $\mathbb{Z}_n$ is an integral domain if and only if n is prime.

We also know from earlier lectures that $\mathbb{Z}_n$ is a field if and only if n is prime. This implies that the prime ideals in $\mathbb{Z}$ are also the maximal ideals.

2. Consider the ring $\mathbb{C}[x, y]$. The ideal $I = (x^2 + y^2 - 1)$ is prime, but not maximal. To see it is prime, suppose $f(x, y)g(x, y) = (x^2 + y^2 - 1)h(x, y)$. Then since $x^2 + y^2 - 1$ is irreducible (cannot be factored over $\mathbb{C}[x, y]$) it must divide one of $f(x, y)$ or $g(x, y)$. To see that it is not a maximal ideal, note that we have an inclusion

$$(x^2 + y^2 - 1) \subset (x - 1, y)$$

because $x^2 + y^2 - 1 = (x + 1)(x - 1) + (y)(y)$. This inclusion is proper since $y \notin I$. Also, $(x - 1, y) \neq \mathbb{C}[x, y]$. Thus I is not maximal. However, $(x - 1, y)$ is maximal.

3. In the ring $\mathbb{Z}[x]$, the ideal (2) is prime but not maximal. To see this, consider

$$\phi : \mathbb{Z}[x] \longrightarrow \mathbb{Z}_2[x]$$

defined by $\phi(f(x)) = f(x) \pmod 2$, i.e. take the coefficients mod 2. Then ϕ is a homomorphism and the kernel is the principal ideal $(2) \subset \mathbb{Z}[x]$. By the 1st Isomorphism Theorem,

$$\mathbb{Z}[x]/\ker(\phi) = \mathbb{Z}[x]/(2) \cong \mathbb{Z}_2[x]$$

Now $\mathbb{Z}_2[x]$ is an integral domain, but not a field (e.g. x is not invertible). Thus (2) is a prime ideal that is not maximal. On the other hand, we may consider the homomorphism

$$\psi : \mathbb{Z}[x] \longrightarrow \mathbb{Z}_2$$

defined by $\psi(f(x)) = f(0) \pmod 2$. Then you can check that the kernel is the ideal $(2, x) = (2) + (x)$. Since $\mathbb{Z}_2$ is a field, this ideal is maximal.

Euclidean domains and UFDs

In this lecture we take a closer look at the Gaussian integers

$$\mathbb{Z}[i] = \{a + b\sqrt{-1} : \ a, b \in \mathbb{Z}\}$$

This ring appears to be closely related to the integers $\mathbb{Z}$; after all, we have only thrown in the element $i = \sqrt{-1}$. So how much of what is true for the ring of integers is true for the Gaussian integers? One of the most important properties of $\mathbb{Z}$ is the following:

▶ **Fundamental Theorem of Arithmetic (FTA): Let n be a positive integer. Then there are (not necessarily distinct) prime numbers $p_1, \ldots, p_m$ such that**

$$n = p_1 p_2 \cdots p_m$$

If $n = q_1 \cdots q_l$ for other primes q_i then $l = m$ and $q_i = p_i$ after re-labelling.

Can we obtain a similar statement for $\mathbb{Z}[i]$? One thing to note is the equation

$$(1 + i)(1 - i) = 1^2 - i^2 = 2$$

and $1 \pm i$ are not units in $\mathbb{Z}[i]$. In other words, "2" is not "prime" in this ring. Thus even if we do have some kind of FTA for $\mathbb{Z}[i]$, the factorizations will be different than those of $\mathbb{Z}$.

First we will formalize the kind of structure we are after. An *irreducible* element $a \in R$ in an integral domain R is an element that is non-zero and not a unit, and such that if $a = bc$, then one of b or c is a unit. Heuristically, a cannot be "decomposed" into smaller pieces. The irreducible elements of $\mathbb{Z}$ are exactly the prime numbers and their negatives.

▶ **An integral domain R is a *Unique Factorization Domain* (UFD) if every non-zero $a \in R$ is the product of (not necessarily distinct) irreducible elements $p_i \in R$:**

$$a = p_1 p_2 \cdots p_m$$

This representation of a is unique in the sense that if $a = q_1 \cdots q_n$ for irreducibles q_i, then after re-ordering the q_i's we have $m = n$ and $q_i = u_i p_i$ for some units u_i.

Thus a UFD is a ring in which we have a generalized version of the FTA.

Now we can rephrase our above question: is the ring $\mathbb{Z}[i]$ a UFD? This question is closely related to PID's because of the following.

▶ **If R is a PID then it is a UFD.**

For the sake of time we will omit the proof, although it is not beyond the scope of the course.

To get "unique factorization" for Gaussian integers, we may hope to show $\mathbb{Z}[i]$ is a PID.

In search of some inspiration, we return to our proof that for R a field, $R[x]$ is a PID. How did we accomplish this? The key to the proof was the division algorithm: for any $f(x), g(x) \in R[x]$ with $g(x)$ non-zero, there are $q(x), r(x) \in R[x]$ such that

$$f(x) = g(x)q(x) + r(x)$$

and such that either $r(x) = 0$ or $\deg(r(x)) < \deg(g(x))$. With this, given any non-zero ideal $I \subset R[x]$ we saw that for any polynomial $f(x) \in I$ that was of minimal degree among all non-zero elements in I, we have $I = (f(x))$, and thus I is principal.

The general structure of this proof can be abstracted. For a general integral domain R, what we need in order to abstract this proof is an analogue of the "degree" of elements.

▶ **An integral domain R is a *Euclidean Domain* if there is a function $N : R\backslash\{0\} \to \mathbb{Z}_{\geqslant 0}$ such that for all $a, b \in R$ with $b \neq 0$ there are $q, r \in R$ such that**

$$a = bq + r$$

where either $r = 0$ or $N(r) < N(b)$.

In other words, a Euclidean domain is an integral domain where we have an abstracted version of the division algorithm. The function N replaces "deg" in our example above.

▶ **If R is a Euclidean domain then R is a PID.**

Proof. As remarked above, the proof is the same as showing that the ring of polynomials over a field is a PID. First, we take a non-zero ideal $I \subset R$ in our Euclidean domain R. Then let $a \in I$ be such that $N(a)$ is minimal among all non-zero elements in I. Finally, you can show that $I = (a)$ just as in the polynomial case, replacing "deg" by "N" throughout. □

▶ **The ring of Gaussian integers $\mathbb{Z}[i]$ is a Euclidean domain.**

Proof. We must find a suitable function N from non-zero elements of $\mathbb{Z}[i]$ to the non-negative integers. For this function we take the complex norm squared:

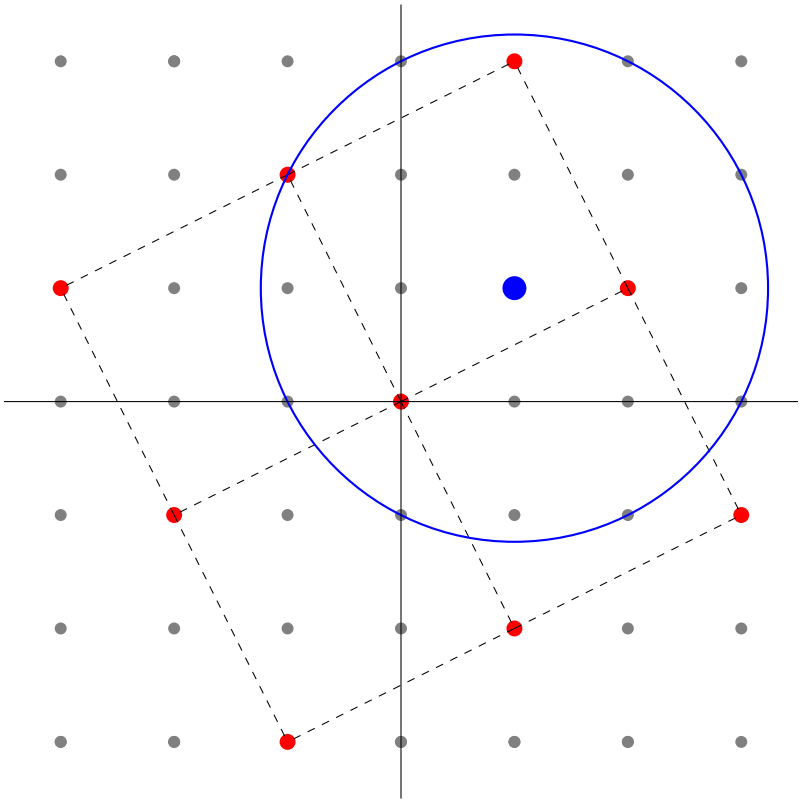
$$N(a + bi) = |a + bi|^2 = a^2 + b^2$$

Now we must show that given $x = a + bi$ and $y = c + di$ in $\mathbb{Z}[i]$ with $y \neq 0$ we can find $q, r \in \mathbb{Z}[i]$ such that $x = yq + r$ and either $r = 0$ or $N(r) < N(y)$. This last condition is equivalent to the condition $|r| < |y|$.

Thus what we are after is: given $x, y \in \mathbb{Z}[i]$, find some $q \in \mathbb{Z}[i]$ such that $|yq - x| < |y|$. For then we can take $r = yq - x$ and we are done.

Let us consider the principal ideal $I = (y) = \{yz : z \in \mathbb{Z}[i]\}$. Note $yi = (c + di)i = ci - d$ is the point in the complex plane which is y rotated by 90° counter-clockwise. Furthermore, any other element in I is of the form $my + nyi$ where $n, m \in \mathbb{Z}$. Thus the elements of I form a square "lattice" in the complex plane.

Now x is inside one of the squares in this lattice. Note that there is a corner of the square whose distance from x is at most $|y|\sqrt{2}/2 < |y|$. (Here we use that a square of side lengths R has diagonal of length $\sqrt{2}R$.) In particular, the disk of radius $|y|$ centered at x contains one of the corners of the square. Therefore there is some point qy in I that lies in this disk around x. Since it is in this disk it satisfies $|qy - x| < |y|$, which is what we wanted.



The argument is illustrated in the figure above for a specific case, in which $y = 2 + i$ and $x = 1 + i$. The red dots are the ideal I, and the blue dot is x. In fact in this case you can see that the origin is in the disk around x, and we can take $q = 0$. □

We have now achieved our original goal, of showing that $\mathbb{Z}[i]$ has "unique factorization":

▶ **The ring of Gaussian integers $\mathbb{Z}[i]$ is a UFD.**

This follows from our result that a Euclidean domain is a UFD.

For example, a factorization of $5 \in \mathbb{Z}[i]$ is given as follows:

$$5 = (2 + i)(2 - i)$$

Furthermore, $(2+i)$ and $(2-i)$ are irreducibles, so this is the "unique factorization" of 5 in this ring. We also have the factorization

$$5 = (1-2i)(1+2i)$$

However the factors are related by units: $(2+i) = i(1-2i)$ and $(2-i) = -i(1+2i)$.

On the other hand, 3 remains "prime" in this ring, and does not factor into smaller irreducible parts. It can be shown that the only prime numbers $p \in \mathbb{Z}$ which are still "prime" (i.e. irreducible) in the ring $\mathbb{Z}[i]$ are the primes p such that $p \equiv 3 \pmod 4$.

Example: Consider the ring $\mathbb{Z}[\sqrt{-5}]$. This is *not* a UFD! In fact we have

$$6 = (2)(3) = (1+\sqrt{-5})(1-\sqrt{-5})$$

These are two different factorizations of 6 in this ring. Note that the units of this ring are $\{1, -1\}$, so these factorizations are truly distinct. The failure of $\mathbb{Z}[\sqrt{-5}]$ to be a UFD is related to (but not implied by) the fact that it is not a PID.

Primes as the sum of two squares

Last lecture we investigated the Gaussian integers $\mathbb{Z}[i]$. A main result was that this ring is a UFD, which means that it has a version of the Fundamental Theorem of Arithmetic, where every element in the ring can be factored in an essentially unique way.

In this lecture we continue this discussion. We begin by taking a step back and asking: why do we care about factorization in general rings? One reason is that it helps us understand concrete arithmetic problems in simpler rings like the integers. We illustrate this point with Fermat's "Sum of two squares" theorem, which is a statement about the integers, but whose proof will use factorization in the Gaussian integers.

We ask the following question: which prime numbers p can be written

$$p = a^2 + b^2$$

where a and b are integers? For example, we have $2 = 1^2 + 1^2$. It is not hard to see that 3 cannot be written as $a^2 + b^2$. On the other hand, we have

$$\begin{aligned} 5 &= 1^2 + 2^2 \\ 13 &= 2^2 + 3^2 \\ 17 &= 1^2 + 4^2 \end{aligned}$$

whereas 7 and 11 cannot be written as sums of two squares. After writing out enough examples you may recognize the pattern that the odd primes p that can be written as the sum of two squares satisfy $p \equiv 1 \pmod 4$.

To see why this is indeed true, suppose p is odd and $p = a^2 + b^2$. Then we consider

$$p = a^2 + b^2 \pmod 4$$

Now $a \pmod 4$ and $b \pmod 4$ are elements in $\mathbb{Z}_4 = \{0, 1, 2, 3\}$. Noting that in $\mathbb{Z}_4$ we have $0^2 = 0$, $1^2 = 1$, $2^2 = 0$, $3^2 = 1$, we must have $a^2 \pmod 4$ and $b^2 \pmod 4$ each equal to either 0 or 1 in $\mathbb{Z}_4$. Since the sum $p \mod 4$ is odd, we must have one of them equal to 0, say $a^2 \pmod 4$, and the other equal to 1, say $b^2 \pmod 4$. But then

$$p = a^2 + b^2 \equiv 0 + 1 \equiv 1 \pmod 4$$

Thus we have shown that if an odd prime p can be written as $p = a^2 + b^2$ then $p \equiv 1 \pmod 4$. This explains why $3, 7, 11, \ldots$ cannot be written as a sum of two squares. More surprising is that the converse is true, i.e. if $p \equiv 1 \pmod 4$ then p can be written as a sum of two squares.

▶ **(Fermat) A prime number $p \neq 2$ is the sum of two integer squares, i.e. $p = a^2 + b^2$ where $a, b \in \mathbb{Z}$, if and only if $p \equiv 1 \pmod 4$.**

Let us first make the connection to Gaussian integers. For $x = a + bi \in \mathbb{Z}[i]$ recall that we have the norm (or rather norm squared) function defined by

$$N(x) = |x|^2 = a^2 + b^2$$

This function is multiplicative in the sense that for all $x, y \in \mathbb{Z}[i]$ we have

$$N(xy) = N(x)N(y)$$

This is just a recasting of the property that $|xy| = |x||y|$ for the usual complex norm.

For any $x \in \mathbb{Z}[i]$ note that $N(x)$ is a non-negative integer. Furthermore, x is a unit if and only if $N(x) = 1$. Indeed, you have already classified the units in this ring: $\mathbb{Z}[i]^\times = \{1, -1, i, -i\}$.

Next, suppose $p = a^2 + b^2$. In the ring $\mathbb{Z}[i]$ this can be factored as follows:

$$p = a^2 + b^2 = (a + bi)(a - bi)$$

and each factor is not a unit. In other words, if $p = a^2 + b^2$ then p is no longer "prime", i.e. irreducible, in the ring $\mathbb{Z}[i]$. In fact, suppose $p \in \mathbb{Z}$ is not irreducible in $\mathbb{Z}[i]$. Then $p = xy$ for some non-units $x, y \in \mathbb{Z}[i]$, because of factorization in the ring $\mathbb{Z}[i]$. Then

$$p^2 = N(p) = N(xy) = N(x)N(y)$$

Since x, y are non-units, $N(x), N(y) > 1$ and hence $N(x) = N(y) = p$. Writing $x = a + bi$ we then have $p = N(x) = a^2 + b^2$. In other words, we have shown: if a prime $p \in \mathbb{Z}$ is *not irreducible* in the ring $\mathbb{Z}[i]$ then $p = a^2 + b^2$ for some integers $a, b \in \mathbb{Z}$.

Therefore, to complete the proof of Fermat's sum of two squares theorem we must show that primes $p \equiv 1 \pmod 4$ are not irreducible in $\mathbb{Z}[i]$.

To prove this last statement we will use the following lemma.

▶ **If a prime p satisfies $p \equiv 1 \pmod 4$ then there is some $m \in \mathbb{Z}$ such that**

$$m^2 + 1 \equiv 0 \pmod p$$

Proof. Recall that $\mathbb{Z}_p$ for p prime has the structure of a field. Note that $(p-1)! \pmod p$ is the product of all non-zero elements in $\mathbb{Z}_p$. For $x \neq \pm 1$ in $\mathbb{Z}_p^\times$ the multiplicative inverse of x is not equal to x. Thus apart from $\pm 1 \bmod p$, elements in $\mathbb{Z}_p^\times$ pair off as inverses. Thus

$$(p-1)! \equiv 1 \cdot 2 \cdots (p-1) \equiv 1 \cdot (-1) \equiv -1 \pmod p$$

This holds for any prime p. (This is sometimes called Wilson's Theorem.) Next suppose $p \equiv 1 \pmod 4$ and write $p = 1 + 4n$. Then we compute

$$\begin{aligned} -1 \equiv (p-1)! &\equiv (1 \cdot 2 \cdots (2n)) \left((p-1) \cdot (p-2) \cdots (p-2n)\right) \\ &\equiv ((2n)!) \left((-1)^{2n}(2n)!\right) \equiv ((2n)!)^2 \pmod p \end{aligned}$$

Now take $m = (2n)!$. Then $m^2 + 1 \equiv 0 \pmod p$, and we are done. □

Now we return to our main goal and show that if a prime number $p \in \mathbb{Z}$ satisfies $p \equiv 1 \pmod 4$ then p is not irreducible in $\mathbb{Z}[i]$.

From the lemma, we have some integer $m \in \mathbb{Z}$ such that p divides $m^2 + 1$. Then in the ring $\mathbb{Z}[i]$ we have that p divides the product of Gaussian integers

$$m^2 + 1 = (m + i)(m - i)$$

Thus we have an expression of the form

$$p \cdot x = (m + i)(m - i) = p_1 \cdots p_k \cdot q_1 \cdots q_l$$

where we have written out factorizations of $(m + i)$ and $(m - i)$ into irreducibles. Now, if p is irreducible, it is some p_i or some q_i up to multiplication by a unit, by unique factorization in $\mathbb{Z}[i]$. This implies p divides one of $m + i$ or $m - i$. However, $m/p \pm i/p$ are not Gaussian integers, so we have a contradiction. Thus we have proved that p is not irreducible. This completes the proof of Fermat's theorem on which primes are the sum of two squares.

With just a bit more work, the following more precise version of a "Fundamental Theorem of Arithmetic" for the Gaussian integers $\mathbb{Z}[i]$ can be proved.

▶ **The irreducible elements in $\mathbb{Z}[i]$ up to mulitplication by units are the following:**

- $1 + i$
- $a+bi$ **where** $a, b \in \mathbb{Z}$ **with** $a > |b| > 0$ **and** $a^2+b^2 = p$ **a prime, with** $p \equiv 1 \pmod 4$
- $p \in \mathbb{Z}_{>0}$ **prime with** $p \equiv 3 \pmod 4$

Thus every element in $\mathbb{Z}[i]$ can be written uniquely as a product of the above elements and a unit, which is one of $1, -1, i, -i$.

Constructions of fields

For the past few weeks we have studied general properties of rings, and described how ring theory interacts with other areas of mathematics such as geometry and number theory. For the rest of the course, we will focus our attention on field theory.

Today we will discuss a number of constructions for fields that will be useful later.

Field of fractions

Let R be any integral domain. It may not be the case that R is a field. However, we can formally "invert" all non-zero elements of R to obtain a field. The result of this construction is $\text{Frac}(R)$, called the *field of fractions of* R, also called the *quotient field*.

To construct this field, we begin with the set of pairs $\{(a, b) : \ a \in R, b \in R, b \neq 0\}$. We define an equivalence relation $\sim$ on this set as follows:

$$(a, b) \sim (c, d) \qquad \iff \qquad ac - bd = 0$$

Let $\text{Frac}(R)$ be the equivalence classes of this relation. It is customary to write the equivalence class of (a, b) as a/b. Then define addition and multiplication as:

$$\frac{a}{b} + \frac{c}{d} = \frac{ad + bc}{bd} \qquad \frac{a}{b} \cdot \frac{c}{d} = \frac{ac}{bd}$$

You can then check that with these operations the set $\text{Frac}(R)$ satisfies the axioms of a commutative ring. Furthermore, if $a/b \in \text{Frac}(R)$ is non-zero, then it has multiplicative inverse b/a. Thus $\text{Frac}(R)$ is a field. The following is an exercise:

▶ **The map $\phi : R \to \text{Frac}(R)$ given by $\phi(a) = a/1$ is a 1-1 ring homomorphism.**

For example, the field of fractions of the integers $\mathbb{Z}$ is naturally isomorphic to the field of rational numbers $\mathbb{Q}$, and we write $\text{Frac}(\mathbb{Z}) = \mathbb{Q}$.

If R is a field, then $\text{Frac}(R)$ is naturally isomorphic to R.

For another example, consider the ring of polynomials $R[x]$ where R is an integral domain such as $\mathbb{Z}$ or $\mathbb{Q}$. Then $R[x]$ is an integral domain, and we have

$$\text{Frac}(R[x]) = \left\{ \frac{f(x)}{g(x)} : \ f(x), g(x) \in R[x], \ g(x) \neq 0 \right\}$$

It is instructive to verify that this field is isomorphic to $\text{Frac}(\text{Frac}(R)[x])$. In other words, we can first take the field of fractions of R, then the field of fractions of $\text{Frac}(R)[x]$, and we get the same result. For example, $\text{Frac}(\mathbb{Z}[x])$ is isomorphic to $\text{Frac}(\mathbb{Q}[x])$.

The field of fractions of R is an abstract field constructed in a way such that there is a natural "inclusion" homomorphism $R \to \text{Frac}(R)$. It gives us a way of viewing any integral domain as sitting inside some larger field. In practice, we may already have R sitting inside a field, and the field of fractions is then isomorphic to something very concrete.

▶ **Suppose a ring R is contained in a field E, and F is the smallest subfield of E containing R. Then F is isomorphic to Frac(R).**

Proof. By assumption, $R \subset F$. We define $\phi : \text{Frac}(R) \to F$ as follows: $\phi(a/b) = ab^{-1}$. This makes sense because $a \in R$ and $b \in R$; in particular, $b \in F$, so b has a multiplicative inverse. Then ϕ is a homomorphism of fields. This is a 1-1 homomorphism, as is every homomorphism between fields. Consider the image of ϕ. This is a subfield of F containing R, and so by assumption it must be equal to F. Thus ϕ is an isomorphism. □

Let $R = \mathbb{Z}[i]$, the Gaussian integers. Consider the following subfield of $\mathbb{C}$:

$$\mathbb{Q}(i) = \{a + bi : \ a, b \in \mathbb{Q}\}$$

It is not difficult to see that this is the smallest subfield of $\mathbb{C}$ which contains $\mathbb{Z}[i]$. Thus Frac$(\mathbb{Z}[i])$ is isomorphic to $\mathbb{Q}(i)$.

Extension fields

Let F be a field, and E is a field containing F, i.e. $F \subset E$. In this situation E is called an *extension field* of F. The extension is *proper* if $E \neq F$.

If $\alpha \in E$ we write $F(\alpha) \subset E$ for the smallest field inside E containing both F and α.

For example, let $F = \mathbb{Q}$, $E = \mathbb{C}$, $\alpha = i = \sqrt{-1}$. Then $\mathbb{Q}(i)$ is the smallest field contained in $\mathbb{C}$ which contains both $\mathbb{Q}$ and i.

For a more interesting example let us consider the field

$$E = \mathbb{Q}(\sqrt{2} + \sqrt{3})$$

This is the smallest field containing both $\mathbb{Q}$ and $\alpha = \sqrt{2} + \sqrt{3}$. Note that

$$(\sqrt{2} + \sqrt{3}) \cdot (\sqrt{2} - \sqrt{3}) = 2 - 3 = -1$$

and so the multiplicative inverse of $\sqrt{2} + \sqrt{3}$ is given by $-\sqrt{2} + \sqrt{3}$. Then

$$\frac{1}{2}\left(\alpha + \alpha^{-1}\right) = \frac{1}{2}\Big((\sqrt{2} + \sqrt{3}) + (-\sqrt{2} + \sqrt{3})\Big) = \sqrt{3}$$

so that $\sqrt{3} \in E$. Similarly, $\sqrt{2} \in E$. In particular, E is an extension field of the fields $\mathbb{Q}(\sqrt{2})$ and $\mathbb{Q}(\sqrt{3})$. Note $\sqrt{2} + \sqrt{3}$ is not in either of these fields, so the extension is proper.

Quotient fields

A very useful construction that can give us fields is the quotient ring R/I of some ideal I in a commutative ring R. Recall that R/I is a field if and only if I is a *maximal* ideal.

A common scenario is when the ring R is a polynomial ring of the form $F[x]$ where F itself is a field. Recall we proved that $F[x]$ is a PID in this case. We need the following:

▶ **An ideal (a) in a PID is maximal if and only if a is irreducible.**

Proof. Suppose (a) is maximal, and a is not irreducible. Then $a = bc$ where b, c are not units. Then $(a) = (bc) = (b)(c)$ where (b), (c) are proper ideals. In particular, $(a) \subset (b)$. By maximality, $(a) = (b)$. Then $a = bu$ for a unit u. We obtain $a = bu = bc$ implying $b(u - c) = 0$. Since the ring is an integral domain and $b \neq 0$, $u = c$, Contradicting that c is not a unit. Thus a is irreducible. Conversely, suppose a is irreducible, and (a) is not maximal. Then (a) is properly contained in (b) for some proper ideal (b). Then $a = bc$ for some c. As a is irreducible, one of b or c is a unit. If b is a unit, (b) is not proper. If c is a unit, then $(a) = (b)$, a contradiction. Thus (a) is a maximal ideal. □

Returning to the case of $F[x]$, where F is a field, note that the group of units of $F[x]$ is exactly the non-zero elements of F. Thus $f(x)$ is irreducible if $f(x)$ is non-zero and not a constant polynomial, and if $f(x) = g(x)h(x)$ implies that one of $g(x)$ or $h(x)$ is a constant polynomial. Recalling that $F[x]/(f(x))$ is a field when $(f(x))$ is maximal, we obtain:

▶ **For F a field and $f(x) \in F[x]$ an irreducible polynomial, $F[x]/(f(x))$ is a field.**

Let us now use this to construct some fields.

Take $F = \mathbb{R}$ and let $f(x) \in \mathbb{R}[x]$ be the irreducible polynomial $f(x) = x^2 + 1$. Then

$$\mathbb{R}[x]/(x^2 + 1)$$

is a field. We consider some elements in this field, where $I = (x^2 + 1)$:

$$\begin{aligned} x^2 + I &= -1 + I \\ x^3 + I &= -x + I \\ x^4 + I &= \quad 1 + I \\ &\vdots \end{aligned}$$

Using these computations we see that any element $g(x) + I$ can be written as $a + bx + I$ where $a, b \in \mathbb{R}$. Noting that x behaves just like $i = \sqrt{-1} \in \mathbb{C}$, we define

$$\phi : \mathbb{C} \longrightarrow \mathbb{R}[x]/(x^2 + 1)$$

by $\phi(a + bi) = a + bx + I$, and you may verify this is an isomorphism.

For another example, let us take $F = \mathbb{Z}_2$ and let $f(x) \in \mathbb{Z}_2[x]$ be $f(x) = x^2 + x + 1$. Suppose this is reducible. Then $f(x) = (a + bx)(c + dx) = ac + (bc + ad)x + bdx^2$ for $a, b, c, d \in \mathbb{Z}_2$. But it is easily seen that no such a, b, c, d can work. Thus $f(x)$ is irreducible in $\mathbb{Z}_2[x]$.

Now we have a field given by the quotient

$$\mathbb{Z}_2[x]/(x^2 + x + 1)$$

Let us write down some elements of this field, where $I = (x^2 + x + 1)$:

$$x^2 + I = x + 1 + I$$
$$x^3 + I = 1 + I$$

In fact we see that every element in this field is equivalent to one of the cosets corresponding to $0, 1, x, 1 + x$. Thus we have a field with 4 elements.

Algebraic field extensions

Recall that if F is a field and E is an extension of F, and $\alpha \in E$, then we write $F(\alpha)$ for the smallest subfield in E which contains F and α.

Of course in this case $F(\alpha)$ is again a field extension of F, and E is an extension of $F(\alpha)$. In short, $F \subset F(\alpha) \subset E$. We call $F(\alpha)$ a *simple extension* of F. We can iterate this notation: for elements $\alpha_1, \ldots, \alpha_n \in E$ we write $F(\alpha_1, \alpha_2, \ldots, \alpha_n)$ for the smallest field in E containing F and the elements $\alpha_1, \ldots, \alpha_n$.

An element $\alpha \in E$ in an extension field E of F is called *algebraic* over F if there is some polynomial $f(x) \in F[x]$ such that $f(\alpha) = 0$. Otherwise α is called *transcendental*.

Examples

1. The element $i = \sqrt{-1}$ is algebraic over $\mathbb{Q}$, because it satisfies $f(i) = 0$ where $f(x) = x^2+1$.

2. The elements $\sqrt{2}$, $\sqrt{3}$, $\sqrt{2}+\sqrt{3}$ are algebraic over $\mathbb{Q}$. The first two are roots of $x^2 - 2$, $x^2 - 3$, respectively. What about the third? Let $\alpha = \sqrt{2}+\sqrt{3}$. We compute

$$\alpha^2 = (\sqrt{2}+\sqrt{3})^2 = 5 + 2\sqrt{6}$$

$$(\alpha^2 - 5)^2 = (2\sqrt{6})^2 = 24$$

Thus $\alpha = \sqrt{2}+\sqrt{3}$ satisfies $f(\alpha) = 0$ where $f(x) = (x^2-5)^2 - 24 = x^4 - 10x^2 + 1$.

3. The numbers π and e are transcendental over $\mathbb{Q}$. This is because they are not the roots of any polynomial with coefficients in $\mathbb{Q}$.

4. Any n^{th} root of a rational number p/q is algebraic over $\mathbb{Q}$, since it satisfies $x^n - p/q = 0$. More generally, any number obtained from the rationals via the operations of multiplication, addition, division, and taking various types of n^{th} roots is algebraic over $\mathbb{Q}$. For example,

$$\sqrt[5]{\frac{\sqrt{2}-1}{\sqrt[3]{4+\sqrt{5}}}}$$

is algebraic over $\mathbb{Q}$. This may not be obvious, but it will follow from a result we will prove.

5. All of the above examples are about algebraic and transcendental numbers over $\mathbb{Q}$. The numbers π and e are algebraic over $\mathbb{R}$, because they are in $\mathbb{R}$! Precisely, they satisfy the polynomial equations $x - \pi = 0$ and $x - e = 0$, and these are of course polynomials in $\mathbb{R}[x]$. As another example, $\sqrt{\pi}$ is algebraic over $\mathbb{Q}(\pi)$.

An extension E of a field F is *algebraic* if every $\alpha \in E$ is algebraic over F, and is called *transcendental* otherwise.

▶ **Suppose $\alpha \in E$ is algebraic over F. Then there is a unique monic irreducible polynomial $p(x) \in F[x]$ such that $p(\alpha) = 0$.**

Proof. Consider the evaluation homomorphism $\phi_\alpha : F[x] \to F(\alpha)$ given by $\phi_\alpha(f(x)) = f(\alpha)$. Since F is a field, $F[x]$ is a PID, and thus the kernel of ϕ_α is an ideal $(f(x))$ where $f(x) \in F[x]$. Note that for a constant polynomial $g(x) = a \in F$ we have $\phi_\alpha(g(x)) = a$, and also for $g(x) = x$ we have $\phi_\alpha(g(x)) = \alpha$. Thus ϕ_α is onto, as the image contains F and α, and $F(\alpha)$ is the smallest field containing F and α. The 1st Isomorphism Theorem then gives an isomorphism between $F[x]/(f(x))$ and $F(\alpha)$, which is a field. Thus $(f(x))$ is a maximal ideal, and $f(x)$ must be an irreducible polynomial. If $f(x)$ is not monic, say $f(x) = a_n x^n + \cdots + a_0$ where $a_n \in F$ is non-zero, then $p(x) = f(x)/a_n$ is a monic irreducible polynomial with $(f(x)) = (p(x))$. Note $p(\alpha) = \phi_\alpha(p(x)) = 0$. Uniqueness is left as an exercise. □

In the above situation, we call $p(x)$ the *minimal polynomial* of α over F. For example, the minimal polynomial of $\sqrt{-1}$ over $\mathbb{Q}$ is $p(x) = x^2 + 1$.

▶ **Let E be a field extension of F and $\alpha \in E$.**

(i) If α is algebraic with minimal polynomial $p(x)$, then $F(\alpha) \cong F[x]/(p(x))$.

(ii) If α is transcendental, then $F(\alpha) \cong$ Frac$(F[x])$.

Proof. (i) Suppose α is algebraic. Let $\phi_\alpha : F[x] \to F(\alpha)$ be the evaluation homomorphism. From the proof of the previous result, ϕ_α is onto and $\ker(\phi_\alpha) = (p(x))$ where $p(x)$ is the minimal polynomial of α. The 1st Isomorphism Theorem then proves (i).

(ii) Now suppose α is transcendental. Then ϕ_α has trivial kernel. Indeed, if $f(x) \in \ker(\phi_\alpha)$ then $f(\alpha) = 0$; and if $f(x)$ is a non-zero polynomial in $F[x]$ this would contradict the assumption that α is transcendental. Next, we can extend ϕ_α to a homomorphism

$$\psi : \text{Frac}(F[x]) \longrightarrow F(\alpha)$$

by setting $\psi(f(x)/g(x)) = f(\alpha)/g(\alpha)$. This makes sense because $f(x)/g(x) \in \text{Frac}(F[x])$ has $g(x)$ a non-zero polynomial, and $g(\alpha) \neq 0$ because α is transcendental. The map ψ is 1-1 because the domain is a field, and is onto because ϕ_α is onto. Thus ψ is an isomorphism. □

▶ **Let F be a field and $f(x) \in F[x]$ a non-constant polynomial. Then there is an extension E of F that contains some $\alpha \in E$ such that $f(\alpha) = 0$.**

Proof. If $f(x)$ is not irreducible, let $p(x)$ be an irreducible polynomial which divides $f(x)$. Then set $E = F[x]/(p(x))$. This is a field, and F naturally is included into it. The equivalence class of x serves as the element α. □

One of the most important tools in field theory is linear algebra. As you have already taken a linear algebra course, we only briefly review some of the basics.

Review of some linear algebra

A *vector space* over a field F is an abelian group V equipped with an operation

$$F \times V \longrightarrow V$$

written $(a, v) \longmapsto av$, called scalar multiplication, which satisfies for all $a, b \in F$ and $v, w \in V$:

$$\begin{aligned} a(bv) &= (ab)v \\ (a+b)v &= av + bv \\ a(v+w) &= av + aw \\ 1v &= v \end{aligned}$$

Elements of V are called vectors, and elements of F in this context are often called scalars. A subset $S \subset V$ is *linearly independent* if when $v_1, \ldots, v_m \in S$ satisfy

$$a_1 v_1 + \cdots + a_m v_m = 0$$

then we must have $a_1 = \cdots = a_m = 0$. Suppose S is a maximal linearly independent subset of V. This means that if $S \subset T$ and T is a linearly independent subset of V, then $S = T$. In this case S is called a *basis* for V, and the *dimension* of V is given by

$$\dim_F V = \#S = \text{ size of a maximal linearly independent subset}$$

If S is a finite set, then $\dim_F V$ is finite, and V is called *finite-dimensional*. Otherwise, V is *infinite-dimensional*. A linearly independent subset $S \subset V$ is a basis if and only if it spans V, i.e. any element of V can be written as a linear combination of elements in S.

Degrees of field extensions

Now let us return to our previous setup, where F is some field, and E is an extension field of F. Then E is a vector space over F: indeed, the scalar multiplication map $E \times F \longrightarrow F$ is just the multiplication of field elements, all of which may be viewed as inside E. We define

$$[E : F] = \dim_F E$$

and call this the *degree* of the field extension E of F. In the next lecture we will study the degrees of field extensions and their relation to algebraicity.

Degrees of field extensions

Last lecture we introduced the notion of algebraic and transcendental elements over a field, and we also introduced the *degree* of a field extension. Recall that for a field extension E of F, we may view E as a vector space over F, and the degree of the extension E is given by

$$\dim_F E = [E : F]$$

Today we will study the relationship between algebraic extensions and degrees of extensions. We first begin with a few examples.

Examples

1. Consider the extension $\mathbb{Q}(\sqrt{2})$ of the field $\mathbb{Q}$. We know that $\mathbb{Q}(\sqrt{2})$ consists of numbers $a+b\sqrt{2}$ where $a, b \in \mathbb{Q}$. Let $v_1 = 1$ and $v_2 = \sqrt{2}$. Then $S = \{v_1, v_2\}$ is a linearly independent subset of $\mathbb{Q}(\sqrt{2})$ as a vector space over $\mathbb{Q}$, and S spans $\mathbb{Q}(\sqrt{2})$. Thus S is a basis for $\mathbb{Q}(\sqrt{2})$ viewed as a vector space over $\mathbb{Q}$. From this we conclude

$$[\mathbb{Q}(\sqrt{2}) : \mathbb{Q}] = \dim_{\mathbb{Q}} \mathbb{Q}(\sqrt{2}) = 2$$

Thus the degree of the extension $\mathbb{Q}(\sqrt{2})$ over $\mathbb{Q}$ is equal to 2.

2. Consider $\mathbb{C}$ as an extension of $\mathbb{R}$. We can write every complex number uniquely as $a + bi$ where $a, b \in \mathbb{R}$. Then $S = \{1, i\}$ is a basis for $\mathbb{C}$ viewed as a vector space over $\mathbb{R}$, and

$$[\mathbb{C} : \mathbb{R}] = \dim_{\mathbb{R}} \mathbb{C} = 2$$

Thus the degree of the extension $\mathbb{C}$ over $\mathbb{R}$ is equal to 2.

3. Consider $\mathbb{Q}(\sqrt{2}+\sqrt{3})$ as an extension of $\mathbb{Q}(\sqrt{2})$. We saw last lecture that $\sqrt{2} \in \mathbb{Q}(\sqrt{2}+\sqrt{3})$ so this is in fact an extension. In fact, we can argue that

$$\mathbb{Q}(\sqrt{2}+\sqrt{3}) = \mathbb{Q}(\sqrt{2}, \sqrt{3})$$

Indeed, clearly $\mathbb{Q}(\sqrt{2}+\sqrt{3}) \subset \mathbb{Q}(\sqrt{2}, \sqrt{3})$, and as we showed that $\sqrt{2}, \sqrt{3} \in \mathbb{Q}(\sqrt{2}+\sqrt{3})$ the reverse inclusion also holds. Every element in this extension can be written uniquely as

$$x + \sqrt{3}y = (a + b\sqrt{2}) + \sqrt{3}(c + d\sqrt{2}) = a + b\sqrt{2} + c\sqrt{3} + d\sqrt{6}$$

where $a, b, c, d \in \mathbb{Q}$. Thus $S = \{1, \sqrt{3}\}$ is a basis of $\mathbb{Q}(\sqrt{2}+\sqrt{3})$ over the field $\mathbb{Q}(\sqrt{2})$, and

$$[\mathbb{Q}(\sqrt{2}+\sqrt{3}) : \mathbb{Q}(\sqrt{2})] = \dim_{\mathbb{Q}(\sqrt{2})} \mathbb{Q}(\sqrt{2}+\sqrt{3}) = 2$$

At the same time we see that as a vector space over $\mathbb{Q}$, the field $\mathbb{Q}(\sqrt{2}, \sqrt{3})$ has basis $\{1, \sqrt{2}, \sqrt{3}, \sqrt{6}\}$ and so is of degree 4 over $\mathbb{Q}$:

$$[\mathbb{Q}(\sqrt{2}+\sqrt{3}) : \mathbb{Q}] = \dim_{\mathbb{Q}} \mathbb{Q}(\sqrt{2}+\sqrt{3}) = 4$$

We say that an extension E of a field F is *finite* if the degree $[E : F]$ is a finite number, i.e. if E is a finite dimensional vector space over the field F.

▶ **If E is a finite extension of F, then E is an algebraic extension over F.**

Proof. Suppose $[E : F] = n$. Let $\alpha \in E$. Consider the elements $1, \alpha, \ldots, \alpha^n$. As there are $n+1$ elements here, and $\dim_F E = n$, they must be linear dependent, i.e. there is a relation

$$a_n\alpha^n + a_{n-1}\alpha^{n-1} + \cdots + a_1\alpha + a_0 1 = 0$$

where not all of the $a_0, \ldots, a_n \in F$ are zero. Then α is a root of the polynomial $f(x) \in F[x]$ given by $f(x) = \sum_{i=0}^{n} a_i x^i$, and so α is algebraic over F. As α was an arbitrary element of E, we conclude that E is an algebraic extension of F. □

▶ **Suppose $F \subset E$ and $E \subset K$ are finite extensions. Then**

$$[K : F] = [K : E][E : F]$$

In particular, K is a finite extension of F.

Proof. Suppose $[E : F] = n$ and $[K : E] = m$. Let $\{v_1, \ldots, v_n\} \subset E$ be a basis for E as a vector space over F, and $\{w_1, \ldots, w_m\} \subset K$ a basis for K as a vector space over E. Then we claim that $S = \{v_i w_j\} \subset K$ where $1 \leqslant i \leqslant n$ and $1 \leqslant j \leqslant m$ is a basis for K as a vector space over F. To establish this we must show that S is linearly independent and also spans K.

We first show S is linearly independent. Suppose we have a relation

$$\sum_{\substack{1 \leqslant i \leqslant n \\ 1 \leqslant j \leqslant m}} a_{ij} \cdot v_i w_j = 0$$

where $a_{ij} \in F$. Then we can write this expression as

$$\sum_{1 \leqslant j \leqslant m} c_j \cdot w_j = 0 \qquad \text{where} \quad c_j = \sum_{1 \leqslant i \leqslant n} a_{ij} v_i \in E$$

Since the w_j are linearly independent in K over E, we must have $c_j = 0$ for $1 \leqslant j \leqslant m$. Then

$$c_j = \sum_{1 \leqslant i \leqslant n} a_{ij} v_i = 0$$

and as the v_i are linearly independent in E over F, we must have, for each j, that $a_{ij} = 0$ for $1 \leqslant i \leqslant n$. Thus all $a_{ij} = 0$. This shows that S is linearly independent over F.

We show S spans K over F. Let $k \in K$. As the w_j are a basis for K over E, we can write

$$k = \sum_{j=1}^{m} c_j w_j$$

for some $c_j \in E$. As the v_i are a basis of E over F, we can write $c_j = \sum_{i=1}^{n} a_{ij} v_i$ for some $a_{ij} \in F$. Then $k = \sum a_{ij} v_i w_j$. This show S spans K over F Hence S is a basis for K over E.

Finally, the basis S contains nm elements and we compute

$$\dim_F K = [K : F] = nm = [K : E][E : F] \qquad \square$$

You might have already noticed this property in the last example we studied above: we have

$$4 = [\mathbb{Q}(\sqrt{2} + \sqrt{3}) : \mathbb{Q}] = [\mathbb{Q}(\sqrt{2} + \sqrt{3}) : \mathbb{Q}(\sqrt{2})][\mathbb{Q}(\sqrt{2}) : \mathbb{Q}] = 2 \cdot 2$$

▶ **Let α be algebraic over F with minimal polynomial $p(x) \in F[x]$. Then**

$$[F(\alpha) : F] = \deg(p(x))$$

Proof. Let $p(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_0$ where $a_i \in F$. Then since α is a root of $p(x)$ we have $\alpha^n + a_{n-1}\alpha^{n-1} + \cdots + a_0 = 0$. We claim $S = \{1, \alpha, \ldots, \alpha^{n-1}\} \subset F(\alpha)$ is a basis of $F(\alpha)$ as a vector space over F. First, if a non-trivial linear combination of these elements with coefficients in F is zero, it would show α is the root of a polynomial of degree $< n$, contradicting the minimality of $p(x)$. Thus S is linearly independent. To see that S spans $F(\alpha)$, make use of the isomorphism $F(\alpha) \cong F[x]/(p(x))$. $\square$

▶ **The following are equivalent statements for a field extension $F \subset E$.**

(i) **E is a finite extension of F.**

(ii) **There are algebraic elements $\alpha_1, \ldots, \alpha_n$ such that $E = F(\alpha_1, \ldots, \alpha_n)$.**

Proof. In (ii), note that we have a sequence of algebraic extensions

$$F \subset F(\alpha_1) \subset F(\alpha_1, \alpha_2) \subset \cdots \subset F(\alpha_1, \ldots, \alpha_n) = E$$

As each of these algebraic extensions is of finite degree by the previous result, we have that (ii) implies (i).

To see that (i) implies (ii), choose $\alpha_1 \in E$ to be an element not in F. Then

$$[E : F] = [E : F(\alpha_1)][F(\alpha_1) : F]$$

Next, choose $\alpha_2 \in E$ not in $F(\alpha_1)$. Then we have

$$[E : F] = [E : F(\alpha_1, \alpha_2)][F(\alpha_1, \alpha_2) : F(\alpha_1)][F(\alpha_1) : F]$$

Continue in this fashion, and choose $\alpha_3 \in E$ not in $F(\alpha_1, \alpha_2)$, and so on. Since $[E : F]$ is finite, this process must eventually terminate at a step in which $E = F(\alpha_1, \ldots, \alpha_n)$. $\square$

▶ **Let $F \subset E$. The set of elements in E algebraic over F form a field.**

Proof. From the previous result, $F(\alpha, \beta)$ is a finite extension of F, and hence is an algebraic extension of F. As this extension contains the elements $\alpha \pm \beta, \alpha\beta, \alpha/\beta$ $(\beta \neq 0)$, these are all algebraic elements over F. This shows that the subset of E of algebraic elements over F is a subfield of E, and is in particular a field. □

We now return to our example from last lecture of the number

$$\sqrt[5]{\frac{\sqrt{2}-1}{\sqrt[3]{4+\sqrt{5}}}}$$

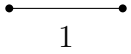
It is at this point easy to deduce that this is an algebraic element over $\mathbb{Q}$. First, we note that an n^{th} root of a number algebraic over $\mathbb{Q}$ is also algebraic over $\mathbb{Q}$. To prove this, if α is algebraic, then it satisfies $p(\alpha) = 0$ for some $p(x) \in \mathbb{Q}[x]$; then $\sqrt[n]{\alpha}$ is a root of $p(x^n) \in \mathbb{Q}[x]$, so it is also algebraic over $\mathbb{Q}$.

Then all that is left to observe is that the number displayed above is obtained from rational numbers and the field operations (addition, subtraction, multiplication, division) and taking n^{th} roots; all of these operations preserve the class of algebraic elements.

Straightedge and compass constructions

In classical geometry, the tools of a straightedge (a ruler without markings) and a compass (which makes circles) are used to construct new shapes out of old ones. Many old problems in geometry ask whether certain operations or quantities can be constructed in this fashion. In this lecture we solve some of these problems by relating them to field theory.

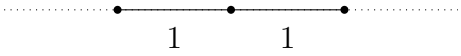
A (real) number α is *constructible* if we can construct a line segment of length $|\alpha|$ in the plane $\mathbb{R}^2$ starting from a line segment of length 1 and using a straightedge and compass.



Let $F \subset \mathbb{R}$ be the set of constructible numbers. At the beginning, we know only that $\{-1, 0, 1\} \subset F$. Indeed, we are allowed to start with a segment of unit length; and we can draw a point in $\mathbb{R}^2$ which is a line segment of length 0.

What is allowed? Using the straightedge, we can draw straight lines as we like through points we have already drawn, but the only arcs we can draw are the ones whose lengths are already in F. Similarly, with the compass we can draw circles of radius which is either specified by two points we have already drawn, or is the length of something in F.

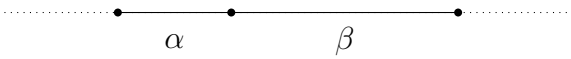
So what other numbers are constructible, i.e. are contained in F? Using our straightedge we can draw unit length arcs in sequence in a straightline:



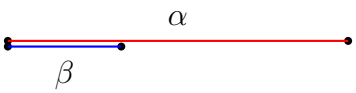
This shows that $\mathbb{Z} \subset F$, i.e. the integers are constructible numbers.

▶ **If $\alpha, \beta \in F$ then $\alpha + \beta$ and $\alpha - \beta$ are also in F.**

The idea is the same as above: given arcs of length α and β we can draw them side by side, parallel and adjacent, using our straightedge. Then we have an arc of length $\alpha + \beta$.



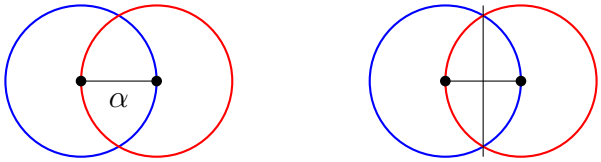
Similarly, if we draw the arcs as starting from the same point, overlapping, then we can extract an arc whose length is the difference $\alpha - \beta$, assuming for simplicity that $\alpha > \beta$.



In the picture, the arcs should really be superimposed. The portion of the arc of length α, in red, which does not overlap the arc of length β, in blue, is a segment of length $\alpha - \beta$.

What about multiplication and division of constructible numbers? This is a bit more work.

Suppose $\alpha \in F$, $\alpha > 0$. Then we can draw an arc of length α. We can then draw two circles each of radius α, centered at the two points of the arc, as in the right picture below.



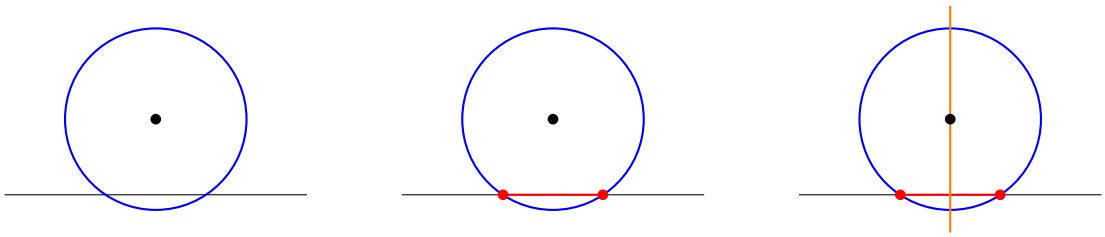
Then using the straightedge we can draw a line passing between the two points of intersection of the circles. We have then bisected the arc! Thus we now know how to draw an arc of length $\alpha/2$. Thus if α is constructible, then $\alpha/2$ is constructible.

We should also record a consequence of our construction: given an arc, we can draw a line perpendicular to that arc passing through the midpoint of the arc.

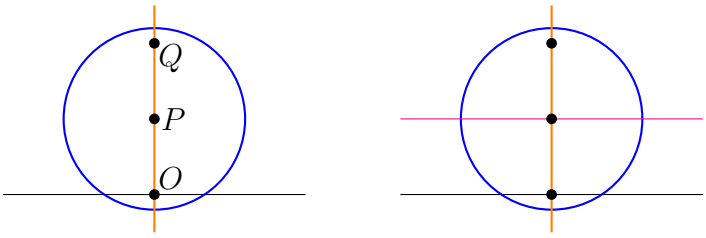
Now consider a point lying above a straight line. We claim that using our tools we can draw a line parallel to the given line, and passing through the given point P. We assume the the distance between the point and the line is constructible.



To do this, first draw a circle centered at the point and of radius large enough so that it intersects the line twice. An arc is formed between the intersection points (in red below).



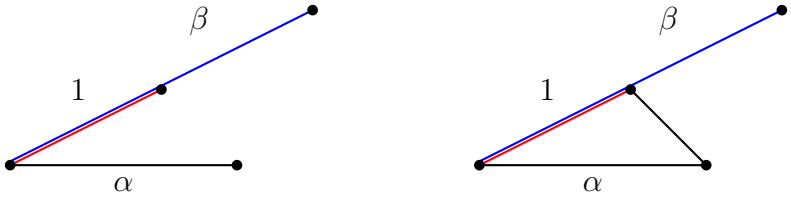
From our previous construction we can draw a line perpendicular to this arc which bisects it. This perpendicular also passes through P. Next, consider the arc from the intersection of the lines, call this point O, to the point P. As the length of OP is constructible, we can draw an arc of the same length starting from P upwards, ending at a point Q.



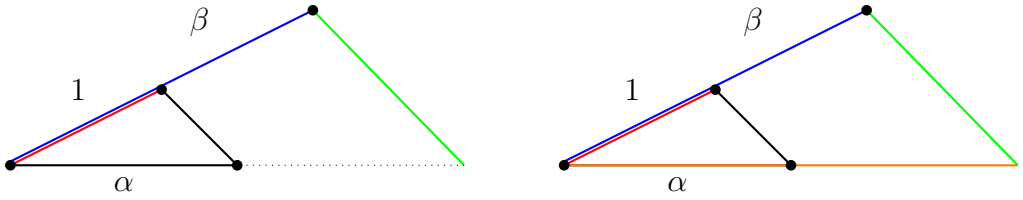
Finally, we know from above that we can draw a line perpendicular to the segment OQ and which passes through its midpoint. This is the desired parallel line.

▶ **If $\alpha, \beta \in F$ then $\alpha\beta$ and α/β ($\beta \neq 0$) are also in F.**

Proof. Assume $\beta > 1$. Draw an arc of length α, an arc of length 1, and an arc of length β all emanating from the same starting point. The 1 and β arcs should be parallel, and the angle between these and the α arc is some arbitrary angle between 0 and 90°.



The situation is depicted above on the left. Then use the straightedge to draw a segment from the endpoint of the unit length arc to the endpoint of the α arc. We then have the right-hand picture above. Next, we know from previous constructions that we can draw a parallel arc (in green) to the arc just drawn, and passing through the endpoint of β.



Now we have an arc starting from the same point as the others, and ending at the intersection of the green arc with the line containing the α arc. This new arc is in orange, above on the right. Call the length x. The two triangles in the picture are similar, so we have $\alpha/1 = x/\beta$. Thus $x = \alpha\beta$. Thus we have constructed an arc of length $\alpha\beta$.

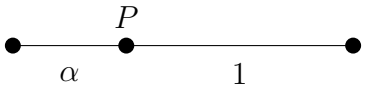
The case $\beta < 1$ and also the case of division are similar, and left as exercises. □

▶ **The set $F \subset \mathbb{R}$ of constructible numbers is a field.**

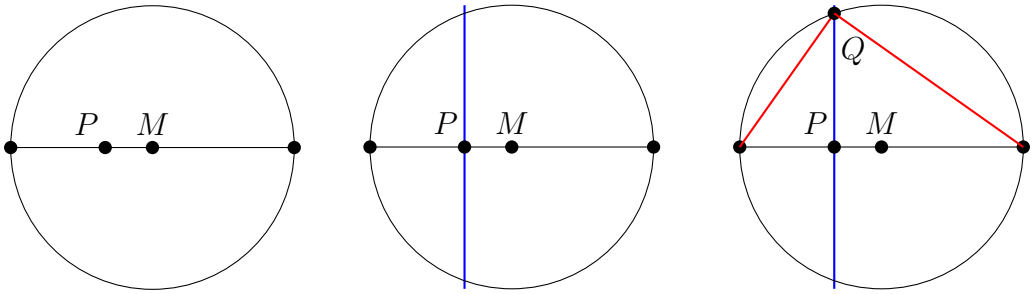
This result just follows from the above work. Thus far it is clear that $\mathbb{Q} \subset F$. However the field F is larger than just the rationals, as the following result indicates.

▶ **If $\alpha \in F$ then $\sqrt{\alpha} \in F$.**

Proof. Start with two arcs of length 1 and α side by side, forming an arc of length $1 + \alpha$.



Bisect this line and draw a circle centered at the midpoint M and with radius $(1 + \alpha)/2$ so that the arc is a diameter of the circle. See the left picture below.



Draw a perpendicular line to the original line which passes through the point P, as in the picture above in the center. (We can draw this line because it is passing through the midpoint of constructible arc of length 2α.) Draw arcs from the endpoints of the original arc to the point of intersection Q, as in the picture above on the right. Then the arc PQ has some length x. Using the similarity of triangles appearing, we have $\alpha/x = x/1$ or $x^2 = \alpha$. Thus we obtain that $x = \sqrt{\alpha}$ is constructible. □

In fact we have essentially found all of the ways of producing constructible numbers.

▶ **A number $\alpha \in \mathbb{R}$ is constructible if and only if there is a sequence of fields**

$$\mathbb{Q} = F_0 \subset F_1 \subset F_2 \subset \cdots \subset F_k$$

such that $F_{i+1} = F_i(\sqrt{\alpha_i})$ where $\alpha_i \in F_i$, and $\alpha \in F_k$.

We will not go through the proof in detail, although it is not too difficult. Roughly, all of the "tricks" one can do to form new constructible numbers are contained in the above results.

▶ **If α is constructible, then $[\mathbb{Q}(\alpha) : \mathbb{Q}] = 2^k$ for some $k \geqslant 0$.**

Proof. From the previous result, we have $\mathbb{Q}(\alpha) = F_k$ for some F_k and a sequence of fields $\mathbb{Q} = F_0 \subset F_1 \subset \cdots \subset F_k$ which are quadratic extensions obtained by adjoing square roots, and in particular have degree 1 or 2. Then we compute $[\mathbb{Q}(\alpha) : \mathbb{Q}]$ to be

$$[\mathbb{Q}(\alpha) : \mathbb{Q}] = [F_k : F_{k-1}] \cdots [F_2 : F_1][F_1 : F_0]$$

which is some power of 2. □

We say a subset of $\mathbb{R}^2$ is *constructible* if it can be draw using straightedge and compass operations as above, using only constructible measurements.

We consider the problem of constructing a regular n-gon, i.e. a symmetric polygon with n sides, using only a straightedge and compass. It is straightforward to argue:

▶ **If the regular n-gon is constructible, then $\cos(2\pi/n)$ is constructible.**

Using this we now prove:

▶ **The regular 18-gon is not constructible.**

Proof. We only need argue that $\cos(2\pi/18) = \cos(360°/18) = \cos(20°)$ is not a constructible number. Using $(\cos(\theta) + \sin(\theta)i)^3 = \cos(3\theta) + \sin(3\theta)i$ we have the identity

$$\cos(3\theta) = 4\cos^3(\theta) - 3\cos(\theta)$$

Setting $\theta = 2\pi/18$ with $\cos(2\pi/18) = \alpha$ gives

$$1/2 = \cos(\pi/3) = \cos(3\theta) = 4\alpha^3 - 3\alpha$$

Thus α is a root of $8x^3 - 6x - 1$. This is an irreducible polynomial, and is the minimal polynomial of α over the field $\mathbb{Q}$. Thus $[\mathbb{Q}(\alpha) : \mathbb{Q}] = \deg(p(x)) = 3$. This is not a power of 2 and so α cannot be constructible. □

Splitting fields and Galois groups

Let $f(x) \in F[x]$ be a polynomial with coefficients in a field F. A field extension E over F is called a *splitting field* for the polynomial $f(x)$ if there are $\alpha_1, \ldots, \alpha_n \in E$ such that $E = F(\alpha_1, \ldots, \alpha_n)$ and the polynomial splits into linear factors

$$f(x) = a(x - \alpha_1) \cdots (x - \alpha_n)$$

for some non-zero constant $a \in F$.

Examples

1. Let $f(x)$ be the polynomial $x^2 + 1 \in \mathbb{R}[x]$. As $f(x) = (x - i)(x + i)$ as a polynomial in $\mathbb{C}[x]$, a splitting field for $f(x)$ is given by $\mathbb{R}(i, -i) = \mathbb{R}(i) = \mathbb{C}$.

2. Let $f(x) = x^2 - 2 \in \mathbb{Q}[x]$. Then $f(x) = (x - \sqrt{2})(x + \sqrt{2})$. Thus a splitting field is given by the field extension $\mathbb{Q}(\sqrt{2})$ over $\mathbb{Q}$.

3. Let $f(x) = (x^2 - 2)(x^2 - 3) \in \mathbb{Q}[x]$. A splitting field for this polynomial is given by the field extension $\mathbb{Q}(\sqrt{2}, \sqrt{3})$ over $\mathbb{Q}$.

4. Let $f(x) = x^3 - 2 \in \mathbb{Q}[x]$. The complex roots of this polynomial are as follows:

$$\sqrt[3]{2}, \qquad \sqrt[3]{2} \cdot e^{2\pi i/3}, \qquad \sqrt[3]{2} \cdot e^{-2\pi i/3}$$

Thus a splitting field for $f(x)$ is given by $E = \mathbb{Q}(\sqrt[3]{2}, \sqrt[3]{2}e^{2\pi i/3}, \sqrt[3]{2}e^{-2\pi i/3})$. Note that

$$\sqrt[3]{2} \cdot e^{-2\pi i/3} = \left(\sqrt[3]{2}\right)^2 \left(\sqrt[3]{2} \cdot e^{2\pi i/3}\right)^{-1}$$

so the third root is already in the field which is generated by the first two roots. In other words, our splitting field can be written $E = \mathbb{Q}(\sqrt[3]{2}, \sqrt[3]{2}e^{2\pi i/3})$. Furthermore, clearly $\sqrt[3]{2}e^{2\pi i/3}$ can be replaced by $e^{2\pi i/3}$ so that $E = \mathbb{Q}(\sqrt[3]{2}, e^{2\pi i/3})$. Finally,

$$e^{2\pi i/3} = -\frac{1}{2} + \frac{\sqrt{3}}{2} i$$

and so $e^{2\pi/3}$ can be replaced by $\sqrt{-3}$. In summary, we have the identification

$$E = \mathbb{Q}(\sqrt[3]{2}, \sqrt{-3})$$

We leave as an exercise that also $E = \mathbb{Q}(\sqrt[3]{2} + e^{2\pi i/3})$.

5. Let $f(x) = x^4 - 2 \in \mathbb{Q}[x]$. The complex roots of this polynomial are $\sqrt[4]{2}, -\sqrt[4]{2}, \sqrt[4]{2}i, -\sqrt[4]{2}i$. We have the splitting field $\mathbb{Q}(\sqrt[4]{2}, \sqrt[4]{2}i) = \mathbb{Q}(\sqrt[4]{2}, i)$.

It is a basic result, which we omit, that for a given non-constant polynomial $f(x) \in F[x]$, a splitting field exists for $f(x)$; and moreover, any two such splitting fields are isomorphic. Thus given $f(x) \in F[x]$ we may speak of *the* splitting field of $f(x)$.

▶ **Given a field extension E over F, we define**

$$\mathbf{Aut}(E/F) = \{\sigma : E \to E \ : \ \sigma \textbf{ is an isomorphism}, \ \sigma(a) = a \textbf{ for all } a \in F\}$$

This is naturally a group, called the automorphism group of the field extension.

Proof. Let $\sigma, \sigma' \in \mathrm{Aut}(E/F)$ so that $\sigma, \sigma' : E \to E$ are isomorphisms each fixing F, i.e. $\sigma(a) = \sigma'(a) = a$ for all $a \in F$. Then the composition $\sigma \circ \sigma'$ is an isomorphism of E and $(\sigma \circ \sigma')(a) = \sigma(\sigma'(a)) = \sigma(a) = a$. Thus $\sigma \circ \sigma' \in \mathrm{Aut}(E/F)$. Further, $a = \mathrm{id}_E(a) = (\sigma^{-1} \circ \sigma)(a) = \sigma^{-1}(\sigma(a)) = \sigma^{-1}(a)$ and thus $\sigma^{-1} \in \mathrm{Aut}(E/F)$. Thus $\mathrm{Aut}(E/F)$ is a group. □

A finite field extension E over a subfield of $\mathbb{C}$ is called *Galois* if it is a splitting field of some polynomial $f(x) \in \mathbb{Q}[x]$. All examples above are Galois extensions. More generally, a finite extension E over an arbitrary field F is *Galois* if it is the splitting field of a polynomial $f(x) \in F[x]$ which has no repeated roots in E. We mainly focus on finite extensions of $\mathbb{Q}$ and other subfields of $\mathbb{C}$, in which case Galois fields can be identified with splitting fields.

▶ **If E is a Galois extension of a field F we write**

$$\mathbf{Gal}(E/F) = \mathbf{Aut}(E/F)$$

and call this the *Galois group* of the field extension.

As a (finite) Galois extension E over F is the splitting field of some polynomial $f(x) \in F[x]$, we also call $\mathrm{Gal}(E/F)$ the *Galois group of the polynomial* $f(x)$.

▶ **Let $E = F(\alpha_1, \ldots, \alpha_n)$ be the splitting field of $f(x) \in F[x]$ whose roots in E are $\alpha_1, \ldots, \alpha_n$. Then there is a 1-1 homomorphism**

$$\phi : \mathbf{Gal}(E/F) \longrightarrow S_n$$

to the symmetric group S_n. If $f(x) \in F[x]$ is irreducible, then the image subgroup $\mathbf{im}(\phi) \subset S_n$ is a transitive subgroup of S_n.

We remark that a *transitive subgroup* $G \subset S_n$ is a subgroup of the symmetric group S_n such that for all $i, j \in \{1, \ldots, n\}$ there is a permutation $f \in G$ such that $f(i) = j$.

Proof. First, write $f(x) = a_n x^n + \cdots + a_0$ where each $a_i \in F$. Consider a root $\alpha_i \in E$ of $f(x)$, and some $\sigma \in \mathrm{Gal}(E/F)$. Using that $\sigma : E \to E$ is a homomorphism fixing F we compute:

$$\begin{aligned} f(\sigma(\alpha_i)) &= a_n(\sigma(\alpha_i))^n + a_{n-1}(\sigma(\alpha_i))^{n-1} + \cdots + a_1\sigma(\alpha_i) + a_0 \\ &= a_n(\sigma(\alpha_i^n)) + a_{n-1}(\sigma(\alpha_i^{n-1})) + \cdots + a_1\sigma(\alpha_i) + a_0 \\ &= \sigma(a_n(\alpha_i^n)) + \sigma(a_{n-1}(\alpha_i^{n-1})) + \cdots + \sigma(a_1\alpha_i) + \sigma(a_0) \\ &= \sigma(a_n\alpha_i^n + a_{n-1}\alpha_i^{n-1} \cdots + a_1\alpha_i + a_0) = \sigma(0) = 0 \end{aligned}$$

Thus $\sigma(\alpha_i)$ is also a root of $f(x)$, and therefore $\sigma(\alpha_i) = \alpha_j$ for some $j \in \{1, \ldots, n\}$. Since E is Galois over F, the roots are distinct, so the index j is uniquely determined.

Next, define $\phi : \mathrm{Gal}(E/F) \to S_n$ by $\phi(\sigma)(i) = j$ if and only if $\sigma(\alpha_i) = \alpha_j$. In other words, we look at how the automorphism $\sigma : E \to E$ permutes the roots $\alpha_1, \ldots, \alpha_n$. It is straightforward from this construction that ϕ is a homomorphism.

Next, recall $E = F(\alpha_1, \ldots, \alpha_n)$. As $\sigma \in \mathrm{Gal}(E/F)$ fixes F, it is entirely determined by where it sends the elements α_i. This shows that ϕ is 1-1.

Finally, suppose $f(x)$ is irreducible. Then $F(\alpha_i) \cong F[x]/f(x)$ for any root α_i of $f(x)$. In particular, for α_i and α_j any two roots of $f(x)$ that lie in E we have an isomorphism $F(\alpha_i) \cong F[x]/f(x) \cong F(\alpha_j)$. This can be extended to an isomorphism $E \to E$ (we omit this step), from which the result follows. □

The above result shows that the Galois group of E over F is isomorphic to a subgroup of S_n, which has order $n!$. We obtain:

▶ **Let E be the splitting field of $f(x) \in F[x]$ whose degree is n. Then**

$$|\mathbf{Gal}(E/F)| \quad \textbf{divides} \quad n!.$$

Examples

1. Let $f(x)$ be the polynomial $x^2 + 1 \in \mathbb{R}[x]$, whose splitting field is $\mathbb{C} = \mathbb{R}(i, -i)$. Label the roots of $f(x)$ by $\alpha_1 = i$ and $\alpha_2 = -i$. Note that complex conjugation $\sigma : \mathbb{C} \to \mathbb{C}$ defined by

$$\sigma(a + bi) = a - bi$$

is an element of $\mathrm{Gal}(\mathbb{C}/\mathbb{R})$ of order 2. Thus $\{\mathrm{id}_{\mathbb{C}}, \sigma\} \subset \mathrm{Gal}(\mathbb{C}/\mathbb{R})$. As the Galois group has order dividing $2! = 2$, this is in fact an equality of groups. Thus $\mathrm{Gal}(\mathbb{C}/\mathbb{R}) \cong \mathbb{Z}_2$.

2. Let $f(x) = x^2 - 2 \in \mathbb{Q}[x]$ with splitting field $\mathbb{Q}(\sqrt{2})$. Similar to the previous example, there is only one non-trivial automorphism $\sigma \in \mathrm{Gal}(\mathbb{Q}(\sqrt{2}), \mathbb{Q})$ and it is given by $\sigma(a + b\sqrt{2}) = a - b\sqrt{2}$. We have $\mathrm{Gal}(\mathbb{Q}(\sqrt{2})/\mathbb{Q}) \cong \mathbb{Z}_2$.

3. Let $f(x) = (x^2 - 2)(x^2 - 3) \in \mathbb{Q}[x]$ with splitting field $E = \mathbb{Q}(\sqrt{2}, \sqrt{3})$. Recall that any element in E can be written uniquely as

$$a + b\sqrt{2} + c\sqrt{3} + d\sqrt{2}\sqrt{3}$$

where $a, b, c, d \in \mathbb{Q}$. Define automorphisms $\sigma_1, \sigma_2, \sigma_3 \in \mathrm{Gal}(E/\mathbb{Q})$ by

$$\begin{aligned}
\sigma_1(a + b\sqrt{2} + c\sqrt{3} + d\sqrt{2}\sqrt{3}) &= a - b\sqrt{2} + c\sqrt{3} - d\sqrt{2}\sqrt{3} \\
\sigma_2(a + b\sqrt{2} + c\sqrt{3} + d\sqrt{2}\sqrt{3}) &= a + b\sqrt{2} - c\sqrt{3} - d\sqrt{2}\sqrt{3} \\
\sigma_3(a + b\sqrt{2} + c\sqrt{3} + d\sqrt{2}\sqrt{3}) &= a - b\sqrt{2} - c\sqrt{3} - d\sqrt{2}\sqrt{3}
\end{aligned}$$

Labelling the roots of $f(x)$ as $\alpha_1 = \sqrt{2}, \alpha_2 = -\sqrt{2}, \alpha_3 = \sqrt{3}, \alpha_4 = -\sqrt{3}$ we have

$$\phi(\sigma_1) = (12), \quad \phi(\sigma_2) = (34), \quad \phi(\sigma_3) = (12)(34)$$

Because no automorphism can interchange $\pm\sqrt{2}$ with $\pm\sqrt{3}$, these are all of them. Thus

$$\mathrm{Gal}(E/\mathbb{Q}) = \{\mathrm{id}_E, \sigma_1, \sigma_2, \sigma_3\} \cong \{e, (12), (34), (12)(34)\} \subset S_4$$

Note that this group is isomorphic to $\mathbb{Z}_2 \times \mathbb{Z}_2$. Note in this example that $f(x)$ is not irreducible, and that the corresponding permutation group is not transitive.

4. Let $f(x) = x^3 - 2 \in \mathbb{Q}[x]$ with splitting field $E = \mathbb{Q}(\sqrt[3]{2}, \sqrt{-3})$. The roots of $f(x)$ are

$$\alpha_1 = \sqrt[3]{2}, \qquad \alpha_2 = \sqrt[3]{2} \cdot e^{2\pi i/3}, \qquad \alpha_3 = \sqrt[3]{2} \cdot e^{-2\pi i/3}$$

A direct computation shows that every permutation of these 3 roots extends to define an automorphism, and thus $\mathrm{Gal}(E/\mathbb{Q}) \cong S_3$.

5. Let $f(x) = x^4 - 2 \in \mathbb{Q}[x]$ with splitting field $E = \mathbb{Q}(\sqrt[4]{2}, i)$. The roots of $f(x)$ are

$$\alpha_1 = \sqrt[4]{2}, \quad \alpha_2 = -\sqrt[4]{2}, \quad \alpha_3 = \sqrt[4]{2}i, \quad \alpha_4 = -\sqrt[4]{2}i$$

Note an automorphism of this extension is determined by where it sends $\alpha_1 = \sqrt[4]{2}$ and i. There are automorphisms $\sigma, \tau \in \mathrm{Gal}(E/\mathbb{Q})$ such that:

$$\sigma(\sqrt[4]{2}) = \sqrt[4]{2}i, \qquad \sigma(i) = i$$

$$\tau(\sqrt[4]{2}) = \sqrt[4]{2}, \qquad \tau(i) = -i$$

We have $\phi(\sigma) = (1324)$ and $\phi(\tau) = (34)$. Any other automorphism is a composition of these two automorphisms. The Galois group is isomorphic to the following subgroup of S_4:

$$\mathrm{Gal}(E/\mathbb{Q}) \cong \{e, (12), (34), (12)(34), (1324), (1423), (13)(24), (14)(23)\} \subset S_4$$

In fact this subgroup is isomorphic to the dihedral group D_4, the symmetries of a square.

The Galois Theorem

Let us recall our setup from last lecture. For simplicity we will work with field extensions over $F \subset \mathbb{C}$, although much of what we will say holds in greater generality. Let E be a finite extension of F. Recall that E is a *Galois* extension, or a *splitting field*, if there is some polynomial $f(x) \in F[x]$ such that over the complex numbers, say, we have

$$f(x) = a(x - \alpha_1) \cdots (x - \alpha_n)$$

where $a \in F$, and $E = F(\alpha_1, \ldots, \alpha_n)$. We introduced for any extension E over F the automorphism group $\mathrm{Aut}(E/F)$ consisting of field isomorphisms $\sigma : E \to E$ which fix F in the sense that $\sigma(a) = a$ for all $a \in F$. In case E is Galois, we write

$$\mathrm{Gal}(E/F) = \mathrm{Aut}(E/F)$$

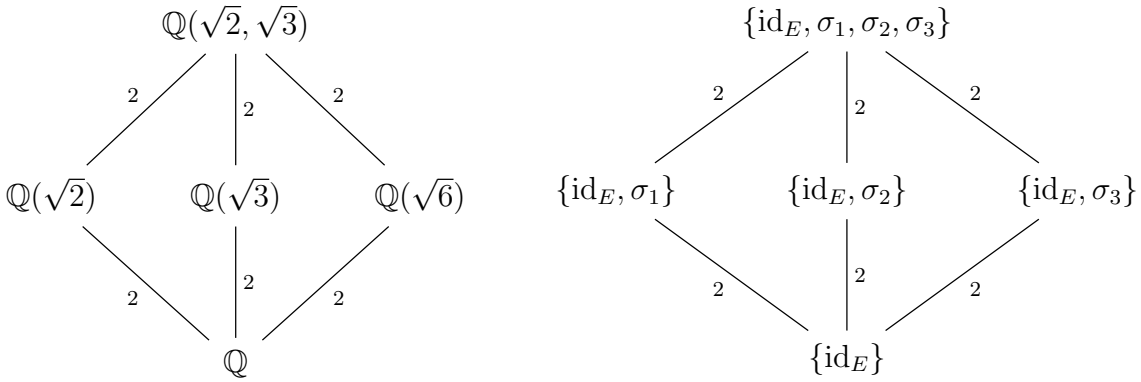
and call this the *Galois group* of the field extension. The main result from last lecture gives us an isomorphism from $\mathrm{Gal}(E/F)$ to a subgroup of the symmetric group S_n, where n is the number of roots as above.

Example: Consider the field $E = \mathbb{Q}(\sqrt{2}, \sqrt{3})$. We saw last lecture that this is a Galois extension over $\mathbb{Q}$: it is the splitting field for the polynomial $f(x) = (x^2 - 2)(x^2 - 3)$ with roots $\sqrt{2}, -\sqrt{2}, \sqrt{3}, -\sqrt{3}$. The Galois group $\mathrm{Gal}(E/\mathbb{Q})$ is isomorphic to $\mathbb{Z}_2 \times \mathbb{Z}_2$. In fact

$$\mathrm{Gal}(E/\mathbb{Q}) = \{\mathrm{id}_E, \sigma_1, \sigma_2, \sigma_3\}$$

where σ_1 has the effect of interchanging $\sqrt{2}$ with $-\sqrt{2}$, but fixes $\sqrt{3}$; σ_2 has the effect of interchanging $\sqrt{3}$ with $-\sqrt{3}$, but fixes $\sqrt{2}$; and $\sigma_3 = \sigma_1 \circ \sigma_2$.

Note that $E = \mathbb{Q}(\sqrt{2}, \sqrt{3})$ can also be viewed as an extension of $\mathbb{Q}(\sqrt{2})$ and of $\mathbb{Q}(\sqrt{3})$, and also $\mathbb{Q}(\sqrt{6})$. We can write all these extensions in a diagram (below on the left); in general, if a line connects E and F and E appears above F, then E is an extension field of F.



We have also indicated the degrees of the extensions; in this case they are all 2. On the right we have drawn instead the diagram of subgroups of the group $\mathrm{Gal}(E/\mathbb{Q}) = \{\mathrm{id}_E, \sigma_1, \sigma_2, \sigma_3\}$. Each subgroup has index 2, as indicated at the lines.

The fundamental theorem of Galois theory states the precise relationship between these two situations, and gives a correspondence between certain field extensions and subgroups.

▶ **(The Galois Theorem) Let E be a finite Galois extension of $F \subset \mathbb{C}$. Then**

$$[E : F] = |\mathbf{Gal}(E/F)|$$

For each field K with $F \subset K \subset E$ we have that E is Galois over K. The assignment $K \mapsto \mathbf{Gal}(E/K)$ induces a 1-1 and onto correspondence

$$\{\textbf{fields } K : \ F \subset K \subset E\} \longrightarrow \{\textbf{subgroups of } \mathbf{Gal}(E/F)\}$$

Further, we have $F \subset K \subset L \subset E$ if and only if we have the sequence of inclusions

$$\{e\} = \mathbf{Gal}(E/E) \subset \mathbf{Gal}(E/L) \subset \mathbf{Gal}(E/K) \subset \mathbf{Gal}(E/F)$$

Finally, a field K with $F \subset K \subset E$ is Galois over F if and only if $\mathbf{Gal}(E/K)$ is a normal subgroup of $\mathbf{Gal}(E/F)$, and if this is the case, we have

$$\mathrm{Gal}(K/F) \cong \frac{\mathrm{Gal}(E/F)}{\mathrm{Gal}(E/K)}$$

Due to a lack of time we will unfortunately omit the proof.

Consider the following problem: given a polynomial $f(x) \in \mathbb{Q}[x]$, we know it has some complete set of roots in the complex numbers, say $\alpha_1, \ldots, \alpha_n$. Thus over $\mathbb{C}$ we can write

$$f(x) = a(x - \alpha_1) \cdots (x - \alpha_n)$$

where $a \in \mathbb{Q}$ and each α_i is some complex number. Can we obtain each root α_i from the rational numbers by taking successive radicals? That is to say, can we write each α_i as an expression involving only the coefficients of $f(x)$, the operations of addition, multiplication, subtraction and division, and also k^{th} roots? The quadratic formula says the answer is "yes" when $f(X)$ is degree 2, and in fact the answer is "yes" for $\deg(f(x)) \leqslant 4$. However, it is not always possible beyond these cases:

▶ **(Abel–Ruffini Theorem) For any $n \geqslant 5$ there are polynomials of degree n such the roots cannot be solved in terms of radicals.**

We sketch a proof of this theorem by exhibiting a quintic which is not solvable in terms of radicals. We begin by recasting what we mean by solvable in terms of field theory: if we can solve for the roots of $f(x)$ in terms of radicals, then there is a chain of field extensions

$$\mathbb{Q} = F_0 \subset F_1 \subset F_2 \subset \cdots \subset F_{k-1} \subset F_k = E$$

such that E is the splitting field of $f(x)$ and each extension

$$F_i \subset F_{i+1}$$

is obtained by appending a radical, in the sense that $F_{i+1} = F_i(\beta_i)$ where $\beta_i^{n_i} \in F_i$. It can be shown that the special structure of this chain of field extensions along with the Galois Theorem implies that $\mathrm{Gal}(E/\mathbb{Q})$ is a *solvable group*. In general, a solvable group is a group G that admits a sequence of subgroups $H_i \subset G$ as depicted below

$$\{e\} = H_0 \subset H_1 \subset H_2 \subset \cdots \subset H_{k-1} \subset H_k = G$$

such that each H_i is normal in H_{i+1} and the factor group H_{i+1}/H_i is abelian. One can prove:

▶**The symmetric group S_n is solvable if and only if $n \leqslant 4$.**

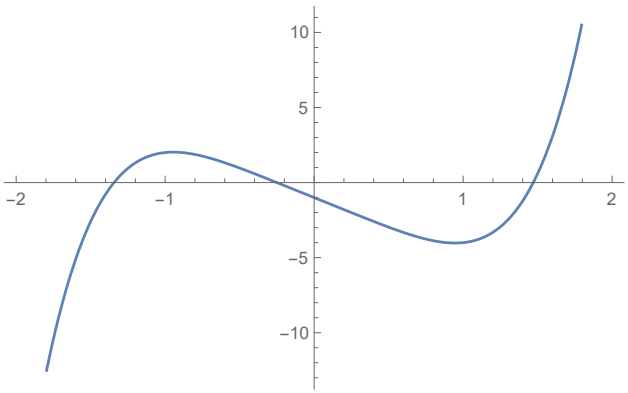
Thus to prove the Abel–Ruffini theorem for quintics, it suffices to find an irreducible polynomial $f(x)$ of degree 5 whose Galois group is isomorphic to S_5.

We know in this case that $\mathrm{Gal}(E/\mathbb{Q})$ is isomorphic to a subgroup G of S_5. Furthermore, from last lecture we know that this subgroup $G \subset S_5$ is *transitive*, that is, given any $i, j \in \{1, 2, 3, 4, 5\}$ we can find $g \in G$ such that $g(i) = j$. A direct computation shows:

▶**If $G \subset S_5$ is transitive and contains a transposition, then $G = S_5$.**

It remains then to show that we can find an irreducible quintic $f(x) \in \mathbb{Q}[x]$ whose Galois group, viewed as a permutation group of the 5 roots, contains a transposition. For this it suffices to find an irreducible quintic with rational coefficients that has exactly 2 complex roots, for then complex conjugation will provide the corresponding transposition!

Example: Consider $f(x) = x^5 - 4x - 1$. One can check this is irreducible.



From its graph, we see that there are exactly 3 real roots. Thus there are 2 complex roots, and complex conjugation gives an element of the Galois group which acts as a transposition on the roots. Thus the Galois group, being a transitive subgroup of S_5 with a transposition, must be all of S_5. This is not solvable, and so the roots of this quintic are not expressible in terms of radicals.

References

[1] Artin, M. *Algebra*, second ed. Pearson Education, 2011.

[2] Bump, D. Rubik's Cube notes. http://sporadic.stanford.edu/bump/match/rubik.pdf. Accessed in 2021.

[3] Herstein, I. N. *Abstract algebra*, third ed. Prentice Hall, Inc., Upper Saddle River, NJ, 1996. With a preface by Barbara Cortzen and David J. Winter.

[4] Judson, T. W. *Abstract Algebra: Theory and Applications*. Orthogonal Publishing, 2026.

[5] Koblitz, N. *A Course in Number Theory and Cryptography*, second ed., vol. 114 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1994.

[6] Lang, S. *Algebra*, third ed., vol. 211 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 2002.