

Introduction to Modern Algebra I

Math 817

and

Introduction to Modern Algebra II

Math 818

University of Nebraska-Lincoln

February 19, 2026

Warning!

Proceed with caution. These notes are under construction and are 100% guaranteed to contain typos. If you find any typos or errors, I will be most grateful to you for letting me know.

Acknowledgements

These notes are heavily based on notes by Mark Walker and Alexandra Seceleanu. Thanks the following students, to who found typos in previous versions of the notes: Gabriel Adams, Cal Heldt, Louis Burns, Daniel Gonzalez, Reese White, Wynter Sanderlin, Stephen Stern, Nicole Xie, and Isaiah Martinez.

Contents

Part I

Groups

Chapter 1

Groups: an introduction

Many mathematical structures consist of a set with special properties. Groups are elementary algebraic structures that allow us to deal with many objects of interest, such as geometric shapes and polynomials.

1.1 Definitions and first examples

Definition 1.1. A **binary operation** on a set S is a function $S \times S \rightarrow S$. If the binary operation is denoted by \cdot , we write $x \cdot y$ for the image of (x, y) under the binary operation \cdot .

Remark 1.2. We often write xy instead of $x \cdot y$ if the operation is clear from context.

Remark 1.3. We say that that a set S is closed under the operation \cdot when we want to emphasize that for any $x, y \in S$ the result xy of the operation is an element of S . But note that closure is really part of the definition of a binary operation on a set, and it is implicitly assumed whenever we consider such an operation.

Definition 1.4. A **group** is a set G equipped with a binary operation \cdot on G called the **group multiplication**, satisfying the following properties:

- **Associativity:** For every $x, y, z \in G$, we have $(x \cdot y) \cdot z = x \cdot (y \cdot z)$.
- **Identity element:** There exists $e \in G$ such that $e \cdot x = x \cdot e = x$ for all $x \in G$.
- **Inverses:** For each $x \in G$, there is an element $y \in G$ such that $xy = e = yx$.

The element e is called the **identity element** or simply **identity** of the group. For each element $x \in G$, an element $y \in G$ such that $xy = e = yx$ is called an **inverse** of x . We may write that (G, \cdot) is a group to mean that G is a group with the operation \cdot .

The **order** of the group G is the number of elements in the underlying set.

Remark 1.5. Although a group is the set *and* the operation, we will usually refer to the group by only naming the underlying set, G .

Remark 1.6. A set G equipped with an associative binary operation is a **semigroup**; if a semigroup also has an identity element, it is a **monoid**.

While we will not be discussing semigroups nor monoids that are not groups in this class, they can be useful and interesting objects. We will however include some fun facts about monoids in the remarks. In particular, there will be no monoids whatsoever in the qualifying exam.

Lemma 1.7. For any group G , we have the following properties:

- (1) The identity is unique: there exists a unique $e \in G$ with $ex = x = xe$ for all $x \in G$.
- (2) Inverses are unique: for each $x \in G$, there exists a unique $y \in G$ such that $xy = e = yx$.

Proof. Suppose e and e' are two identity elements; that is, assume e and e' satisfy $ex = x = xe$ and $e'x = x = xe'$ for all $x \in G$. Then

$$e = ee' = e'.$$

Now given $x \in G$, suppose y and z are two inverses for x , meaning that $yx = xy = e$ and $zx = xz = e$. Then

$$\begin{aligned} z &= ez && \text{since } e \text{ is the identity} \\ &= (yx)z && \text{since } y \text{ is an inverse for } x \\ &= y(xz) && \text{by associativity} \\ &= ye && \text{since } z \text{ is an inverse for } x \\ &= y && \text{since } e \text{ is the identity. } \square \end{aligned}$$

Remark 1.8. Note that our proof of Lemma 1.7 also applies to show that the identity element of a monoid is unique.

Given a group G , we can refer to *the* identity of G . Similarly, given an element $x \in G$, we can refer to *the* inverse of x .

Notation 1.9. Given an element x in a group G , we write x^{-1} to denote its unique inverse.

Remark 1.10. In a monoid G with identity e , an element x might have a **left inverse**, which is an element y satisfying $yx = e$. Similarly, x might have a **right inverse**, which is an element z satisfying $xz = e$. An element in a monoid might have several distinct right inverses, or several distinct left inverses, but if it has both a left and a right inverse, then it has a unique left inverse and a unique right inverse, and those elements coincide.

Exercise 1. Give an example of a monoid M and an element in M that has a left inverse but not a right inverse.

Definition 1.11. Let G be a group, $x \in G$, and $n \geq 1$ be an integer. We write x^n to denote the element obtained by multiplying x with itself n times:

$$x^n := \underbrace{x \cdots x}_{n \text{ times}}.$$

Exercise 2 (Properties of group elements). Let G be a group and let $x, y, z, a_1, \dots, a_n \in G$. Show that the following properties hold:

- (1) If $xy = xz$, then $y = z$.
- (2) If $yx = zx$, then $y = z$.
- (3) $(x^{-1})^{-1} = x$.
- (4) $(a_1 \dots a_n)^{-1} = a_n^{-1} \dots a_1^{-1}$.
- (5) $(x^{-1}yx)^n = x^{-1}y^n x$ for any integer $n \geq 1$.
- (6) $(x^{-1})^n = (x^n)^{-1}$.

Notation 1.12. Given a group G , an element $x \in G$, and a positive integer n , we write $x^{-n} := (x^n)^{-1}$.

Note that by Exercise 2, $x^{-n} = (x^{-1})^n$.

Exercise 3. Let G be a group and consider $x \in G$. Show that $x^a x^b = x^{a+b}$.

Definition 1.13. A group G is **abelian** if \cdot is commutative, meaning that $x \cdot y = y \cdot x$ for all $x, y \in G$.

Often, but not always, the group operation for an abelian group is written as $+$ instead of \cdot . In this case, the identity element is usually written as 0 and the inverse of an element x is written as $-x$.

Example 1.14.

- (1) The **trivial group** is the group with a single element $\{e\}$. This is an abelian group.
- (2) The pairs $(\mathbb{Z}, +)$, $(\mathbb{Q}, +)$, $(\mathbb{R}, +)$ and $(\mathbb{C}, +)$ are abelian groups.
- (3) For any n , let \mathbb{Z}/n denote the integers modulo n . Then $(\mathbb{Z}/n, +)$ is an abelian group where $+$ denotes addition modulo n .
- (4) For any field F , such as \mathbb{Q} , \mathbb{R} , \mathbb{C} or \mathbb{Z}/p for a prime p , the set $F^\times := F \setminus \{0\}$ is an abelian group under multiplication. We will later formally define what a field is, but these fields might already be familiar to you.

Example 1.15. Let F be any field. If you are not yet familiar with fields, the real or complex numbers are excellent examples. Consider a positive integer n , and let

$$\mathrm{GL}_n(F) := \{\text{invertible } n \times n \text{ matrices with entries in } F\}.$$

An invertible matrix is one that has a two-sided (multiplicative) inverse. It turns out that if an $n \times n$ matrix M has a left inverse N then that inverse N is automatically a right inverse too, and vice-versa; this is a consequence of a more general fact we mentioned in Remark 1.10.

It is not hard to see that $\mathrm{GL}_n(F)$ is a nonabelian group under matrix multiplication. Note that $(\mathrm{GL}_1(F), \cdot)$ is simply (F^\times, \cdot) .

Even if the group is not abelian, the set of elements that commute with every other element is particularly important.

Definition 1.16. Let G be a group. The **center** of G is the set

$$Z(G) := \{x \in G \mid xy = yx \text{ for all } y \in G\}.$$

Remark 1.17. Note that the center of any group always includes the identity. Whenever $Z(G) = \{e_G\}$, we say that the center of G is trivial.

Remark 1.18. Note that G is abelian if and only if $Z(G) = G$.

One might describe a group by giving a presentation.

Informal definition 1.19. A **presentation** for a group is a way to specify a group in the following format:

$$G = \langle \text{set of generators} \mid \text{set of relations} \rangle.$$

A set S is said to **generate** or be a **set of generators** for G if every element of the group can be expressed in some way as a product of finitely many of the elements of S and their inverses (with repetitions allowed). A **relation** is an identity satisfied by some expressions involving the generators and their inverses. We usually record just enough relations so that every valid equation involving the generators is a consequence of those listed here and the axioms of a group.

Remark 1.20. We can only take products of finitely many of our generators and their inverses because we do not have a way to make sense of infinite products.

Note, however, that the set of generators and the set of relations are allowed to be infinite.

Example 1.21. The group \mathbb{Z} has one generator, the element 1, which satisfies no relations.

Example 1.22. The following is a presentation for the group \mathbb{Z}/n of integers modulo n :

$$\mathbb{Z}/n = \langle x \mid x^n = e \rangle.$$

Definition 1.23. A group G is called **cyclic** if it is generated by a single element. A group G is **finitely generated** if it is generated by finitely many elements.

Example 1.24. We saw above that \mathbb{Z} and \mathbb{Z}/n are cyclic groups.

Exercise 4. Prove that every cyclic group is abelian.

Exercise 5. Prove that $(\mathbb{Q}, +)$ and $\mathrm{GL}_2(\mathbb{Z}_2)$ are not cyclic groups.

In general, given a presentation, it is very difficult to prove certain expressions are not actually equal to each other. In fact,

There is no algorithm that, given any group presentation as an input, can decide whether the group is actually the trivial group with just one element.

and perhaps more strikingly

There exist a presentation with finitely many generators and finitely many relations such that whether or not the group is actually the trivial group with just one element is *independent of the standard axioms of mathematics!*

We will now dedicate the next few sections to some classes of examples are very important.

1.2 Permutation groups

Definition 1.25. For any set X , the **permutation group** on X is the set $\text{Perm}(X)$ of all bijective functions from X to itself equipped with the binary operation given by composition of functions.

Notation 1.26. For an integer $n \geq 1$, we write $[n] := \{1, \dots, n\}$ and $S_n := \text{Perm}([n])$. An element of S_n is called a **permutation on n symbols**, sometimes also called a permutation on n letters or n elements.

We can write an element σ of S_n as a table of values:

i	1	2	3	...	n
$\sigma(i)$	$\sigma(1)$	$\sigma(2)$	$\sigma(3)$...	$\sigma(n)$

We may also represent this using arrows, as follows:

$$\begin{aligned} 1 &\longmapsto \sigma(1) \\ 2 &\longmapsto \sigma(2) \\ &\vdots \\ n &\longmapsto \sigma(n). \end{aligned}$$

Remark 1.27. To count the elements $\sigma \in S_n$, note that

- there are n choices for $\sigma(1)$;
- once $\sigma(1)$ has been chosen, we have $n - 1$ choices for $\sigma(2)$;

⋮

- once $\sigma(1), \dots, \sigma(n-1)$ have been chosen, there is a unique possible value for $\sigma(n)$, which is the only value left.

Thus the group S_n has $n!$ elements.

It is customary to use cycle notation for permutations.

Definition 1.28. If i_1, \dots, i_m are distinct integers between 1 and n , then $\sigma = (i_1 i_2 \dots i_m)$ denotes the element of S_n determined by

$$\sigma(i_1) = i_2, \quad \sigma(i_2) = i_3, \quad \dots, \quad \sigma(i_{m-1}) = i_m, \quad \text{and} \quad \sigma(i_m) = i_1,$$

and which fixes all elements of $[n] \setminus \{i_1, \dots, i_m\}$, meaning that

$$\sigma(j) = j \quad \text{for all } j \in [n] \text{ with } j \notin \{i_1, \dots, i_m\}.$$

Such a permutation is called a **cycle** or an **m-cycle** when we want to emphasize its length. In particular, we say that σ has length m .

Remark 1.29. A 1-cycle is the identity permutation.

Notation 1.30. A 2-cycle is often called a **transposition**.

Remark 1.31. The cycles $(i_1 \dots i_m)$ and $(j_1 \dots j_m)$ represent the same cycle if and only if the two lists i_1, \dots, i_m and j_1, \dots, j_m are cyclical rearrangements of each other. For example, $(1\ 2\ 3) = (2\ 3\ 1)$ but $(1\ 2\ 3) \neq (2\ 1\ 3)$.

Remark 1.32. Consider the m -cycle $\sigma = (i_1 \dots i_m)$. Then for any integer k , we have

$$\sigma^k(i_j) = i_{j+k} \pmod{m}.$$

Here we interpret $j + k \pmod{m}$ to denote the unique integer $0 \leq s < m$ such that

$$s \equiv j + k \pmod{m}.$$

Notation 1.33. We denote the product (composition) of the cycles $(i_1 \dots i_s)$ and $(j_1 \dots j_t)$ by juxtaposition; more precisely, $(i_1 \dots i_s)(j_1 \dots j_t)$ denotes the composition of the two cycles, read from right to left.

Example 1.34. We claim that the permutation group $\text{Perm}(X)$ is nonabelian whenever the set X has 3 or more elements. Indeed, given three distinct elements $x, y, z \in S$, consider the transpositions (xy) and (yz) . Now consider the permutations $(yz)(xy)$ and $(yz)(xy)$, where the composition is read from right to left, such as function composition. Then

$$\begin{array}{ccc} (yz)(xy) : & \begin{array}{c} x \xrightarrow{(xy)} y \xrightarrow{(yz)} z \\ y \xrightarrow{(xy)} x \xrightarrow{(yz)} x \\ z \xrightarrow{(xy)} z \xrightarrow{(yz)} y \end{array} & (xy)(yz) : \begin{array}{c} x \xrightarrow{(yz)} x \xrightarrow{(xy)} y \\ y \xrightarrow{(yz)} z \xrightarrow{(xy)} z \\ z \xrightarrow{(yz)} y \xrightarrow{(xy)} x \end{array} \end{array}$$

Note that $(yz)(xy) \neq (xy)(yz)$, since for example the first one takes x to z while the second one takes x to y .

Lemma 1.35. *Disjoint cycles commute; that is, if*

$$\{i_1, i_2, \dots, i_m\} \cap \{j_1, j_2, \dots, j_k\} = \emptyset$$

then the cycles

$$\sigma_1 = (i_1 i_2 \cdots i_m) \quad \text{and} \quad \sigma_2 = (j_1 j_2 \cdots j_k)$$

satisfy $\sigma_1 \circ \sigma_2 = \sigma_2 \circ \sigma_1$.

Proof. We need to show $\sigma_1(\sigma_2(l)) = \sigma_2(\sigma_1(l))$ for all $l \in [n]$. If $l \notin \{i_1, \dots, i_m, j_1, \dots, j_k\}$, Then $\sigma_1(l) = l = \sigma_2(l)$, so

$$\sigma_1(\sigma_2(l)) = \sigma_1(l) = l \quad \text{and} \quad \sigma_2(\sigma_1(l)) = \sigma_2(l) = l.$$

If $l \in \{j_1, \dots, j_k\}$, then $\sigma_2(l) \in \{j_1, \dots, j_k\}$ and hence, since the subsets are disjoint, l and $\sigma_2(l)$ are not in the set $\{i_1, i_2, \dots, i_m\}$. It follows that σ_1 preserves l and $\sigma_2(l)$, and thus

$$\sigma_1(\sigma_2(l)) = \sigma_2(l) \quad \text{and} \quad \sigma_2(\sigma_1(l)) = \sigma_2(l).$$

The case when $l \in \{i_1, \dots, i_m\}$ is analogous. \square

Theorem 1.36. *Each $\sigma \in S_n$ can be written as a product of disjoint cycles, and such a factorization is unique up to the order of the factors.*

Remark 1.37. For the uniqueness part of Theorem 1.36, one needs to establish a convention regarding 1-cycles: we need to decide whether the 1-cycles will be recorded. If we decide not to record 1-cycles, this gives the shorter version of our factorization into cycles. If all the 1-cycles are recorded, this gives a longer version of our factorization, but this option has the advantage that it makes it clear what the size n of our group S_n is. We will follow the first convention: we will write only m -cycles with $m \geq 2$. Under this convention, the identity element of S_n is the empty product of disjoint cycles. We will, however, sometimes denote the identity by (1) for convenience.

Proof. Fix a permutation σ . The key idea is to look at the *orbits* of σ : for each $x \in [n]$, its orbit by σ is the subset of $[n]$ of the form

$$O_x = \{\sigma(x), \sigma^2(x), \sigma^3(x), \dots\} = \{\sigma^i(x) \mid i \geq 1\}.$$

Notice that the orbits of two elements x and y are either the same orbit, which happens precisely when $y \in O_x$, or disjoint. Since $[n]$ is a finite set, and σ is a bijection of σ , we will eventually have $\sigma^i(x) = \sigma^j(x)$ for some $j > i$, but then

$$\sigma^{j-i}(x) = \sigma^{i-i}(x) = \sigma^0(x) = x.$$

Thus we can find the smallest positive integer n_x such that $\sigma^{n_x}(x) = x$. Now for each $x \in [n]$, we consider the cycle

$$\tau_x = (\sigma(x) \ \sigma^2(x) \ \sigma^3(x) \cdots \sigma^{n_x}(x)).$$

Now let S be a set of indices for the distinct τ_x , where note that we are not including the τ_x that are 1-cycles. We claim that we can factor σ as

$$\sigma = \prod_{i \in S} \tau_i.$$

To show this, consider any $x \in [n]$. It must be of the form $\sigma^j(i)$ for some $i \in S$, given that our choice of S was exhaustive. On the right hand side, only τ_i moves x , and indeed by definition of τ_i we have

$$\tau_i(x) = \sigma^{j+1}(i) = \sigma(\sigma^j(i)) = \sigma(x).$$

This proves that

$$\sigma = \prod_{i \in S} \tau_i.$$

As for uniqueness, note that if $\sigma = \tau_1 \cdots \tau_s$ is a product of disjoint cycles, then each $x \in [n]$ is moved by at most one of the cycles τ_i , since the cycles are all disjoint. Fix i such that τ_i moves x . We claim that

$$\tau_x = (\sigma(x) \ \sigma^2(x) \ \sigma^3(x) \cdots \sigma^{n_x}(x)).$$

This will show that our product of disjoint cycles giving σ is the same (unique) product we constructed above. To do this, note that we do know that there is some integer s such that $\tau_x^s(x) = e$, and

$$\tau_x = (\tau_x(x) \ \tau_x^2(x) \ \tau_x^3(x) \cdots \tau_x^s(x)).$$

Thus we need only to prove that

$$\tau_x^k(x) = \sigma^k(x)$$

for all integers $k \geq 1$. Now by Lemma 1.35, disjoint cycles commute, and thus for each integer $k \geq 1$ we have

$$\sigma^k = \tau_1^k \cdots \tau_s^k.$$

But τ_j fixes x whenever $j \neq i$, so

$$\sigma^k = \tau_i^k(x).$$

We conclude that the integer n_x we defined before is the length of the cycle τ_i , and that

$$\tau_i = (x \tau_i(x) \tau_i^2(x) \cdots \tau_i^{n_x-1}(x)) = (x \sigma(x) \sigma^2(x) \cdots \sigma^{n_x-1}(x)).$$

Thus this decomposition of σ as a product of disjoint cycles is the same decomposition we described above. \square

Example 1.38. Consider the permutation $\sigma \in S_5$ given by

$$\begin{aligned} 1 &\mapsto 3 \\ 2 &\mapsto 4 \\ 3 &\mapsto 5 \\ 4 &\mapsto 2 \\ 5 &\mapsto 1. \end{aligned}$$

Its decomposition into a product of disjoint cycles is

$$(135)(24).$$

Definition 1.39. The **cycle type** of an element $\sigma \in S_n$ is the unordered list of lengths of cycles that occur in the unique decomposition of σ into a product of disjoint cycles.

Example 1.40. The element

$$(3\ 4)(1\ 5)(2\ 6\ 7)(9\ 8\ 11)(15\ 16\ 17\ 105\ 114)$$

of S_{156} has cycle type $2, 2, 3, 3, 5$. Note here that the n of S_n is not recorded, but is implicit.

It is also useful to write permutations as products of (not necessarily disjoint) transpositions. First, we need the following exercise:

Exercise 6. Show that

$$(i_1\ i_2\ \cdots\ i_p) = (i_1\ i_p)(i_1\ i_{p-2})(i_1\ i_3)(i_1\ i_2)$$

for any $p \geq 2$.

Corollary 1.41. *The group S_n is generated by transpositions: every permutation is a product of transpositions.*

Proof. Given any permutation, we can decompose it as a product of cycles by Theorem 1.36. Thus it suffices to show that each cycle can be written as a product of permutations. For a cycle $(i_1 i_2 \cdots i_p)$, one can show that

$$(i_1 i_2 \cdots i_p) = (i_1 i_2)(i_2 i_3) \cdots (i_{p-2} i_{p-1})(i_{p-1} i_p),$$

which we leave as an exercise (see Exercise 6). □

Remark 1.42. Note however that when we write a permutation as a product of transpositions, such a product is no longer necessarily unique.

Example 1.43. If $n \geq 2$, the identity in S_n can be written as $(12)(12)$. In fact, any transposition is its own inverse, so we can write the identity as $(ij)(ij)$ for any $i \neq j$.

Exercise 7. Show that

$$(cd)(ab) = (ab)(cd) \quad \text{and} \quad (bc)(ab) = (ac)(bc)$$

for all distinct a, b, c, d in $[n]$.

Theorem 1.44. *Given a permutation $\sigma \in S_n$, the parity of the number of transpositions in any representation of σ as a product of transpositions depends only on σ .*

Proof. Suppose that σ is a permutation that can be written as a production of transpositions β_i and λ_j in two ways,

$$\sigma = \beta_1 \cdots \beta_s = \lambda_1 \cdots \lambda_t$$

where s is even and t is odd. As we noted in Example 1.43, every transposition is its own inverse, so we conclude that

$$e_{S_n} = \beta_1 \cdots \beta_s \lambda_t \cdots \lambda_1,$$

which is a product of $s + t$ transpositions. This is an odd number, so it suffices to show that it is not possible to write the identity as a product of an odd number of transpositions.

So suppose that the identity can be written as the product $(a_1 b_1) \cdots (a_k b_k)$, where each $a_i \neq b_i$. First, note that a single transposition *cannot* be the identity, and thus $k \neq 1$. So assume, for the sake of an argument by induction, that for a fixed k , we know that every product of fewer than k transpositions that equals the identity must use an even number of transpositions. We might as well have $k \geq 3$, since we 2 is even.

Now note that since $k > 1$, and our product is the identity, then some transposition $(a_i b_i)$ with $i > 1$ must move a_1 ; otherwise, b_1 would be sent to a_1 , and our product would not be the identity.

Now notice that the two rules in Exercise 7 allow us to rewrite the overall product without changing the number of transpositions in such a way that the transposition $(a_2 b_2)$ moves a_1 , meaning a_2 or b_2 is a_1 . So let us assume that our product of transpositions has already been put in this form. Note also that $(a_i b_i) = (b_i a_i)$, so we might as well assume without loss of generality that $a_2 = a_1$. We will consider the cases when $b_2 = b_1$ and $b_2 \neq b_1$.

Case 1: When $b_1 = b_2$, our product is

$$(a_1b_1)(a_1b_1)(a_3b_3) \cdots (a_kb_k),$$

but $(a_1b_1)(a_1b_1)$ is the identity, so we can rewrite our product using only $k - 2$ transpositions. By induction hypothesis, $k - 2$ is even, and thus k is even.

Case 2: When $b_1 \neq b_2$, we can use Exercise 7 to write

$$(a_1b_1)(a_1b_2) = (a_1b_1)(b_2a_1) = (a_1b_2)(b_1b_2).$$

Notice here that it matters that a_1 , b_1 , and b_2 are all distinct, so that we can apply Exercise 7. So our product, which equals the identity, is

$$(a_1b_2)(b_1b_2)(a_3b_3) \cdots (a_kb_k).$$

The advantage of this shuffling is that while we have only changed the first two transpositions, we have decreased the number of transpositions that move a_1 . We must now have some other transposition that moves a_1 , and we can repeat the argument to keep decreasing the number of transpositions in our product that move a_1 . Each time we do this, we cannot keep landing in case 2 indefinitely, as each time we lower the number of transpositions moving a_1 . So eventually we will land in case 1, which allows us to lower the total number of transpositions, and using the induction hypothesis we will show that k must be even. \square

Definition 1.45. Consider a permutation $\sigma \in S_n$. If $\sigma = \tau_1 \cdots \tau_s$ is a product of transpositions, the **sign** of σ is given by $(-1)^s$. Permutations with sign 1 are called **even** and those with sign -1 are called **odd**. This is also called the parity of the permutation.

Theorem 1.44 tells us that the sign of a permutation is well-defined.

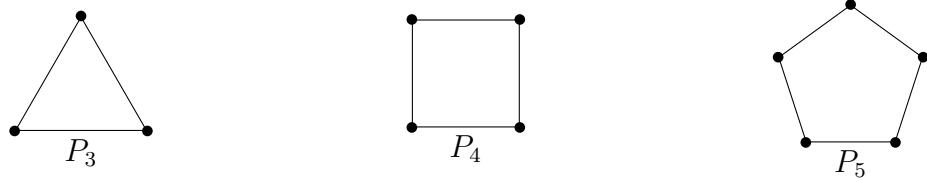
Example 1.46. The identity permutation is even. Every transposition is odd.

Example 1.47. The 3-cycle (123) can be rewritten as $(12)(23)$, a product of 2 transpositions, so the sign of (123) is 1.

Exercise 8. Show that every permutation is a product adjacent transpositions, meaning transpositions of the form $(i \ i + 1)$.

1.3 Dihedral groups

For any integer $n \geq 3$, let P_n denote a regular n -gon. For concreteness sake, let us imagine P_n is centered at the origin with one of its vertices located along the positive y -axis. Note that the size of the polygon will not matter. Here are some examples:



Definition 1.48. The **dihedral group** D_n is the set of symmetries of the regular n -gon P_n equipped with the binary operation given by composition.

Remark 1.49. There are competing notations for the group of symmetries of the n -gon. Some authors prefer to write it as D_{2n} , since, as we will show, that is the order of the group. Democracy has dictated that we will be denoting it by D_n , which indicates that we are talking about the symmetries of the n -gon. Some authors like to write $D_{2 \times n}$, always keeping the 2, for example with $D_{2 \times 3}$, to satisfy both camps.

Let us make this more precise. Let $d(-, -)$ denote the usual Euclidean distance between two points on the plane \mathbb{R}^2 . An **isometry** of the plane is a function $f : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ that is bijective and preserves the Euclidean distance, meaning that

$$d(f(A), f(B)) = d(A, B) \quad \text{for all } A, B \in \mathbb{R}^2.$$

Though not obvious, it is a fact that if f preserves the distance between every pair of points in the plane, then it must be a bijection.

A **symmetry** of P_n is an isometry of the plane that maps P_n to itself. By this I do not mean that f fixes each point of P_n , but rather that we have an equality of sets $f(P_n) = P_n$, meaning every point of P_n is mapped to a (possibly different) point of P_n and every point of P_n is the image of some point in P_n via f .

We are now ready to give the formal definition of the dihedral groups:

Remark 1.50. Let us informally verify that this really is a group. If f and g are in D_n , then $f \circ g$ is an isometry (since the composition of any two isometries is again an isometry) and

$$(f \circ g)(P_n) = f(g(P_n)) = f(P_n) = P_n,$$

so that $f \circ g \in D_n$. This proves composition is a binary operation on D_n . Now note that associativity of composition is a general property of functions. The identity function on \mathbb{R}^2 , denoted $\text{id}_{\mathbb{R}^2}$, belongs to D_n and it is the identity element of D_n . Finally, the inverse function of an isometry is also an isometry. Using this, we see that every element of D_n has an inverse.

Later on we will need the following elementary fact, which we leave as an exercise:

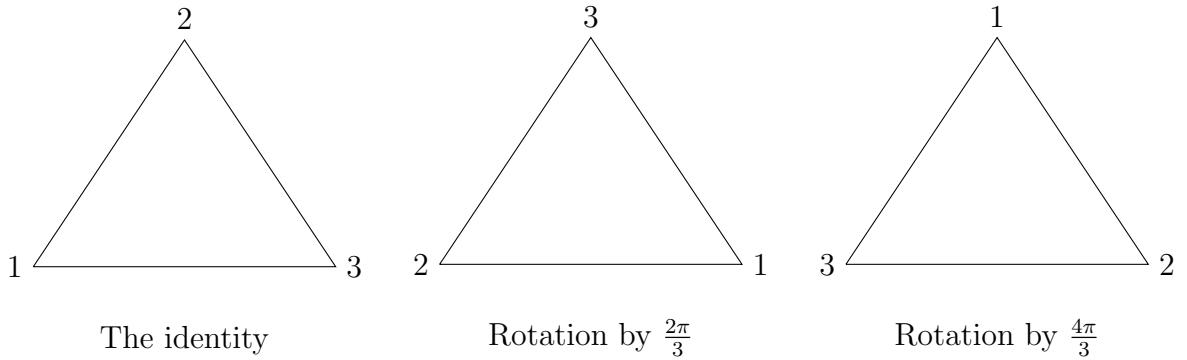
Lemma 1.51. Every point on a regular polygon is completely determined, among all points on the polygon, by its distances to two adjacent vertices of the polygon.

Exercise 9. Prove Lemma 1.51.

Definition 1.52 (Rotations in D_n). Assume that the regular n -gon P_n is drawn in the plane with its center at the origin and one vertex on the x axis. Let r denote the rotation about the origin by $\frac{2\pi}{n}$ radians counterclockwise; this is an element of D_n . Its inverse is the clockwise rotation by $\frac{2\pi}{n}$. This gives us rotations r^i , where r^i is the counterclockwise rotation by $\frac{2\pi i}{n}$, for each $i = 1, \dots, n$. Notice that when $i = n$ this is simply the identity map.

Each symmetry of P_n is completely determined by the images of the vertices. In particular, it is sometimes convenient to label the vertices of P_n with $1, 2, \dots, n$, and to indicate each symmetry by indicating the images of the vertices, as in the following example.

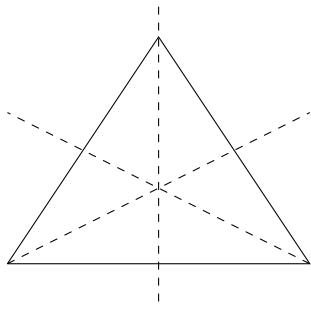
Example 1.53. Here are the rotations of D_3 :



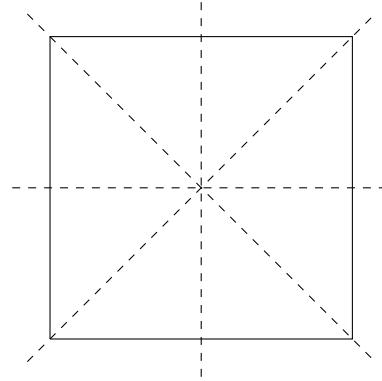
Definition 1.54 (Reflections in D_n). For any line of symmetry of P_n , reflection about that line gives an element of D_n . When n is odd, the line connecting a vertex to the midpoint of the opposite side of P_n is a line of symmetry. When n is even, there are two types of reflections: the ones about the line connecting two opposite vertices, and the ones across the line connecting midpoints of opposite sides.

In both cases, these give us a total of n reflections.

Example 1.55.



The reflection lines in D_3



The reflection lines in D_4

Let us summarize the content of this page:

Notation 1.56. Fix $n \geq 3$. We will consider two special elements of D_n :

- Let r denote the symmetry of P_n given by counterclockwise rotation by $\frac{2\pi}{n}$.
- Let s denote a reflection symmetry of P_n that fixes at least one of the vertices of P_n , as described in Definition 1.54. Let V_1 be a vertex of P_n that is fixed by s , and label the remaining vertices of P_n with V_2, \dots, V_n by going counterclockwise from V_1 .

From now on, whenever we are talking about D_n , the letters r and s will refer only to these specific elements. Finally, we will sometimes denote the identity element of D_n by id , since it is the identity map.

Theorem 1.57. *The dihedral group D_n has $2n$ elements.*

Proof. First, we show that D_n has order at most $2n$. Any element $\sigma \in D_n$ takes the polygon P_n to itself, and must in particular send vertices to vertices and preserve adjacencies, meaning that any two adjacent vertices remain adjacent after applying σ . Fix two adjacent vertices A and B . By Lemma 1.51, the location of every other point P on the polygon after applying σ is completely determined by the locations of $\sigma(A)$ and $\sigma(B)$. There are n distinct possibilities for $\sigma(A)$, since it must be one of the n vertices of the polygon. But once $\sigma(A)$ is fixed, $\sigma(B)$ must be a vertex adjacent to $\sigma(A)$, so there are at most 2 possibilities for $\sigma(B)$. This gives us at most $2n$ elements in D_n .

Now we need only to present $2n$ distinct elements in D_n . We have described n reflections and n rotations for D_n ; we need only to see that they are all distinct. First, note that the only rotation that fixes any vertices of P_n is the identity. Moreover, if we label the vertices of P_n in order with $1, 2, \dots, n$, say by starting in a fixed vertex and going counterclockwise through each adjacent vertex, then the rotation by an angle of $\frac{2\pi i}{n}$ sends V_1 to V_{i+1} for each $i < n$, showing these n rotations are distinct. Now when n is odd, each of the n reflections fixes exactly one vertex, and so they are all distinct and disjoint from the rotations. Finally, when n is even, we have two kinds of reflections to consider. The reflections through a line connecting opposite vertices have exactly two fixed vertices, and are completely determined by which two vertices are fixed; since rotations have no fixed points, none of these matches any of the rotations we have already considered. The other reflections, the ones through the midpoint of two opposite sides, are completely determined by (one of) the two pairs of adjacent vertices that they switch. No rotation switches two adjacent vertices, and thus these give us brand new elements of D_n .

In both cases, we have a total of $2n$ distinct elements of D_n given by the n rotations and the n reflections. \square

Remark 1.58. Given an element of D_n , we now know that it must be a rotation or a reflection. The rotations are the elements of D_n that preserve orientation, while the reflections are the elements of D_n that reverse orientation.

Remark 1.59. Any reflection is its own inverse. In particular, $s^2 = \text{id}$.

Remark 1.60. Note that $r^j(V_1) = V_{1+j \pmod n}$ for any j . Thus if $r^j = r^i$ for some $1 \leq i, j \leq n$, then we must have $i = j$.

In fact, we have seen that $r^n = \text{id}$ and that the rotations $\text{id}, r, r^2, \dots, r^{n-1}$ are all distinct, so $|r| = n$. In particular, the inverse of r is r^{n-1} .

Lemma 1.61. *Following Notation 1.56, we have $srs^{-1} = r^{-1}$.*

Proof. First, we claim that rs is a reflection. To see this, observe that $s(V_1) = V_1$, so

$$rs(V_1) = r(V_1) = V_2$$

and

$$rs(V_2) = r(V_n) = V_1.$$

This shows that rs must be a reflection, since it reverses orientation. Reflections have order 2, so $rsrs = (rs)^2 = \text{id}$ and hence $srs = r^{-1}$. \square

Remark 1.62. Given $|r| = n$ and $|s| = 2$, as noted in Remark 1.59 and Remark 1.60, we can rewrite Lemma 1.61 as

$$srs = r^{n-1}.$$

Exercise 10. Show that $sr^i s^{-1} = r^{n-i}$ for all i .

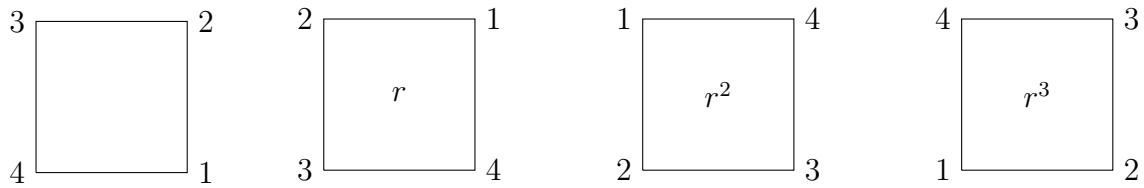
Theorem 1.63. *Every element in D_n can be written uniquely as r^j or $r^j s$ for $0 \leq j \leq n-1$.*

Proof. Let α be an arbitrary symmetry of P_n . Note α must fix the origin, since it is the center of mass of P_n , and it must send each vertex to a vertex because the vertices are the points on P_n at largest distance from the origin. Thus $\alpha(V_1) = V_j$ for some $1 \leq j \leq n$ and therefore the element $r^{-j}\alpha$ fixes V_1 and the origin. The only elements that fix V_1 are the identity and s . Hence either $r^{-j}\alpha = \text{id}$ or $r^{-j}\alpha = s$. We conclude that $\alpha = r^j$ or $\alpha = r^j s$.

Notice that we have shown that D_n has exactly $2n$ elements, and that there are $2n$ distinct expressions of the form r^j or $r^j s$ for $0 \leq j \leq n-1$. Thus each element of D_n can be written in this form in a unique way. \square

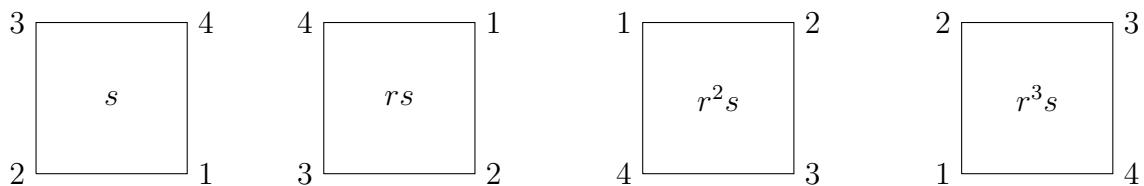
Remark 1.64. The elements s, rs, \dots, r^{n-1} are all reflections since they reverse orientation. Alternatively, we can check these are all reflections by checking they have order 2. As we noted before, the elements $\text{id}, r, \dots, r^{n-1}$ are rotations, and preserve orientation.

Example 1.65. The 8 elements of D_4 , the group of symmetries of the square, are



The identity Rotation by $\frac{2\pi}{4} = \frac{\pi}{2}$ Rotation by $\frac{4\pi}{4} = \pi$ Rotation by $\frac{6\pi}{4} = \frac{3\pi}{2}$

and the reflections



Let us now give a presentation for D_n .

Theorem 1.66. *Let $r : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ denote counterclockwise rotation around the origin by $\frac{2\pi}{n}$ radians and let $s : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ denote reflection about the x -axis respectively. Set*

$$X_{2n} = \langle r, s \mid r^n = 1, s^2 = 1, srs^{-1} = r^{-1} \rangle.$$

Then $D_n = X_{2n}$, that is,

$$D_n = \langle r, s \mid r^n = 1, s^2 = 1, srs^{-1} = r^{-1} \rangle.$$

Proof. Theorem 1.63 shows that $\{r, s\}$ is a set of generators for D_n . Moreover, we also know that the relations listed above $r^n = 1, s^2 = 1, srs^{-1} = r^{-1}$ hold; the first two are easy to check, and the last one is Lemma 1.61. The only concern we need to deal with is that we may not have discovered all the relations of D_n ; or rather, we need to check that we have found enough relations so that any other valid relation follows as a consequence of the ones listed.

Let

$$X_{2n} = \langle r, s \mid r^n = 1, s^2 = 1, srs^{-1} = r^{-1} \rangle.$$

Assume that D_n has more relations than X_{2n} does. Then D_n would be a group of cardinality strictly smaller than X_{2n} , meaning that $|D_n| < |X_{2n}|$.¹ We will show below that in fact $|X_{2n}| \leq 2n = |D_n|$, thus obtaining a contradiction.

Now we show that X_{2n} has at most $2n$ elements using just the information contained in the presentation. By definition, since r and s generated X_{2n} then every element $x \in X_{2n}$ can be written as

$$x = r^{m_1} s^{n_1} r^{m_2} s^{n_2} \cdots r^{m_j} s^{n_j}$$

for some j and (possibly negative) integers $m_1, \dots, m_j, n_1, \dots, n_j$.² As a consequence of the last relation, we have

$$sr = r^{-1}s,$$

and its not hard to see that this implies

$$sr^m = r^{-m}s$$

for all m . Thus, we can slide an s past a power of r , at the cost of changing the sign of the power. Doing this repeatedly gives that we can rewrite x as

$$x = r^M s^N.$$

By the first relation, $r^n = 1$, from which it follows that $r^a = r^b$ if a and b are congruent modulo n . Thus we may assume $0 \leq M \leq n - 1$. Likewise, we may assume $0 \leq N \leq 1$. This gives a total of at most $2n$ elements, and we conclude that X_{2n} must in fact be D_n . \square

Note that we have *not* shown that

$$X_{2n} = \langle r, s \mid r^n, s^2, srs^{-1} = r^{-1} \rangle$$

has at least $2n$ elements using just the presentation. But for this particular example, since we know the group presented is the same as D_n , we know from Theorem 1.63 that it has exactly $2n$ elements.

¹This will become more clear once we properly define presentations.

²Note that, m_1 could be 0, so that expressions beginning with a power of s are included in this list.

1.4 The quaternions

For our last big example we mention the group of quaternions, written Q_8 .

Definition 1.67. The **quaternion group** Q_8 is a group with 8 elements

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

satisfying the following relations: 1 is the identity element, and

$$\begin{aligned} i^2 &= -1, & j^2 &= -1, & k^2 &= -1, & ij &= k, & jk &= i, & ki &= j, \\ (-1)i &= -i, & (-1)j &= -j, & (-1)k &= -k, & (-1)(-1) &= 1. \end{aligned}$$

To verify that this really is a group is rather tedious, since the associative property takes forever to check. Here is a better way: in the group $\mathrm{GL}_2(\mathbb{C})$, define elements

$$I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad A = \begin{bmatrix} \sqrt{-1} & 0 \\ 0 & -\sqrt{-1} \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \quad C = \begin{bmatrix} 0 & \sqrt{-1} \\ \sqrt{-1} & 0 \end{bmatrix}$$

where $\sqrt{-1}$ denotes the complex number whose square is -1 , to avoid confusion with the symbol $i \in Q_8$. Let $-I, -A, -B, -C$ be the negatives of these matrices.

Then we can define an injective map $f : Q_8 \rightarrow \mathrm{GL}_2(\mathbb{C})$ by assigning

$$\begin{aligned} 1 &\mapsto I, & -1 &\mapsto -I \\ i &\mapsto A, & -i &\mapsto -A \\ j &\mapsto B, & -j &\mapsto -B \\ k &\mapsto C, & -k &\mapsto -C. \end{aligned}$$

It can be checked directly that this map has the nice property (called being a *group homomorphism*) that

$$f(xy) = f(x)f(y) \text{ for any elements } x, y \in Q_8.$$

Let us now prove associativity for Q_8 using this information:

Claim: For any $x, y, z \in Q_8$, we have $(xy)z = x(yz)$.

Proof. By using the property $f(xy) = f(x)f(y)$ as well as associativity of multiplication in $\mathrm{GL}_2(\mathbb{C})$ (marked by *) we obtain

$$f((xy)z) = f(xy)f(z) = (f(x)f(y))f(z) \stackrel{*}{=} f(x)(f(y)f(z)) = f(x)f(yz) = f(x(yz)).$$

Since f is injective and $f((xy)z) = f(x(yz))$, we deduce $(xy)z = x(yz)$. \square

The subset $\{\pm I, \pm A, \pm B, \pm C\}$ of $\mathrm{GL}_2(\mathbb{C})$ is a *subgroup* (a term we define carefully later), meaning that it is closed under multiplication and taking inverses. (For example, $AB = C$ and $C^{-1} = -C$.) This proves it really is a group and one can check it satisfies an analogous list of identities as the one satisfied by Q_8 .

This is an excellent motivation to talk about group homomorphisms.

1.5 Group homomorphisms

A group homomorphism is a function between groups that preserves the group structure.

Definition 1.68. Let (G, \cdot_G) and (H, \cdot_H) be groups. A (group) **homomorphism** from G to H is a function $f : G \rightarrow H$ such that

$$f(x \cdot_G y) = f(x) \cdot_H f(y).$$

Note that a group homomorphism does not necessarily need to be injective nor surjective, it can be any function as long as it preserves the product.

Definition 1.69. Let G and H be groups. A homomorphism $f : G \rightarrow H$ is an **isomorphism** if there exists a homomorphism $g : H \rightarrow G$ such that

$$f \circ g = \text{id}_H \text{ and } g \circ f = \text{id}_G.$$

If $f : G \rightarrow H$ is an isomorphism, G and H are called **isomorphic**, and we denote this by writing $G \cong H$. An isomorphism $G \rightarrow G$ is called an **automorphism** of G . We denote the set of all automorphisms of G by $\text{Aut}(G)$.

Remark 1.70. Two groups G and H are isomorphic if we can obtain H from G by renaming all the elements, without changing the group structure. One should think of an isomorphism $f : G \xrightarrow{\cong} H$ of groups as saying that the multiplication tables of G and H are the same up to renaming the elements. The multiplication rule \cdot_G for G can be visualized as a table with both rows and columns labeled by elements of G , and with $x \cdot_G y$ placed in row x and column y . The isomorphism f sends x to $f(x)$, y to $f(y)$, and the table entry $x \cdot_G y$ to the table entry $f(x) \cdot_H f(y)$. The inverse map f^{-1} does the opposite.

Remark 1.71. Suppose that $f : G \rightarrow H$ is an isomorphism. As a function, f has an inverse, and thus it must necessarily be a bijective function. Our definition, however, requires more: the inverse must in fact also be a group homomorphism. Note that many books define group homomorphism by simply requiring it to be a homomorphism that is bijective: and we will soon show that this is in fact equivalent to the definition we gave. There are however good reasons to define it as we did: in many contexts, such as sets, groups, rings, fields, or topological spaces, the correct meaning of the word “isomorphism” in “a morphism that has a two-sided inverse”. This explains our choice of definition.

Exercise 11. Let G be a group. Show that $\text{Aut}(G)$ is a group under composition.

Example 1.72.

- (a) For any group G , the identity map $\text{id}_G : G \rightarrow G$ is a group isomorphism.
- (b) For all groups G and H , the constant map $f : G \rightarrow H$ with $f(g) = e_H$ for all $g \in G$ is a homomorphism, which we sometimes refer to as the **trivial homomorphism**.

(c) The exponential map and the logarithm map

$$\begin{array}{ccc} \exp: (\mathbb{R}, +) & \longrightarrow & (\mathbb{R} \setminus \{0\}, \cdot) \\ & & x \longmapsto e^x \end{array} \quad \begin{array}{ccc} \ln: (\mathbb{R}_{>0}, \cdot) & \longrightarrow & (\mathbb{R}, +) \\ & & y \longmapsto \ln y \end{array}$$

are both isomorphisms, so $(\mathbb{R}, +) \cong (\mathbb{R}_{>0}, \cdot)$. In fact, these maps are inverse to each other.

(d) The function $f: \mathbb{Z} \rightarrow \mathbb{Z}$ given by $f(x) = 2x$ is a group homomorphism that is injective but not surjective.

(e) For any positive integer n and any field F , the determinant map

$$\begin{array}{ccc} \det: \mathrm{GL}_n(F) & \longrightarrow & (F \setminus \{0\}, \cdot) \\ & & A \longmapsto \det(A) \end{array}$$

is a group homomorphism. For $n \geq 2$, the determinant map is not injective (you should check this!) and so it cannot be an isomorphism. It is however surjective: for each $c \in F \setminus \{0\}$, the diagonal matrix

$$\begin{pmatrix} c & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{pmatrix}$$

has determinant c .

(f) Fix an integer $n > 1$, and consider the function $f: (\mathbb{Z}, +) \rightarrow (\mathbb{C}^*, \cdot)$ given by $f(n) = e^{\frac{2\pi i}{n}}$. This is a group homomorphism, but it is neither surjective nor injective. It is not surjective because the image only contains complex numbers x with $|x| = 1$, and it is not injective because $f(0) = f(n)$.

Group homomorphisms preserve the group structure. In particular, group homomorphisms preserve the identity and all inverses.

Lemma 1.73 (Properties of homomorphisms). *If $f: G \rightarrow H$ is a homomorphism of groups, then*

$$f(e_G) = e_H.$$

Moreover, for any $x \in G$ we have

$$f(x^{-1}) = f(x)^{-1}.$$

Proof. By definition,

$$f(e_G)f(e_G) = f(e_Ge_G) = f(e_G).$$

Multiplying both sides by $f(e_G)^{-1}$, we get

$$f(e_G) = e_H.$$

Now given any $x \in G$, we have

$$f(x^{-1})f(x) = f(x^{-1}x) = f(e) = e,$$

and thus $f(x^{-1}) = f(x)^{-1}$. □

Remark 1.74. Let G be a cyclic group generated by the element g . Then any homomorphism $f: G \rightarrow H$ is completely determined by $f(g)$, since any other element $h \in G$ can be written as $h = g^n$ for some integer n , and

$$f(g^n) = f(g)^n.$$

More generally, given a group G a set S of generators for G , any homomorphism $f: G \rightarrow H$ is completely determined by the images of the generators in S : the element $g = s_1 \cdots s_m$, where s_i is either in S or the inverse of an element of S , has image

$$f(g) = f(s_1 \cdots s_m) = f(s_1) \cdots f(s_m).$$

Note, however, that not all choices of images for the generators might actually give rise to a homomorphism; we need to check that the map determined by the given images of the generators is well-defined.

Definition 1.75. The **image** of a group homomorphism $f: G \rightarrow G$ is

$$\text{im}(f) := \{f(g) \mid g \in G\}.$$

Notice that $f: G \rightarrow H$ is surjective if and only if $\text{im}(f) = H$.

Definition 1.76. The **kernel** of a group homomorphism $f: G \rightarrow G$ is

$$\ker(f) := \{g \in G \mid f(g) = e_H\}.$$

Remark 1.77. Given any group homomorphism $f: G \rightarrow G$, we must have $e_G \in \ker f$ by Lemma 1.73.

When the kernel of f is as small as possible, meaning $\ker(f) = \{e\}$, we say that f the kernel of f is trivial. A homomorphism is injective if and only if it has a trivial kernel.

Lemma 1.78. A group homomorphism $f: G \rightarrow H$ is injective if and only if $\ker(f) = \{e_G\}$.

Proof. First, note that $e_G \in \ker f$ by Lemma 1.73. If f is injective, then e_G must be the only element that f sends to e_H , and thus $\ker(f) = \{e_G\}$.

Now suppose $\ker(f) = \{e_G\}$. If $f(g) = f(h)$ for some $g, h \in G$, then

$$f(h^{-1}g) = f(h^{-1})f(g) = f(h)^{-1}f(g) = e_H.$$

But then $h^{-1}g \in \ker(f)$, so we conclude that $h^{-1}g = e_G$, and thus $g = h$. \square

Example 1.79. First, number the vertices of P_n from 1 to n in any manner you like. Now define a function $f: D_n \rightarrow S_n$ as follows: given any symmetry $\alpha \in D_n$, set $f(\alpha)$ to be the permutation of $[n]$ that records how α permutes the vertices of P_n according to your labelling. So $f(\alpha) = \sigma$ where σ is the permutation that for all $1 \leq i \leq n$, if α sends the i th vertex to the j th one in the list, then $\sigma(i) = j$. This map f is a group homomorphism.

Now suppose $f(\alpha) = \text{id}_{S_n}$. Then α must fix all the vertices of P_n , and thus α must be the identity element of D_n . We have thus shown that the kernel of f is trivial. By Lemma 1.78, this proves f is injective.

We defined isomorphisms to be homomorphisms that have an inverse that is also a homomorphism. We are now ready to show that this can simplified: an isomorphism is a bijective group homomorphism.

Lemma 1.80. Suppose $f : G \rightarrow H$ is a group homomorphism. Then f is an isomorphism if and only if f is bijective.

Proof. (\Rightarrow) A function $f : X \rightarrow Y$ between two sets is bijective if and only if it has an inverse, meaning that there is a function $g : Y \rightarrow X$ such that $f \circ g = \text{id}_Y$ and $g \circ f = \text{id}_X$. Our definition of group isomorphism implies that this must hold for any isomorphism (and more!), as we noted in Remark 1.71.

(\Leftarrow) If f is bijective homomorphism, then as a function it has a *set-theoretic* two-sided inverse g , as remarked in Remark 1.71. But we need to show that this inverse g is actually a homomorphism. For any $x, y \in H$, we have

$$\begin{aligned} f(g(xy)) &= xy && \text{since } fg = \text{id}_G \\ &= f(g(x))f(g(y)) && \text{since } fg = \text{id}_G \\ &= f(g(x)g(y)) && \text{since } f \text{ is a group homomorphism.} \end{aligned}$$

Since f is injective, we must have $g(xy) = g(x)g(y)$. Thus g is a homomorphism, and f is an isomorphism. \square

Exercise 12. Let $f : G \rightarrow H$ be an isomorphism. Show that for all $x \in G$, we have $|f(x)| = |x|$.

In other words, isomorphisms preserve the order of an element. This is an example of an isomorphism invariant.

Definition 1.81. An **isomorphism invariant** (of a group) is a property P (of groups) such that whenever G and H are isomorphic groups and G has the property P , then H also has the property P .

Theorem 1.82. The following are isomorphism invariants:

- (a) the order of the group,
- (b) the set of all the orders of elements in the group,
- (c) the property of being abelian,
- (d) the order of the center of the group,
- (e) being finitely generated.

Recall that by definition two sets have the same cardinality if and only if they are in bijection with each other.

Proof. Let $f : G \rightarrow H$ be any a group isomorphism.

- (a) Since f is a bijection by Remark 1.71, we conclude that $|G| = |H|$.

(b) We wish to show that $\{|x| \mid x \in G\} = \{|y| \mid y \in H\}$.

(\subseteq) follows from Exercise 12: given any $x \in G$, we have $|x| = |f(x)|$, which is the order of an element in H .

(\supseteq) follows from the previous statement applied to the group isomorphism f^{-1} : given any $y \in H$, we have $f^{-1}(y) \in G$ and $|y| = |f^{-1}(y)|$ is the order of an element of G .

(c) For any $y_1, y_2 \in H$ there exist some $x_1, x_2 \in G$ such that $f(x_i) = y_i$. Then we have

$$y_1 y_2 = f(x_1) f(x_2) = f(x_1 x_2) \stackrel{*}{=} f(x_2 x_1) = f(x_2) f(x_1) = y_2 y_1,$$

where $*$ indicates the place where we used that G is abelian.

(d) Exercise. The idea is to show f induces an isomorphism $Z(G) \cong Z(H)$.

(e) Exercise. Show that if S generates G then $f(S) = \{f(s) \mid s \in S\}$ generates H . \square

The easiest way to show that two groups are not isomorphic is to find an isomorphism invariant that they do not share.

Remark 1.83. Let G and H be two groups. If P is an isomorphism invariant, and G has P while H does not have P , then G is not isomorphic to H .

Example 1.84.

- (1) We have $S_n \cong S_m$ if and only if $n = m$, since $|S_n| = n!$ and $|S_m| = m!$ and the order of a group is an isomorphism invariant.
- (2) Since $\mathbb{Z}/6$ is abelian and S_3 is not abelian, we conclude that $\mathbb{Z}/6 \not\cong S_3$.
- (3) You will show in Problem Set 2 that $|Z(D_{24})| = 2$, while S_n has trivial center. We conclude that $D_{24} \not\cong S_4$.

Chapter 2

Group actions: a first look

We come to one of the central concepts in group theory: the action of a group on a set. Some would say this is the main reason one would study groups, so we want to introduce it early both as motivation for studying group theory but also because the language of group actions will be very helpful to us.

2.1 What is a group action?

Definition 2.1. For a group (G, \cdot) and set S , an **action** of G on S is a function

$$G \times S \rightarrow S,$$

typically written as $(g, s) \mapsto g \cdot s$, such that

- (1) $g \cdot (h \cdot s) = (gh) \cdot s$ for all $g, h \in G$ and $s \in S$.
- (2) $e_G \cdot s = s$ for all $s \in S$.

Remark 2.2. To make the first axiom clearer, we will write \cdot for the action of G on S and no symbol (concatenation) for the multiplication of two elements in the group G .

A group action is the same thing as a group homomorphism.

Lemma 2.3 (Permutation representation). *Consider a group G and a set S .*

- (1) *Suppose \cdot is an action of G on S . For each $g \in G$, let $\mu_g: S \rightarrow S$ denote the function given by $\mu_g(s) = g \cdot s$. Then the function*

$$\begin{aligned} \rho: G &\longrightarrow \text{Perm}(S) \\ g &\longmapsto \mu_g \end{aligned}$$

is a well-defined homomorphism of groups.

- (2) *Conversely, if $\rho: G \rightarrow \text{Perm}(S)$ is a group homomorphism, then the rule*

$$g \cdot s := (\rho(g))(s)$$

defines an action of G on S .

Proof. (1) Assume we are given an action of G on S . We first need to check that for all g , μ_g really is a permutation of S . We will show this by proving that μ_g has a two-sided inverse; in fact, that inverse is $\mu_{g^{-1}}$. Indeed, we have

$$\begin{aligned}
 (\mu_g \circ \mu_{g^{-1}})(s) &= \mu_g(\mu_{g^{-1}}(s)) && \text{by the definition of composition} \\
 &= g \cdot (g^{-1} \cdot s) && \text{by the definitinion for } \mu_g \text{ and } \mu_{g^{-1}} \\
 &= (gg^{-1}) \cdot s && \text{by the definition of a group action} \\
 &= e_G \cdot s && \text{by the definition of a group} \\
 &= s && \text{by the definition of a group action}
 \end{aligned}$$

thus $\mu_g \circ \mu_{g^{-1}} = \text{id}_S$, and a similar argument shows that $\mu_{g^{-1}} \circ \mu_g = \text{id}_S$ (exercise!). This shows that μ_g has an inverse, and thus it is bijective; it must then be a permutation of S .

Finally, we wish to show that ρ is a homomorphism of groups, so we need to check that $\rho(gh) = \rho(g) \circ \rho(h)$. Equivalently, we need to prove that $\mu_{gh} = \mu_g \circ \mu_h$. Now for all s , we have

$$\begin{aligned}
 \mu_{gh}(s) &= (gh) \cdot s && \text{by definition of } \mu \\
 &= g \cdot (h \cdot s) && \text{by definition of a group action} \\
 &= \mu_g(\mu_h(s)) && \text{by definition of } \mu_g \text{ and } \mu_h \\
 &= (\mu_g \circ \mu_h)(s).
 \end{aligned}$$

This proves that ρ is a homomorphism.

(2) On the other hand, given a homomorphism ρ , the function

$$\begin{aligned}
 G \times S &\longrightarrow S \\
 (g, s) &\longmapsto g \cdot s = \rho(g)(s)
 \end{aligned}$$

is an action, because

$$\begin{aligned}
 h \cdot (g \cdot s) &= \rho(h)(\rho(g)(s)) && \text{by definition of } \rho \\
 &= (\rho(h) \circ \rho(g))(s) \\
 &= \rho(gh)(s) && \text{since } \rho \text{ is a homomorphism} \\
 &= (gh) \cdot s && \text{by definition of } \rho,
 \end{aligned}$$

and

$$e_G s = \rho(e_G)(s) = \text{id}(s) = s. \quad \square$$

Definition 2.4. Given a group G acting on a set S , the group homomorphism ρ associated to the action as defined in Lemma 2.3 is called the **permutation representation** of the action.

Definition 2.5. Let G be a group acting on a set S . The equivalence relation on S induced by the action of G , written \sim_G , is defined by $s \sim_G t$ if and only if there is a $g \in G$ such that $t = g \cdot s$. The equivalence classes of \sim_G are called **orbits**: the equivalence class

$$\text{Orb}_G(s) := \{g \cdot s \mid g \in G\}$$

is the orbit of s . The set of equivalence classes with respect to \sim_G is written S/G .

Lemma 2.6. Let G be a group acting on a set S . Then

- (a) The relation \sim_G really is an equivalence relation.
- (b) For any $s, t \in S$ either $\text{Orb}_G(s) = \text{Orb}_G(t)$ or $\text{Orb}_G(s) \cap \text{Orb}_G(t) = \emptyset$.
- (c) The orbits of the action of G form a partition of S : $S = \bigcup_{s \in S} \text{Orb}_G(s)$.

Proof. Assume G acts on S .

(a) We really need to prove three things: that \sim_G is reflexive, symmetric, and transitive.

(Reflexive): We have $x \sim_G x$ for all $x \in S$ since $x = e_G \cdot x$.

(Symmetric): If $x \sim_G y$, then $y = g \cdot x$ for some $g \in G$, and thus

$$g^{-1} \cdot y = g^{-1} \cdot (g \cdot x) = (g^{-1}g) \cdot x = e \cdot x = x,$$

which shows that $y \sim_G x$.

(Transitive): If $x \sim_G y$ and $y \sim_G z$, then $y = g \cdot x$ and $z = h \cdot y$ for some $g, h \in G$ and hence $z = h \cdot (g \cdot x) = (hg) \cdot x$, which gives $x \sim_G z$.

Parts (b) and (c) are formal properties of the equivalence classes for any equivalence relation. \square

Corollary 2.7. Suppose a group G acts on a finite set S . Let s_1, \dots, s_k be a complete set of orbit representatives — that is, assume each orbit contains exactly one member of the list s_1, \dots, s_k . Then

$$|S| = \sum_{i=1}^k |\text{Orb}_G(s_i)|.$$

Proof. This is an immediate corollary of the fact that the orbits form a partition of S . \square

Remark 2.8. Let G be a group acting on S . The associated group homomorphism ρ is injective if and only if it has trivial kernel, by Lemma 1.78. This is equivalent to the statement $\mu_g = \text{id}_S \implies g = e_G$. The later can be written in terms of elements of S : for each $g \in G$,

$$g \cdot s = s \quad \text{for all } s \in S \implies g = e_G.$$

Definition 2.9. Let G be a group acting on a set S . The action is **faithful** if the associated group homomorphism is injective. Equivalently, the action is faithful if and only if

$$g \cdot s = s \quad \text{for all } s \in S \implies g = e_G.$$

The action is **transitive** if for all $p, q \in S$ there is $g \in G$ such that $q = g \cdot p$. Equivalently, the action is transitive if there is only one orbit, meaning that

$$\text{Orb}_G(p) = S \text{ for all } p \in S.$$

2.2 Examples of group actions

Example 2.10 (Trivial action). For any group G and any set S , $g \cdot s := s$ defines an action, the **trivial action**. The associated group homomorphism is the map

$$\begin{aligned} G &\longrightarrow \text{Perm}(S) \\ g &\longmapsto \text{id}_S. \end{aligned}$$

A trivial action is not faithful unless the group G is trivial; in fact, the corresponding group homomorphism is trivial.

Example 2.11. The group D_n acts on the vertices of P_n , which we will label with V_1, \dots, V_n in a counterclockwise fashion, with V_1 on the positive x -axis, as in Notation 1.56. Note that D_n acts on $\{V_1, \dots, V_n\}$: for each $g \in D_n$ and each integer $1 \leq j \leq n$, we set

$$g \cdot V_j = V_i \quad \text{if and only if} \quad g(V_j) = V_i.$$

This satisfies the two axioms of a group action (check!).

Let $\rho: D_n \rightarrow \text{Perm}(\{V_1, \dots, V_n\}) \cong S_n$ be the associated group homomorphism. Note that ρ is injective, because if an element of D_n fixes all n vertices of a polygon, then it must be the identity map. More generally, if an isometry of \mathbb{R}^2 fixes any three noncolinear points, then it is the identity. To see this, note that given three noncolinear points, every point in the plane is uniquely determined by its distance from these three points (exercise!).

The action of D_n on the n vertices of P_n is faithful; in fact, we saw before that each $\sigma \in D_n$ is completely determined by what it does to any two adjacent vertices.

Example 2.12 (group acting on itself by left multiplication). Let G be any group and define an action \cdot of G on G (regarded as just a set) by the rule

$$g \cdot x := gx.$$

This is an action, since multiplication is associative and $e_G \cdot x = x$ for all x ; it is known as the **left regular action** of G on itself.

The left regular action of G on itself is faithful, since if $g \cdot x = x$ for all x (or even for just one x), then $g = e$. It follows that the associated homomorphism is injective. This action is also transitive: given any $g \in G$, $g = g \cdot e$, and thus $\text{Orb}_G(e) = G$.

Example 2.13 (conjugation). Let G be any group and fix an element $g \in G$. Define the conjugation action of G on itself by setting

$$g \cdot x := gxg^{-1} \text{ for any } g, x \in G.$$

The action of G on itself by conjugation is not necessarily faithful. In fact, we claim that the kernel of the permutation representation $\rho: G \rightarrow \text{Perm}(G)$ for the conjugation action is the center $Z(G)$. Indeed,

$$\begin{aligned} g \in \ker \rho &\iff g \cdot x = x \text{ for all } x \in G \iff gxg^{-1} = x \text{ for all } x \in G \\ &\iff gx = xg \text{ for all } x \in G \iff g \in Z(G). \end{aligned}$$

If G is nontrivial, this action is *never* transitive unless G is trivial: note that $\text{Orb}_G(e) = \{e\}$.

Chapter 3

Subgroups

Every time we define a new abstract structure consisting of a set S with some extra structure, we then want to consider subsets of S that inherit that special structure. It is now time to discuss subgroups.

3.1 Definition and examples

Definition 3.1. A nonempty subset H of a group G is a **subgroup** of G if H is a group under the multiplication law of G . If H is a subgroup of G , we write $H \leq G$, or $H < G$ if we want to indicate that H is a subgroup of G but $H \neq G$.

Remark 3.2. Note that if H is a subgroup of G , then necessarily H must be closed for the product in G , meaning that for any $x, y \in H$ we must have $xy \in H$.

Remark 3.3. Let H be a subgroup of G . Since H itself is a group, it has an identity element e_H , and thus

$$e_H e_H = e_H$$

in H . But the product in H is just a restriction of the product of G , so this equality also holds in G . Multiplying by e_H^{-1} , we conclude that $e_H = e_G$.

In summary, if H is any subgroup of G , then we must have $e_G \in H$.

Example 3.4. Any group G has two **trivial subgroups**: G itself, and $\{e_G\}$.

Any subgroup H of G that is neither G nor $\{e_G\}$ is a **nontrivial subgroup**. A group might not have any nontrivial subroups.

Example 3.5. The group $\mathbb{Z}/2$ has no nontrivial subgroup.

Example 3.6. The following are strings of subgroups with the obvious group structure:

$$\mathbb{Z} < \mathbb{Q} < \mathbb{R} < \mathbb{C} \quad \text{and} \quad \mathbb{Z}^\times < \mathbb{Q}^\times < \mathbb{R}^\times < \mathbb{C}^\times.$$

To prove that a certain subset H of G forms a subgroup, it is very inefficient to prove directly that H forms a group under the same operation as G . Instead, we use one of the following two tests:

Lemma 3.7 (Subgroup tests). *Let H be a subset of a group G .*

- Two-step test: If H is nonempty and closed under multiplication and taking inverses, then H is a subgroup of G . More precisely, if for all $x, y \in H$, we have $xy \in H$ and $x^{-1} \in H$, then H is a subgroup of G .
- One-step test: If H is nonempty and $xy^{-1} \in H$ for all $x, y \in H$, then H is a subgroup of G .

Proof. We prove the One-step test first. Assume H is nonempty and for all $x, y \in H$ we have $xy^{-1} \in H$. Since H is nonempty, there is some $h \in H$, and hence $e_G = hh^{-1} \in H$. Since $e_{Gx} = x = xe_G$ for any $x \in G$, and hence for any $x \in H$, then e_G is an identity element for H . For any $h \in H$, we have $h^{-1} = eh^{-1} \in H$, and since in G we have $h^{-1}h = e = hh^{-1} \in H$ and this calculation does not change when we restrict to H , we can conclude that every element of H has an inverse inside H . For every $x, y \in H$ we must have $y^{-1} \in H$ and thus

$$xy = x(y^{-1})^{-1} \in H$$

so H is closed under the multiplication operation. This means that the restriction of the group operation of G to H is a well-defined group operation. This operation is associative by the axioms for the group G . The axioms of a group have now been established for (H, \cdot) .

Now we prove the Two-Step test. Assume H is nonempty and closed under multiplication and taking inverses. Then for all $x, y \in H$ we must have $y^{-1} \in H$ and thus $xy^{-1} \in H$. Since the hypothesis of the One-step test is satisfied, we conclude that H is a subgroup of G . \square

Lemma 3.8 (Examples of subgroups). *Let G be a group.*

- If H is a subgroup of G and K is a subgroup of H , then K is a subgroup of G .
- Let J be any (index) set. If H_α is a subgroup of G for all $\alpha \in J$, then $H = \bigcap_{\alpha \in J} H_\alpha$ is a subgroup of G .
- If $f : G \rightarrow H$ is a homomorphism of groups, then $\text{im}(f)$ is a subgroup of H .
- If $f : G \rightarrow H$ is a homomorphism of groups, and K is a subgroup of G , then

$$f(K) := \{f(g) \mid g \in K\}$$

is a subgroup of H .

- If $f : G \rightarrow H$ is a homomorphism of groups, then $\ker(f)$ is a subgroup of G .
- The center $Z(G)$ is a subgroup of G .

Proof.

- By definition, K is a group under the multiplication in H , and the multiplication in H is the same as that in G , so K is a subgroup of G .
- First, note that H is nonempty since $e_G \in H_\alpha$ for all $\alpha \in J$. Moreover, given $x, y \in H$, for each α we have $x, y \in H_\alpha$ and hence $xy^{-1} \in H_\alpha$. It follows that $xy^{-1} \in H$. By the Two-Step test, H is a subgroup of G .

- (c) Since G is nonempty, then $\text{im}(f)$ must also be nonempty; for example, it contains $f(e_G) = e_H$. If $x, y \in \text{im}(f)$, then $x = f(a)$ and $y = f(b)$ for some $a, b \in G$, and hence

$$xy^{-1} = f(a)f(b)^{-1} = f(ab^{-1}) \in \text{im}(f).$$

By the Two-Step Test, $\text{im}(f)$ is a subgroup of H .

- (d) The restriction $g: K \rightarrow H$ of f to K is still a group homomorphism, and thus $f(K) = \text{img}$ is a subgroup of H .
(e) Using the One-step test, note that if $x, y \in \ker(f)$, meaning $f(x) = f(y) = e_G$, then

$$f(xy^{-1}) = f(x)f(y)^{-1} = e_G.$$

This shows that if $x, y \in \ker(f)$ then $xy^{-1} \in \ker(f)$, so $\ker(f)$ is closed for taking inverses. By the Two-Step test, $\ker(f)$ is a subgroup of G .

- (f) The center $Z(G)$ is the kernel of the permutation representation $G \rightarrow \text{Perm}(G)$ for the conjugation action, so $Z(G)$ is a subgroup of G since the kernel of a homomorphism is a subgroup. \square

Example 3.9. For any field F , the **special linear group**

$$\text{SL}_n(F) := \{A \mid A = n \times n \text{ matrix with entries in } F, \det(A) = 1_F\}$$

is a subgroup of the general linear group $\text{GL}_n(F)$. To prove this, note that $\text{SL}_n(F)$ is the kernel of the determinant map $\det: \text{GL}_n(F) \rightarrow F^\times$, which is one of the homomorphisms in Example 1.72. By Lemma 3.8, this implies that $\text{SL}_n(F)$ is indeed a subgroup of $\text{GL}_n(F)$.

Definition 3.10. Let $f: G \rightarrow H$ be a group homomorphism and $K \leq H$. The **preimage** of K is given by

$$f^{-1}(K) := \{g \in G \mid f(g) \in K\}$$

Exercise 13. Prove that if $f: G \rightarrow H$ is a group homomorphism and $K \leq H$, then the preimage of K is a subgroup of G .

Exercise 14. The set of rotational symmetries $\{r^i \mid i \in \mathbb{Z}\} = \{\text{id}, r, r^2, \dots, r^{n-1}\}$ of P_n is a subgroup of D_n .

In fact, this is the subgroup generated by r .

Definition 3.11. Given a group G and a subset X of G , the **subgroup of G generated by X** is

$$\langle X \rangle := \bigcap_{\substack{H \leq G \\ H \supseteq X}} H.$$

If $X = \{x\}$ is a set with one element, then we write $\langle X \rangle = \langle x \rangle$ and we refer to this as the **cyclic subgroup generated by x** . More generally, when $X = \{x_1, \dots, x_n\}$ is finite, we may write $\langle x_1, \dots, x_n \rangle$ instead of $\langle X \rangle$. Finally, given two subsets X and Y of G , we may sometimes write $\langle X, Y \rangle$ instead of $\langle X \cup Y \rangle$.

Remark 3.12. Note that by Lemma 3.8, $\langle X \rangle$ really is a subgroup of G . By definition, the subgroup generated by X is the smallest (with respect to containment) subgroup of G that contains X , meaning that $\langle X \rangle$ is contained in any subgroup that contains X .

Remark 3.13. Do not confuse this notation with giving generators and relations for a group; here we are forgoing the relations and focusing only on writing a list of generators. Another key difference is that we have picked elements in a given group G , but the subgroup they generate might not be G itself, but rather some other subgroup of G .

Lemma 3.14. *For a subset X of G , the elements of $\langle X \rangle$ can be described as:*

$$\langle X \rangle = \{x_1^{j_1} \cdots x_m^{j_m} \mid m \geq 0, j_1, \dots, j_m \in \mathbb{Z} \text{ and } x_1, \dots, x_m \in X\}.$$

Note that the product of no elements is by definition the identity.

Proof. Let

$$S = \{x_1^{j_1} \cdots x_m^{j_m} \mid m \geq 0, j_1, \dots, j_m \in \mathbb{Z} \text{ and } x_1, \dots, x_m \in X\}.$$

Since $\langle X \rangle$ is a subgroup that contains X , it is closed under products and inverses, and thus must contain all elements of S . Thus $X \supseteq S$.

To show $X \subseteq S$, we will prove that the set S is a subgroup of G using the One-step test:

- $S \neq \emptyset$ since we allow $m = 0$ and declare the empty product to be e_G .
- Let a and b be elements of S , so that they can be written as $a = x_1^{j_1} \cdots x_m^{j_m}$ and $b = y_1^{i_1} \cdots y_n^{i_n}$. Then

$$ab^{-1} = x_1^{j_1} \cdots x_m^{j_m} (y_1^{i_1} \cdots y_n^{i_n})^{-1} = x_1^{j_1} \cdots x_m^{j_m} y_n^{-i_n} \cdots y_1^{-i_1} \in S.$$

Therefore, $S \leq G$ and $X \subseteq S$ (by taking $m = 1$ and $j_1 = 1$) and by the minimality of $\langle X \rangle$ we conclude that $\langle X \rangle \subseteq S$. \square

Example 3.15. Lemma 3.14 implies that for an element x of a group G , $\langle x \rangle = \{x^j \mid j \in \mathbb{Z}\}$.

Example 3.16. We showed in Theorem 1.63 that $D_n = \langle r, s \rangle$, so D_n is the subgroup of D_n generated by $\{r, s\}$. But do not mistake this for a presentation with no relations! In fact, these generators satisfy lots of relations, such as $srs = r^{-1}$, which we proved in Lemma 1.61.

Example 3.17. For any $n \geq 1$, we proved in Problem Set 2 that S_n is generated by the collection of adjacent transpositions $(i \ i+1)$.

Theorem 3.18 (Cayley's Theorem). *Every finite group is isomorphic to a subgroup of S_n .*

Proof. Suppose G is a finite group of order n and label the group elements of G from 1 to n in any way you like. The left regular action of G on itself determines a permutation representation $\rho: G \rightarrow \text{Perm}(G)$, which is injective. Note that since G has n elements, $\text{Perm}(G)$ is the group of permutations on n elements, and thus $\text{Perm}(G) \cong S_n$. By Lemma 3.8, $\text{im}(\rho)$ is a subgroup of S_n . If we restrict ρ to its image, we get an isomorphism $\rho: G \rightarrow \text{im}(\rho)$. Hence $G \cong \text{im}(\rho)$, which is a subgroup of S_n . \square

Remark 3.19. From a practical perspective, this is a nearly useless theorem. It is, however, a beautiful fact.

3.2 Subgroups vs isomorphism invariants

Some properties of a group G pass onto all its subgroups, but not all. In this section, we collect some facts examples illustrating some of the most important properties.

Theorem 3.20 (Lagrange's Theorem). *If H is a subgroup of a finite group G , then $|H|$ divides $|G|$.*

You will prove Lagrange's Theorem in the next problem set.

Exercise 15. Let G be a finite group Suppose that A and B are subgroups of G such that $\gcd(|A|, |B|) = 1$. Show that $A \cap B = \{e\}$.

Example 3.21 (Infinite group with finite subgroup). The group $\mathrm{SL}_2(\mathbb{R})$ is infinite, but the matrix

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

has order 2 and it generates the subgroup $\langle A \rangle = \{A, I\}$ with two elements.

Example 3.22 (Nonabelian group with abelian subgroup). The dihedral group D_n , with $n \geq 3$, is nonabelian, while the subgroup of rotations (see Exercise 14) is abelian (for example, because it is cyclic; see Lemma 3.27 below).

To give an example of a finitely generated group with an infinitely generated group, we have to work a bit harder.

Example 3.23 (Finitely generated group with infinitely generated subgroup). Consider the subgroup G of $\mathrm{GL}_2(\mathbb{Q})$ generated by

$$A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}.$$

Let H be the subgroup of $\mathrm{GL}_2(\mathbb{Q})$ given by

$$H = \left\{ \begin{pmatrix} 1 & \frac{n}{2^m} \\ 0 & 1 \end{pmatrix} \in G \mid n, m \in \mathbb{Z} \right\}.$$

We leave it as an exercise to check that this is indeed a subgroup of $\mathrm{GL}_2(\mathbb{Q})$. Note that for all integers n and m we have

$$A^n = \begin{pmatrix} 1 & n \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad B^m = \begin{pmatrix} 2^m & 0 \\ 0 & 1 \end{pmatrix},$$

and

$$B^{-m} A^n B^m = \begin{pmatrix} 1 & \frac{n}{2^m} \\ 0 & 1 \end{pmatrix} \in H.$$

Therefore, H is a subgroup of G , and in fact

$$H = \langle B^{-m} A^n B^m \mid n, m \in \mathbb{Z} \rangle.$$

While $G = \langle A, B \rangle$ is finitely generated by construction, we claim that H is not. The issue is that

$$\begin{pmatrix} 1 & \frac{a}{2^b} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & \frac{c}{2^d} \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & \frac{a}{2^b} + \frac{c}{2^d} \\ 0 & 1 \end{pmatrix},$$

so the subgroup generated by any finite set of matrices in H , say

$$\left\langle \begin{pmatrix} 1 & \frac{n_1}{2^{m_1}} \\ 0 & 1 \end{pmatrix}, \dots, \begin{pmatrix} 1 & \frac{n_t}{2^{m_t}} \\ 0 & 1 \end{pmatrix} \right\rangle$$

does not contain

$$\begin{pmatrix} 1 & \frac{1}{2^N} \\ 0 & 1 \end{pmatrix} \in H$$

with $N = \max_i\{|m_i|\} + 1$. Thus H is infinitely generated.

In the previous example, we constructed a group with two generators that has an infinitely generated subgroup. We will see in the next section that we couldn't have done this with less generators; in fact, the subgroups of a cyclic group are all cyclic.

Below we collect some important facts about the relationship between finite groups and their subgroups, including some explained by the examples above and others which we leave as an exercise.

Order of the group:

- Every subgroup of a finite group is finite.
- There exist infinite groups with finite subgroups; see Example 3.21.
- Lagrange's Theorem: If H is a subgroup of a finite group G , then $|H|$ divides $|G|$.

Orders of elements:

- If $H \subseteq G$, then the set of orders of elements of H is a subset of the set of orders of elements of G .

Abelianity:

- Every subgroup of an abelian group is abelian.
- There exist nonabelian groups with abelian subgroups; see Example 3.22.
- Every cyclic (sub)group is abelian.

Generators:

- There exist a finitely generated group G and a subgroup H of G such that H is not finitely generated; see Example 3.23.
- Every infinitely generated group has finitely generated subgroups.¹
- Every subgroup of a cyclic group is cyclic; see Theorem 3.29.

¹This one is a triviality: we are just noting that even if the group is infinitely generated, we can always consider the subgroup generated by our favorite element, which is, by definition, finitely generated.

3.3 Cyclic groups

Recall the definition of a cyclic group.

Definition 3.24. If G is a group generated by a single element, meaning that there exists $x \in G$ such that $G = \langle x \rangle$, then G is a **cyclic group**.

Remark 3.25. Given a cyclic group G , we may be able to pick different generators for G . For example, \mathbb{Z} is a cyclic group, and both 1 or -1 are a generator. More generally, for any element x in a group G

$$\langle x \rangle = \langle x^{-1} \rangle.$$

Example 3.26. The main examples of cyclic groups, in additive notation, are the following:

- The group $(\mathbb{Z}, +)$ is cyclic with generator 1 or -1 .
- The group $(\mathbb{Z}/n, +)$ of congruences modulo n is cyclic, since it is for example generated by $[1]$. Below we will find all the choices of generators for this group.

In fact, we will later prove that up to isomorphism these are the *only* examples of cyclic groups.

Let us record some facts important facts about cyclic groups which you have proved in problem sets:

Lemma 3.27. *Every cyclic group is abelian.*

Lemma 3.28. *Let G be a group and $x \in G$. If $x^m = e$ then $|x|$ divides m .*

Now we can use these to say more about cyclic groups.

Theorem 3.29. *Let $G = \langle x \rangle$, where x has finite order n . Then*

(a) $|G| = |x| = n$ and $G = \{e, x, \dots, x^{n-1}\}$.

(b) For any integer k , then $|x^k| = \frac{n}{\gcd(k, n)}$. In particular,

$$\langle x^k \rangle = G \iff \gcd(n, k) = 1.$$

(c) There is a bijection

$$\begin{array}{ccc} \{\text{divisors of } |G|\} & \longleftrightarrow & \{\text{subgroups of } G\} \\ d \longmapsto \Psi & & \langle x^{\frac{|G|}{d}} \rangle \\ |H| \longleftarrow \Phi & & H \end{array}$$

Thus all subgroups of G are cyclic, and there is a unique subgroup of each order.

Proof. (a) By Lemma 3.14, we know $G = \{x^i \mid i \in \mathbb{Z}\}$. Now we claim that the elements

$$e = x^0, x^1, \dots, x^{n-1}$$

are all distinct. Indeed, if $x^i = x^j$ for some $0 \leq i < j < n$, then $x^{j-i} = e$ and $1 \leq j-i < n$, contradicting the minimality of the order n of x . In particular, this shows that $|G| \geq n$.

Now take any $m \in \mathbb{Z}$. By the Division Algorithm, we can write $m = qn + r$ for some integers q, r with $0 < r \leq n$. Then

$$x^m = x^{qn+r} = (x^n)^q x^r = x^r.$$

This shows that every element in G can be written in the form x^r with $0 \leq r < n$, so

$$G = \{x^0, x^1, \dots, x^{n-1}\} \quad \text{and} \quad |G| = n.$$

(b) Let k be any integer. Set $y := x^k$ and $d := \gcd(n, k)$, and note that $n = da, k = db$ for some $a, b \in \mathbb{Z}$ such that $\gcd(a, b) = 1$. We have

$$y^a = x^{ka} = x^{dba} = (x^n)^b = e,$$

so $|y|$ divides a by Lemma 3.28. On the other hand, $x^{k|y|} = y^{|y|} = e$, so again by Lemma 3.28 we have n divides $k|y|$. Now

$$da = n \text{ divides } k|y| = db|y|$$

and thus

$$a \text{ divides } b|y|.$$

But $\gcd(a, b) = 1$, so we conclude that a divides $|y|$. Since $|y|$ also divides a and both a and $|y|$ are positive, we conclude that

$$|y| = a = \frac{n}{\gcd(k, n)}.$$

(c) Consider any subgroup H of G with $H \neq \{e\}$, and set

$$k := \min\{i \in \mathbb{Z} \mid i > 0 \text{ and } g^i \in H\}.$$

On the one hand, $H \supseteq \langle g^k \rangle$, since $H \ni g^k$ and H is closed for products. Moreover, given any other positive integer i , we can again write $i = kq + r$ for some integers q, r with $0 \leq r < k$, and

$$g^r = g^{i-kq} = g^i(g^k)^q \in H,$$

so by minimality of r we conclude that $r = 0$. Therefore, $k|r$, and thus we conclude that

$$H = \langle g^k \rangle.$$

Now to show that Ψ is a bijection, we only need to prove that Φ is a well-defined function and a two-sided inverse for Ψ , and this we leave as an exercise. \square

Corollary 3.30. Let G be any finite group and consider $x \in G$. Then $|x|$ divides $|G|$.

Proof. The subgroup $\langle x \rangle$ of G generated by x is a cyclic group, and since G is finite so is $\langle x \rangle$. By Theorem 3.29, $|x| = |\langle x \rangle|$, and by Lagrange's Theorem 3.20, the order of $\langle x \rangle$ divides the order of G . \square

There is a sort of quasi-converse to Theorem 3.29:

Exercise 16. Show that if G is a finite group G has a unique subgroup of order d for each positive divisor d of $|G|$, then G must be cyclic.

We can say a little more about the bijection in Theorem 3.29. Notice how smaller subgroups (with respect to containment) correspond to smaller divisors of G . We can make this observation rigorous by talking about partially ordered sets.

Definition 3.31. An **order relation** on a set S is a binary relation \leq that satisfies the following properties:

- Reflexive: $s \leq s$ for all $s \in S$.
- Antisymmetric: if $a \leq b$ and $b \leq a$, then $a = b$.
- Transitive: if $a \leq b$ and $b \leq c$, then $a \leq c$.

A **partially ordered set** or **poset** consists of a set S endowed with an order relation \leq , which we might indicate by saying that the pair (S, \leq) is a partially ordered set.

Given a poset (S, \leq) and a subset $T \subseteq S$, an **upper bound** for T is an element $s \in S$ such that $t \leq s$ for all $t \in T$, while a **lower bound** is an element $s \in S$ such that $s \leq t$ for all $t \in T$. An upper bound s for T is called a **supremum** if $s \leq u$ for all upper bounds u of T , while a lower bound t for T is an **infimum** if $t \leq l$ for all lower bounds t for T . A **lattice** is a poset in which every two elements have a unique supremum and a unique infimum.

Remark 3.32. Note that the word *unique* can be removed from the definition of lattice. In fact, if a subset $T \subseteq S$ has a supremum, then that supremum is necessarily unique. Indeed, given two suprema s and t , then by definition $s \leq t$, since s is a supremum and t is an upper bound for T , but also $t \leq s$ since t is a supremum and s is an upper bound for T . By antisymmetry, we conclude that $s = t$.

Example 3.33. The set of all positive integers is a poset with respect to divisibility, setting $a \leq b$ whenever $a|b$. In fact, this is a lattice: the supremum of a and b is $\text{lcm}(a, b)$ and the infimum of a and b is $\text{gcd}(a, b)$.

Example 3.34. Given a set S , the **power set** of S , meaning the set of all subsets of S , is a poset with respect to containment, where the order is defined by $A \leq B$ whenever $A \subseteq B$. In fact, this is a lattice: the supremum of A and B is $A \cup B$ and the infimum of A and B is $A \cap B$.

Exercise 17. Show that the set of all subgroups of a group G is a poset with respect to containment, setting $A \leq B$ if $A \subseteq B$.

Lemma 3.35. *The set of all subgroups of a group G is a lattice with respect to containment.*

Proof. Let A and B be subgroups of G . We need to prove that A and B have an infimum and a supremum. We claim that $A \cap B$ is the infimum and $\langle A, B \rangle$ is the supremum. First, these are both subgroups of G , by Lemma 3.8 in the case $A \cap B$ and by definition for the other. Moreover, $A \cap B$ is a lower bound for A and B and $\langle A, B \rangle$ is an upper bound by definition. Finally, if $H \leq A$ and $H \leq B$, then every element of H is in both A and B , and thus it must be in $A \cap B$, so $H \leq A \cap B$. Similarly, if $A \leq H$ and $B \leq H$, then $\langle A, B \rangle \subseteq H$. \square

Remark 3.36. The isomorphism Ψ in Theorem 3.29 satisfies the following property: if $d_1 \mid d_2$ then $\Psi(d_1) \subseteq \Psi(d_2)$. In other words, Ψ preserves the poset structure. This means that Ψ is a **lattice isomorphism** between the lattice of divisors of $|G|$ and the lattice of subgroups of G . Of course the inverse map $\Phi = \Psi^{-1}$ is also a lattice isomorphism.

Lemma 3.37 (Universal Mapping Property of a Cyclic Group). *Let $G = \langle x \rangle$ be a cyclic group and let H be any other group.*

- (1) *If $|x| = n < \infty$, then for each $y \in H$ such that $y^n = e$, there exists a unique group homomorphism $f: G \rightarrow H$ such that $f(x) = y$.*
- (2) *If $|x| = \infty$, then for each $y \in H$, there exists a unique group homomorphism $f: G \rightarrow H$ such that $f(x) = y$.*

In both cases this unique group homomorphism is given by $f(x^i) = y^i$ for any $i \in \mathbb{Z}$.

Remark 3.38. We will later discuss a universal mapping property of any presentation. This is a particular case of that universal mapping property of a presentation, since a cyclic group is either presented by $\langle x \mid x^n = e \rangle$ or $\langle x \mid - \rangle$.

Proof. Recall that either $G = \{e, x, x^2, \dots, x^{n-1}\}$ has exactly n elements if $|x| = n$ or $G = \{x^i \mid i \in \mathbb{Z}\}$ with no repetitions if $|x| = \infty$.

Uniqueness: We have already noted that any homomorphism is uniquely determined by the images of the generators of the domain in Remark 1.74, and that f must then be given by $f(x^i) = f(x)^i = y^i$.

Existence: In either case, define $f(x^i) = y^i$. We must show this function is a well-defined group homomorphism. To see that f is well-defined, suppose $x^i = x^j$ for some $i, j \in \mathbb{Z}$. Then, since $x^{i-j} = e_G$, using Lemma 3.28 we have

$$\begin{cases} n \mid i - j & \text{if } |x| = n \\ i - j = 0 & \text{if } |x| = \infty \end{cases} \implies \begin{cases} y^{i-j} = y^{nk} & \text{if } |x| = n \\ y^{i-j} = y^0 & \text{if } |x| = \infty \end{cases} \implies y^{i-j} = e_H \implies y^i = y^j.$$

Thus, if $x^i = x^j$ then $f(x^i) = y^i = y^j = f(x^j)$. In particular, if $x^k = e$, then $f(x^k) = e$, and f is well-defined.

The fact that f is a homomorphism is immediate:

$$f(x^i x^j) = f(x^{i+j}) = y^{i+j} = y^i y^j = f(x^i) f(x^j).$$

\square

Definition 3.39. The **infinite cyclic group** is the group

$$C_\infty := \{a^i \mid i \in \mathbb{Z}\}$$

with multiplication $a^i a^j = a^{i+j}$.

For any natural number n , the **cyclic group of order n** is the group

$$C_n := \{a^i \mid i \in \{0, \dots, n-1\}\}$$

with multiplication $a^i a^j = a^{i+j} \pmod{n}$.

Remark 3.40. The presentations for these groups are

$$C_\infty = \langle a \mid - \rangle \quad \text{and} \quad C_n = \langle a \mid a^n = e \rangle.$$

Theorem 3.41 (Classification Theorem for Cyclic Groups). *Every infinite cyclic group is isomorphic to C_∞ . Every cyclic group of order n is isomorphic to C_n .*

Proof. Suppose $G = \langle x \rangle$ with $|x| = n$ or $|x| = \infty$, and set

$$H = \begin{cases} C_n & \text{if } |x| = n \\ C_\infty & \text{if } |x| = \infty. \end{cases}$$

By Lemma 3.37, there are homomorphisms $f: G \rightarrow H$ and $g: G \rightarrow H$ such that $f(x) = a$ and $g(a) = x$. Now $g \circ f$ is an endomorphism of G mapping x to x . But the identity map also has this property, and so the uniqueness clause in Lemma 3.37 gives us $g \circ f = \text{id}_G$. Similarly, $f \circ g = \text{id}_H$. We conclude that f and g are isomorphisms. \square

Example 3.42. For a fixed $n \geq 1$,

$$\mu_n := \{z \in \mathbb{C} \mid z^n = 1\}$$

is a subgroup of $(\mathbb{C} \setminus \{0\}, \cdot)$. Since $\|z^n\| = \|z\|^n = 1$ for any $z \in \mu_n$, then we can write $z = e^{ri}$ for some real number r . Moreover, the equality $1 = z^n = e^{nri}$ implies that nr is an integer multiple of 2π . It follows that

$$\mu_n = \{1, e^{2\pi i/n}, e^{4\pi i/n}, \dots, e^{(n-1)2\pi i/n}\}$$

and that $e^{2\pi i/n}$ generates μ_n . Thus μ_n is cyclic of order n . This group is therefore isomorphic to C_n , via the map

$$\begin{aligned} C_n &\longrightarrow \mu_n \\ a^j &\longmapsto e^{2j\pi i/n}. \end{aligned}$$

Exercise 18. Let $p > 0$ be a prime. Show that every group of order p is cyclic.

Chapter 4

Quotient groups

Recall from your undergraduate algebra course the construction for the integers modulo n : one starts with an equivalence relation \sim on \mathbb{Z} , considers the set \mathbb{Z}/n of all equivalence classes with respect to this equivalence relation, and verifies that the operations on \mathbb{Z} give rise to well defined binary operations on the set of equivalence classes.

This idea still works if we replace \mathbb{Z} by an arbitrary group, but one has to be somewhat careful about what equivalence relation is used.

4.1 Equivalence relations on a group and cosets

Let G be a group and consider an equivalence relation \sim on G . Let G/\sim denote the set of equivalence classes for \sim and write $[g]$ for the equivalence class that the element $g \in G$ belongs to, that is

$$[x] := \{g \in G \mid g \sim x\}.$$

When does G/\sim acquire the structure of a group under the operation

$$[x] \cdot [y] := [xy] ?$$

Right away, we should be worried about whether this operation is well-defined, meaning that it is independent of our choice of representatives for each class. That is, if $[x] = [x']$ and $[y] = [y']$ then must $[xy] = [x'y']$? In other words, if $x \sim x'$ and $y \sim y'$, must $xy \sim x'y'$?

Definition 4.1. We say an equivalence relation \sim on a group G is **compatible with multiplication** if $x \sim y$ implies $xz \sim yz$ and $zx \sim zy$ for all $x, y, z \in G$.

Lemma 4.2. *For a group G and equivalence relation \sim , the rule $[x] \cdot [y] = [xy]$ is well-defined and makes G/\sim into a group if and only if \sim is compatible with multiplication.*

Proof. To say that the rule $[x] \cdot [y] = [xy]$ is well-defined is to say that for all $x, x', y, y' \in G$ we have

$$[x] = [x'] \text{ and } [y] = [y'] \implies [x][y] = [x'][y'].$$

So $[xy] = [x'y']$ if and only if whenever $x \sim x'$ and $y \sim y'$, then $xy \sim x'y'$.

Assume \sim is compatible with multiplication. Then $x \sim x'$ implies $xy \sim x'y$ and $y \sim y'$ implies $x'y \sim x'y'$, hence by transitivity $xy \sim x'y'$. Thus $[x] \cdot [y] = [xy]$ is well-defined.

Conversely, assume the rule $[x] \cdot [y] = [xy]$ is well-defined, so that

$$[x] = [x'] \text{ and } [y] = [y'] \implies [x][y] = [x'][y'].$$

Setting $y = y'$ gives us

$$x \sim x' \implies xy \sim x'y.$$

Setting $x = x'$ gives us

$$y \sim y' \implies xy \sim xy'.$$

Hence \sim is compatible with multiplication.

So now assume that the multiplication rule is well-defined, which we have now proved is equivalent to saying that \sim is compatible with the multiplication in G . We need to prove that G/\sim really is a group. Indeed, since G itself is a group then given any $x, y, z \in G$ we have

$$[x] \cdot ([y] \cdot [z]) = [x] \cdot [yz] = [x(yz)] = [(xy)z] = [xy][z] = ([x][y])[z]$$

Moreover, for all $x \in G$ we have

$$[e_G][x] = [e_Gx] = [x] \quad \text{and} \quad [x][e_G] = [xe_G] = [x],$$

so that $[e_G]$ is an identity for G/\sim . Finally,

$$[x][x^{-1}] = [e_G] = e_{G/\sim},$$

so that every element in G/\sim has an inverse; in fact, this shows that $[x]^{-1} = [x^{-1}]$. \square

Definition 4.3. Let G be a group and let \sim be an equivalence relation on G that is compatible with multiplication. The **quotient group** is the set G/\sim of equivalence classes, with group multiplication $[x] \cdot [y] = [xy]$.

Example 4.4. Let $G = \mathbb{Z}$ and fix an integer $n \geq 1$. Let \sim be the equivalence relation given by congruence modulo n , so $a \sim b \iff a \equiv b \pmod{n}$. Then

$$(\mathbb{Z}, +)/\sim = (\mathbb{Z}/n, +).$$

But how do we come up with equivalence relations that are compatible with the group law?

Definition 4.5. Let H be a subgroup of a group G . The **left action of H on G** is given by

$$h \cdot g = hg \quad \text{for } h \in H, g \in G.$$

The equivalence relation \sim_H on G induced by the left action of H is given by

$$a \sim_H b \text{ if and only if } b = ha \text{ for some } h \in H.$$

The equivalence class of $g \in G$, also called the **orbit** of g , and also called the **right coset** of H in G containing g , is

$$Hg := \{hg \mid h \in H\}.$$

There is also a **left coset** of H in G containing g , defined by

$$gH := \{gh \mid h \in H\}.$$

Example 4.6. Let $G = \mathbb{Z}$ and $H = \langle n \rangle = n\mathbb{Z} = \{nk \mid k \in \mathbb{Z}\}$. Then

$$x \sim_{n\mathbb{Z}} y \iff x = y + nk \text{ for some } k \in \mathbb{Z} \iff x \equiv y \pmod{n}.$$

Therefore the equivalence relation $\sim_{n\mathbb{Z}}$ is the same as congruence modulo n and the right and left cosets of $n\mathbb{Z}$ in \mathbb{Z} are the congruence classes of integers modulo n .

Lemma 4.7. Let $H \leq G$. The following facts about left cosets are equivalent for $x, y \in G$:

1. The elements x and y belong to the same left coset of H in G .
2. $x = yh$ for some $h \in H$.
3. $y = xh$ for some $h \in H$.
4. $y^{-1}x \in H$.
5. $x^{-1}y \in H$.
6. $xH = yH$.

Analogously, the following facts about right cosets are equivalent for all $x, y \in G$:

1. The elements x and y belong to the same right coset of H in G .
2. There exists $h \in H$ such that $x = hy$.
3. There exists $h \in H$ such that $y = hx$.
4. We have $yx^{-1} \in H$.
5. We have $xy^{-1} \in H$.
6. We have $Hx = Hy$.

Proof. We will only prove the statements about left cosets, since the statements about right cosets are analogous.

(1. \Rightarrow 2.) Suppose that x and y belong to the same left coset gH of H in G . Then $x = ga$ and $y = gb$ for some $a, b \in H$, so $g = yb^{-1}$ and therefore $x = yb^{-1}h = ya$ where $h = b^{-1}a \in H$.

(2. \Leftrightarrow 3.) We have $x = yh$ for some $h \in H$ if and only if $y = xh^{-1}$ and $h^{-1} \in H$.

(2. \Leftrightarrow 4.) We have $x = yh$ for some $h \in H$ if and only if $y^{-1}x = h \in H$.

(4. \Leftrightarrow 5.) Note that $y^{-1}x \in H \Leftrightarrow (y^{-1}x)^{-1} \in H \Leftrightarrow x^{-1}y \in H$.

(2. \Rightarrow 6.) Suppose $x = ya$ for some $a \in H$. Then by 2. \Rightarrow 3. we also have $y = xb$ for some $b \in H$. Note that for all $h \in H$, we also have $ah \in h$ and $bh \in H$. Then

$$xH = \{xh \mid h \in H\} = \underbrace{\{y(ah) \mid h \in H\}}_{\in H} \subseteq yH$$

and

$$yH = \{yh \mid h \in H\} = \underbrace{\{x(bh) \mid h \in H\}}_{\in H} \subseteq xH.$$

Therefore, $xH = yH$.

(6. \Rightarrow 1.) Since $e_G = e_H \in H$, we have $x = xe_G \in xH$ and $y = ye_G \in yH$. If $xH = yH$ then, x and y belong to the same left coset. \square

Remark 4.8. Note that Lemma 4.7 says in particular that \sim_H is compatible with multiplication.

Lemma 4.9. For $H \leq G$, the collection of left cosets of H in G form a partition of G , and similarly for the collection of right cosets:

$$\bigcup_{x \in G} xH = G$$

and for all $x, y \in G$, either $xH = yH$ or $xH \cap yH = \emptyset$.

The analogous statement for right cosets also holds. Moreover, all left and right cosets have the same cardinality: for any $x \in G$,

$$|xH| = |Hx| = |H|.$$

Proof. Since the left (respectively, right) cosets are the equivalence classes for an equivalence relation, the first part of the statement is just a special case of a general fact about equivalence relation.

Let us nevertheless write a proof for the assertions for right cosets. Every element $g \in G$ belongs to at least one right coset, since $e \in H$ gives us $g \in Hg$. Thus

$$\bigcup_{x \in G} xH = G.$$

Now we need to show any two cosets are either identical or disjoint: if Hx and Hy share an element, then it follows from 1. \Rightarrow 6. of Lemma 4.7 that $Hx = Hy$. This proves that the right cosets partition G .

To see that all right cosets have the same cardinality as H , consider the function

$$\rho: H \rightarrow Hg \quad \text{defined by} \quad \rho(h) = hg.$$

This function ρ is surjective by construction. Moreover, if $\rho(h) = \rho(h')$ then $hg = h'g$ and thus $h = h'$. Thus ρ is also injective, and therefore a bijection, so $|Hg| = |H|$. \square

Definition 4.10. The number of left cosets of a subgroup H in a finite group G is denoted by $[G : H]$ and called the **index** of H in G . Equivalently, the index $[G : H]$ is the number of right cosets of H .

We can now write a fancier version of Lagrange's Theorem 3.20; we leave the proof as an exercise.

Corollary 4.11 (Lagrange's Theorem revisited). *If G is a finite group and $H \leq G$, then*

$$|G| = |H| \cdot [G : H].$$

In particular, $|H|$ is a divisor of $|G|$.

Another way to write this: if G is finite and H is any subgroup of G , then

$$[G : H] = \frac{|G|}{|H|}.$$

Example 4.12. For $G = D_n$ and $H = \langle s \rangle = \{e, s\}$, the left cosets gH of H in G are

$$\{e, s\}, \quad \{r, rs\}, \quad \{r^2, r^2s\}, \dots, \{r^{n-1}, r^{n-1}s\}$$

and the right cosets Hg are

$$\{e, s\}, \quad \{r, r^{-1}s\}, \quad \{r^2, r^{-2}s\}, \dots, \{r^{n-1}, r^{-n+1}s\}.$$

Note that these lists are *not* the same, but they do have the same length. For example, r is in the left coset $\{r, rs\}$, while its right coset is $\{r, r^{-1}s\}$. We have $|G| = 2n$, $|H| = 2$ and $[G : H] = n$.

Keeping $G = D_n$ but now letting $K = \langle r \rangle$, the left cosets are K and

$$sK = \{s, sr, \dots, sr^{n-1}\} = \{s, r^{n-1}s, r^{n-2}s, \dots, rs\}$$

and the right cosets are K and

$$Ks = \{s, r^{n-1}s, r^{n-2}s, \dots, rs\}.$$

In this case $sK = Ks$, and the left and right cosets are exactly the same. We have $|G| = 2n$, $|H| = n$ and $[G : H] = 2$.

4.2 Normal subgroups

Definition 4.13. A subgroup N of a group G is **normal** in G , written $N \trianglelefteq G$, if

$$gNg^{-1} = N \quad \text{for all } g \in G.$$

Example 4.14.

- (1) The trivial subgroups $\{e\}$ and G of a group G are always normal.
- (2) Any subgroup of an abelian group is normal.
- (3) For any group G , $Z(G) \trianglelefteq G$.

Remark 4.15. The relation of being a normal subgroup is not transitive. For example, for

$$V = \{e, (12)(34), (13)(24), (14)(23)\}$$

one can show that $V \trianglelefteq S_4$ (see Lemma 4.21 below), and since V is abelian (because you proved before that all groups with 4 elements are abelian!), the subgroup $H = \{e, (12)(34)\}$ is normal in V . But H is *not* normal in S_4 , since for example

$$(13)[(12)(34)](13)^{-1} = (32)(14) \notin H.$$

Lemma 4.16. Assume N is a subgroup of G . The following conditions are equivalent.

- (a) N is a normal subgroup of G , meaning that $gNg^{-1} = N$ for all $g \in G$.
- (b) We have $gNg^{-1} \subseteq N$ for all $g \in G$, meaning that $gng^{-1} \in N$ for all $n \in N$ and $g \in G$.
- (c) The right and left cosets of N agree. More precisely, $gN = Ng$ for all $g \in G$.
- (d) We have $gN \subseteq Ng$ for all $g \in G$.
- (e) We have $Ng \subseteq gN$ for all $g \in G$.

Proof. Note that $gNg^{-1} = N$ if and only if $gN = Ng$ and hence (1) \iff (3).

The implication (a) \Rightarrow (b) is immediate. Conversely, if $gNg^{-1} \subseteq N$ for all g , then

$$N = g^{-1}(gNg^{-1})g \subseteq g^{-1}Ng.$$

Thus (b) implies (a).

Finally, (b), (d), and (e) are all equivalent since

$$gNg^{-1} \subseteq N \iff gN \subseteq Ng$$

and

$$g^{-1}Ng \subseteq N \iff Ng \subseteq gN. \quad \square$$

Exercise 19. Kernels of group homomorphisms are normal.

We will see later that, conversely, all normal subgroups are kernels of group homomorphisms.

Exercise 20. Any subgroup of index two is normal.

Exercise 21. Preimages of normal subgroups are normal, that is, if $f : G \rightarrow H$ is a group homomorphism and $K \trianglelefteq H$, then $f^{-1}(K) \trianglelefteq G$.

Remark 4.17. Let $A \leq B$ be subgroups of a group G . If A is a normal subgroup of G , then in particular for all $b \in B$ we have

$$bab^{-1} \in A,$$

since $b \in B \subseteq G$. Therefore, A is a normal subgroup of B .

Example 4.18. Let us go back to Example 4.12, where we considered the group $G = D_n$ and the subgroups

$$H = \langle s \rangle = \{e, s\} \quad \text{and} \quad K = \langle r \rangle.$$

We showed that the left and right cosets of H are not the same, and thus H is not a normal subgroup of G . We also showed that the left and right cosets of K are in fact the same, which proves that K is a normal subgroup of G . Note that H is nevertheless a very nice group – it is cyclic and thus abelian – despite not being a normal subgroup of G . This indicates that whether a subgroup H is a normal subgroup of G has a lot more to do about the relationship between H and G than the properties of H as a group on its own.

Definition 4.19. The **alternating group** A_n is the subgroup of S_n generated by all products of two transpositions.

Remark 4.20. Recall that we proved in Theorem 1.44 that the sign of a permutation is well-defined. Notice also that the inverse of an even permutation must also be even, and the product of any two even permutations is even, and thus A_n can also be described as the set of all even permutations.

Lemma 4.21. For all $n \geq 2$, $A_n \trianglelefteq S_n$.

Proof. Consider the sign map $\text{sign}: S_n \rightarrow \mathbb{Z}/2$ that takes each permutation to its sign, meaning

$$\text{sign}(\sigma) = \begin{cases} 1 & \text{if } \sigma \text{ is even} \\ -1 & \text{if } \sigma \text{ is odd.} \end{cases}$$

This a group homomorphism (exercise!), and by construction the kernel of sign is A_n . By Exercise 19, we conclude that A_n must be a normal subgroup of S_n .

Alternatively, we can prove Lemma 4.21 by showing that A_n is a subgroup of S_n of index 2, and using Exercise 20. \square

The last condition in Lemma 4.16 implies that for all $g \in G$ and $n \in N$, we have $gn = n'g$ for some $n' \in N$, which is precisely what was needed to make the group law on G/\sim_H well-defined. Recall that

$$a \sim_H b \text{ if and only if } b = ha \text{ for some } h \in H.$$

Lemma 4.22. Let G be a group. An equivalence relation \sim on G is compatible with multiplication if and only if $\sim = \sim_N$ for some normal subgroup $N \trianglelefteq G$.

Proof. (\Rightarrow) Suppose \sim is compatible with multiplication, and set $N := \{g \in G \mid g \sim e\}$. Then we claim that $N \trianglelefteq G$ and $\sim = \sim_N$.

To see that $N \trianglelefteq G$, let $n \in N$ and $g \in G$. Since $n \in N$, then $n \sim e$, and thus since \sim is compatible with multiplication we conclude that for all $g \in G$ we have

$$gng^{-1} \sim geg^{-1} = e \in N.$$

This shows that $gng^{-1} \subseteq N$ for any $n \in N$ and any $g \in G$, and thus N is a normal subgroup of G by Lemma 4.16.

It remains to check that $\sim = \sim_N$. Given any $a, b \in G$, since \sim is compatible with multiplication then

$$a \sim b \implies ab^{-1} \sim bb^{-1} = e \implies ab^{-1} \in N.$$

Thus there exists some $h \in H$ such that

$$ab^{-1} = h \implies a = hb. \iff a \sim_H b.$$

(\Leftarrow) If $\sim = \sim_N$, then in particular \sim is compatible with multiplication. Let $x, y, z \in G$ such that $x \sim_N y$. Then $y = nx$ for some $n \in N$, so $yz = nxz$ and

$$zy = znx = zn(z^{-1}z)x = (znz^{-1})zx = n'zx$$

for some $n' \in N$, where the last equality uses the normal subgroup property. We deduce that $yz \sim_N xz$ and $zy \sim_N zx$. \square

4.3 Quotient groups

Definition 4.23. Let N be a normal subgroup of a group G . The **quotient group** G/N is the group G/\sim_N , where \sim_N is the equivalence relation induced by the left action of N on G . Thus G/N is the set of left cosets of N in G , and the multiplication is given by

$$xN \cdot yN := (xy)N.$$

The identity elements is $e_G N = N$ and for each $g \in G$, the inverse of gN is $(gN)^{-1} = g^{-1}N$.

Remark 4.24. Note that, by Lemma 4.9, G/N is also the set of right cosets of N in G with multiplication given by

$$Nx \cdot Ny := N(xy).$$

In order to prove statements about a quotient G/N , it is often useful to rewrite those statements in terms of elements in the original group G , but one needs to be careful when translating.

Remark 4.25. Given a group G and a normal subgroup N , equality in the quotient does not mean that the representatives are equal. By Lemma 4.7,

$$gN = hN \iff gh^{-1} \in N.$$

In particular, $gN = N$ if and only if $g \in N$.

Remark 4.26. Note that $|G/N| = [G : N]$. By Lagrange's Theorem, if G is finite then

$$|G/N| = \frac{|G|}{|N|}.$$

Example 4.27. We saw in Example 4.18 that the subgroup $N = \langle r \rangle$ of D_n is normal. The quotient D_n/N has just two elements, N and sN , and hence it must be cyclic of order 2, since that is the only one group of order 2. In fact, note that $|N| = n$ and $|D_n| = 2n$, so by Lagrange's Theorem

$$|D_n/N| = \frac{2n}{n} = 2.$$

Example 4.28. The **infinite dihedral group** D_∞ is the set

$$D_\infty = \{r^i, r^i s \mid i \in \mathbb{Z}\}$$

together with the multiplication operation defined by

$$r^i \cdot r^j = r^{i+j}, \quad r^i \cdot (r^j s) = r^{i+j} s, \quad (r^i s) \cdot r^j = r^{i-j} s, \quad \text{and} \quad (r^i s)(r^j s) = r^{i-j}.$$

One can show that D_∞ is the group with presentation

$$D_\infty = \langle r, s \mid s^2 = e, srs = r^{-1} \rangle.$$

Then $\langle r^n \rangle \trianglelefteq D_\infty$ and $D_\infty/\langle r^n \rangle \cong D_n$ via the map $r\langle r^n \rangle \mapsto r$ and $s\langle r^n \rangle \mapsto s$.

Remark 4.29. In Example 4.28 above, both groups D_∞ and $\langle r^n \rangle$ are infinite, but

$$[D_\infty : \langle r^n \rangle] = |D_\infty / \langle r^n \rangle| = |D_n| = 2n.$$

This shows that the quotient of an infinite group by an infinite subgroup can be a finite group.

The quotient of an infinite group by an infinite subgroup can also be infinite. In contrast, a quotient of any finite group must necessarily be finite.

Lemma 4.30. *Let G be a group and consider a normal subgroup N of G . Then the map*

$$\begin{aligned} G &\xrightarrow{\pi} G/N \\ g &\longmapsto \pi(g) = gN \end{aligned}$$

is a surjective group homomorphism with $\ker(\pi) = N$.

Proof. Surjectivity is immediate from the definition. Now we claim that π is a group homomorphism:

$$\begin{aligned} \pi(gg') &= (gg')N && \text{by definition of } \pi \\ &= gN \cdot g'N && \text{by definition of the multiplication on } G/N \\ &= \pi(g)\pi(g') && \text{by definition of } \pi. \end{aligned}$$

Finally, by Lemma 4.7, we have

$$\ker(\pi) = \{g \in G \mid gN = e_G N\} = N. \quad \square$$

Definition 4.31. Let G be any group and N be a normal subgroup of G . The group homomorphism

$$\begin{aligned} G &\xrightarrow{\pi} G/N \\ g &\longmapsto \pi(g) = gN \end{aligned}$$

is called the **canonical (quotient) map**, the **canonical surjection**, or the **canonical projection** of G onto G/N .

The canonical projection is a surjective homomorphism. We might indicate that in our notation by writing $\pi: G \twoheadrightarrow G/N$. More generally

Notation 4.32. If $f: A \rightarrow B$ is a surjective function, we might write $f: A \twoheadrightarrow B$ to denote that surjectivity.

Normal subgroups are precisely those that can be realized as kernels of a group homomorphism.

Corollary 4.33. *A subgroup N of a group G is normal in G if and only if N is the kernel of a homomorphism with domain G .*

Proof. By Exercise 19, the kernel of any group homomorphism is a normal subgroup; we have just shown in Lemma 4.30 that every normal subgroup can be realized as the kernel of a group homomorphism. \square

Definition 4.34. Let G be any group. For $x, y \in G$, the **commutator** of x and y is the element

$$[x, y] := xyx^{-1}y^{-1}.$$

The **commutator subgroup** or **derived subgroup** of G , denoted by G' or $[G, G]$, is the subgroup generated by all commutators of elements in G . More precisely,

$$[G, G] := \langle [x, y] \mid x, y \in G \rangle.$$

Remark 4.35. Note that $[x, y] = e$ if and only if $xy = yx$. More generally, $[G, G] = \{e_G\}$ if and only if G is abelian.

The commutator subgroup measures how far G is from being abelian: if the commutator is as small as possible, then G is abelian, so a larger commutator indicates the group is somehow further from being abelian.

Remark 4.36 (The commutator is a normal subgroup). A typical element of $[G, G]$ has the form

$$[x_1, y_1] \cdots [x_k, y_k] \quad \text{for } k \geq 1 \text{ and } x_1, \dots, x_k, y_1, \dots, y_k \in G.$$

We do not need to explicitly include inverses since

$$[x, y]^{-1} = yxy^{-1}x^{-1} = [y, x].$$

Exercise 22. Show that $[G, G]$ is a normal subgroup of G .

Definition 4.37. Let G be a group and $[G, G]$ be its commutator subgroup. The associated quotient group

$$G^{\text{ab}} := G/[G, G]$$

is called the **abelianization** of G .

Remark 4.38. In this remark we will write G' instead of $[G, G]$ for convenience. The abelianization G/G' of any group G is an abelian, since

$$[xG', yG'] = [x, y]G' = G' = e_{G/G'}$$

for all $x, y \in G$.

Exercise 23. Let G be any group. The abelianization of G is the *largest* quotient of G that is abelian, in the sense that if G/N is abelian for some normal subgroup N , then $[G, G] \subseteq N$.

It is now time to prove the famous (and very useful!) Isomorphism Theorems.

4.4 The Isomorphism Theorems for groups

Theorem 4.39 (Universal Mapping Property (UMP) of a Quotient Group). *Let G be a group and N a normal subgroup. Given any group homomorphism $f: G \rightarrow H$ with $N \subseteq \ker(f)$, there exists a unique group homomorphism*

$$\bar{f}: G/N \rightarrow H$$

such that the triangle

$$\begin{array}{ccc} & G & \\ \pi \searrow & & \searrow f \\ G/N & \dashrightarrow & H \\ & \bar{f} & \end{array}$$

commutes, meaning that $\bar{f} \circ \pi = f$.

Moreover, $\text{im}(f) = \text{im}(\bar{f})$. In particular, if f is surjective, then \bar{f} is also surjective. Finally,

$$\ker(\bar{f}) = \ker(f)/N := \{gN \mid f(g) = e_H\}.$$

Proof. Suppose that such a homomorphism \bar{f} exists. Since $f = \pi \circ \bar{f}$, then \bar{f} has to be given by

$$\bar{f}(gN) = \bar{f}(\pi(g)) = f(g).$$

In particular, \bar{f} is necessarily unique. To show existence, we just need to show that this formula determines a well-defined homomorphism. Given $xN = yN$, we have

$$y^{-1}x \in N \subseteq \ker(f)$$

and so

$$f(y)^{-1}f(x) = f(y^{-1}x) = e \implies f(y) = f(x).$$

This shows that \bar{f} is well-defined. Moreover, for any $x, y \in G$, we have

$$\bar{f}((xN)(yN)) = \bar{f}((xy)N) = f(xy) = f(x)f(y) = \bar{f}(xN)\bar{f}(yN).$$

Thus \bar{f} is a group homomorphism.

The fact that $\text{im } f = \text{im } \bar{f}$ is immediate from the formula for \bar{f} given above, and hence f is surjective if and only if \bar{f} is surjective.

Finally, we have

$$xN \in \ker(\bar{f}) \iff \bar{f}(xN) = e_H \iff f(x) = e_H \iff x \in \ker(f).$$

Therefore, if $xN \in \ker(\bar{f})$ then $xN \in \ker(f)/N$. On the other hand, if $xN \in \ker(f)/N$ for some $x \in G$, then $xN = yN$ for some $y \in \ker(f)$ and hence $x = yz$ for some $z \in N$. Since $N \subseteq \ker(f)$, then $x, y \in \ker(f)$, and thus we conclude that $x = yz \in \ker(f)$. \square

In short, the UMP of quotient groups says that to give a homomorphism from a quotient G/N is the same as to give a homomorphism from G with kernel containing N .

Corollary 4.40. *Let G be any group and let A be an abelian group. Any group homomorphism $f: G \rightarrow A$ must factor uniquely through the abelianization G^{ab} of G : there exists a unique homomorphism \bar{f} such that f factors as the composition*

$$f: G \xrightarrow{\pi} G/[G, G] \xrightarrow{\bar{f}} A.$$

Proof. Let $\pi: G \rightarrow G^{\text{ab}} = G/[G, G]$ be the canonical projection. Since A is abelian, then

$$f([x, y]) = [f(x), f(y)] = e$$

for all $x, y \in G$, and thus $[G, G] \subseteq \ker(f)$. By Theorem 4.39, the homomorphism f must uniquely factor as

$$f: G \xrightarrow{\pi} G/[G, G] \xrightarrow{\bar{f}} A. \quad \square$$

The slogan for the previous result is that any homomorphism from a group G to any abelian group factors uniquely through the abelianization $G/[G, G]$ of G .

We are now ready for the First (and most important) Isomorphism Theorem.

Theorem 4.41 (First Isomorphism Theorem). *If $f: G \rightarrow H$ is a homomorphism of groups, then $\ker(f) \trianglelefteq G$ and the map \bar{f} defined by*

$$\begin{aligned} G/\ker(f) &\xrightarrow{\bar{f}} H \\ g \cdot \ker(f) &\longmapsto f(g) \end{aligned}$$

induces an isomorphism

$$\bar{f}: G/\ker(f) \xrightarrow{\cong} \text{im}(f).$$

In particular, if f is surjective, then f induces an isomorphism $\bar{f}: G/\ker(f) \xrightarrow{\cong} H$.

Proof. The fact that the kernel is a normal subgroup is Exercise 19. Let us first restrict the target of f to $\text{im}(f)$, so that we can assume without loss of generality that f is surjective. By Theorem 4.39, there exists a (unique) homomorphism \bar{f} such that $\bar{f} \circ \pi = f$, where $\pi: G \rightarrow G/\ker(f)$ is the canonical projection. Moreover, the kernel $\ker(\bar{f})$ of \bar{f} consists of just one element, the coset $\ker(f)$ of the identity, and so \bar{f} is injective. Moreover, Theorem 4.39 also says that the image of \bar{f} equals the image of f . We conclude that \bar{f} is an isomorphism. \square

Example 4.42. Let F be a field and consider $G = \text{GL}_n(F)$ for some integer $n \geq 1$. We claim that $H = \text{SL}_n(F)$, the square matrices with determinant 1, is a normal subgroup of $G = \text{GL}_n(F)$. Indeed, given $A \in \text{GL}_n(F)$ and $B \in \text{SL}_n(F)$, then

$$\det(ABA^{-1}) = \det(A)\underbrace{\det(B)}_1\det(A)^{-1} = \det(A)\det(A)^{-1} = 1,$$

so $ABA^{-1} \in H$. The map

$$\det: \text{GL}_n(F) \rightarrow (F^\times, \cdot)$$

is a surjective group homomorphism whose kernel is by definition of $\text{SL}_n(F)$. By the First Isomorphism Theorem,

$$\text{GL}_n(F)/\text{SL}_n(F) \cong (F^\times, \cdot).$$

Example 4.43. Note that $N = (\{\pm 1\}, \cdot)$ is a subgroup of $G = (\mathbb{R} \setminus \{0\}, \cdot)$, and N is normal in G since G is abelian. We claim that G/N is isomorphic to $(\mathbb{R}_{>0}, \cdot)$. To prove this, define

$$f: \mathbb{R}^\times \rightarrow \mathbb{R}_{>0}$$

to be the absolute value function, so that $f(r) = |r|$. Then f is a surjective homomorphism and its kernel is N . The [First Isomorphism Theorem](#) gives

$$G/N \cong (\mathbb{R}_{>0}, \cdot).$$

Example 4.44. We showed in Example 4.27 that $D_n/\langle r \rangle$ is isomorphic to the cyclic group of order 2. Let us now reprove that fact using the [First Isomorphism Theorem](#).

Recall that $(\{\pm 1\}, \cdot)$ is a group with \cdot the usual multiplication. Define $f: D_n \rightarrow \{\pm 1\}$ by

$$f(\alpha) = \begin{cases} 1 & \text{if } \alpha \text{ preserves orientation} \\ -1 & \text{if } \alpha \text{ reverses orientation} \end{cases} = \begin{cases} 1 & \text{if } \alpha \text{ is a rotation} \\ -1 & \text{if } \alpha \text{ is a reflection.} \end{cases}$$

One can show (exercise!) that this is a surjective homomorphism with kernel $\ker f = \langle r \rangle$, and hence by the [First Isomorphism Theorem](#)

$$D_n/\langle r \rangle \cong (\{\pm 1\}, \cdot).$$

To set up the Second Isomorphism Theorem, we need some more background first.

Definition 4.45. Given subgroups H and K of a group G , we define the subset HK of G by

$$HK := \{hk \mid h \in H, k \in K\}.$$

Note that HK is in general only a subset of G , not a subgroup.

Remark 4.46. Given subgroups H and K of a group G , note that H and K are both subgroups of HK . For example, any element $h \in H$ is in HK because $e \in K$ and $h = he \in HK$.

Exercise 24. Let H and K be subgroups of G .

- (1) The subset HK is a subgroup of G if and only if $HK = KH$.
- (2) If at least one of H or K is a normal subgroup of G , then

$$HK \leq G \quad \text{and} \quad HK = KH = \langle H \cup K \rangle.$$

Warning! The identity $HK = KH$ does not mean that every pair of elements from H and K must commute, as the example below will show; this is only an equality of sets.

Example 4.47. In D_n , consider the subgroups $H = \langle s \rangle$ and $K = \langle r \rangle$. The work we did in Example 4.12 shows that

$$HK = KH = D_2,$$

but r and s do not commute. The fact that $HK = KH$ can also be justified by observing that $K \trianglelefteq D_n$ (see Example 4.18) and using Exercise 24.

Theorem 4.48 (Second Isomorphism Theorem). *Let G be a group, $H \leq G$, and $N \trianglelefteq G$. Then*

$$HN \leq G, \quad N \cap H \trianglelefteq H, \quad N \trianglelefteq HN$$

and there is an isomorphism

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

given by

$$h \cdot (N \cap H) \mapsto hN.$$

Proof. We leave the facts that $HN \leq G$ and $N \cap H \trianglelefteq H$ as exercises. Since $N \trianglelefteq G$, then $N \trianglelefteq HN$. Let $\pi: HN \rightarrow \frac{HN}{N}$ be the canonical projection. Define

$$\begin{aligned} H &\xrightarrow{f} \frac{HN}{N} \\ h &\mapsto f(h) = hN. \end{aligned}$$

This is a homomorphism, since it is the composition of homomorphisms

$$f: H \subseteq HN \xrightarrow{\pi} \frac{HN}{N},$$

where the first map is just the inclusion. Moreover, f is surjective since

$$hnN = hN = f(h)$$

for all $h \in H$ and $n \in N$. The kernel of f is

$$\ker(f) = \{h \in H \mid hN = N\} = H \cap N.$$

The result now follows from the [First Isomorphism Theorem](#) applied to f . □

Corollary 4.49. *If H and N are finite subgroups of G and $N \trianglelefteq G$, then*

$$|HN| = \frac{|H| \cdot |N|}{|H \cap N|}.$$

Proof. By Theorem 4.48,

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

The result now follows from Remark 4.26, which is really just an application of [Lagrange's Theorem](#):

$$\frac{|H|}{|N \cap H|} = \frac{|HN|}{|N|}. \quad \square$$

In fact, the corollary is also true without requiring that N is normal.

Example 4.50. Fix a field F and an integer $n \geq 1$. Let $G = \mathrm{GL}_n(F)$ and $N = \mathrm{SL}_n(F)$, and recall that we showed in Example 4.42 that N is a normal subgroup of G . Let H be the set of diagonal invertible matrices, which one can show is also a subgroup of G . One can show that every invertible matrix A can be written as a product of a diagonal matrix and a matrix of determinant 1, and thus $HN = G$. By the [Second Isomorphism Theorem](#),

$$H/(N \cap H) \cong G/N$$

and since we showed in Example 4.42 that

$$G/N \cong (F^\times, \cdot),$$

where $F^\times = F \setminus \{0\}$, we get

$$H/(N \cap H) \cong (F^\times, \cdot).$$

Before we prove what is known as the Third Isomorphism Theorem, we need to get a better understanding of the subgroups of a quotient group. That is the content of what is known as the Lattice Isomorphism Theorem, sometimes (rarely?) called the Fourth Isomorphism Theorem.

Theorem 4.51 (The Lattice Isomorphism Theorem). *Let G be a group and N a normal subgroup of G , and let $\pi : G \twoheadrightarrow G/N$ be the quotient map. There is an order-preserving bijection of posets (a lattice isomorphism)*

$$\begin{array}{ccc} \{ \text{subgroups of } G \text{ that contain } N \} & \xrightleftharpoons[\Phi]{\Psi} & \{ \text{subgroups of } G/N \} \\ H & \longmapsto & \Psi(H) = H/N \\ \Phi(A) = \pi^{-1}(A) = \{x \in G \mid \pi(x) \in A\} & \longleftarrow & A \end{array}$$

Then this bijection enjoys the following properties:

(1) Subgroups correspond to subgroups:

$$H \leq G \iff H/N \leq G/N.$$

(2) Normal subgroups correspond to normal subgroups:

$$H \trianglelefteq G \iff H/N \trianglelefteq G/N.$$

(3) Indices are preserved:

$$[G : H] = [G/N : H/N].$$

(4) Intersections and unions are preserved:

$$H/N \cap K/N = (H \cap K)/N \quad \text{and} \quad \langle H/N \cup K/N \rangle = \langle H \cup K \rangle/N.$$

Proof. We showed in Lemma 4.30 that the quotient map $\pi: G \rightarrow G/N$ is a surjective group homomorphism. It will be useful to rewrite the maps in the statement of the theorem in terms of π . Notice that $\Psi(H) = H/N = \{hN \mid h \in H\} = \pi(H)$. Note that Ψ does indeed land in the correct codomain, since by Lemma 3.8 images of subgroups through group homomorphisms are subgroups, and thus $\pi(H) \leq G/N$ for each $H \leq G$. Thus Ψ is well-defined. We claim Φ also lands in the correct codomain. Indeed, by Exercise 13 preimages of subgroups through group homomorphisms are subgroups, and thus in particular for each $A \leq G$ we have $\pi^{-1}(A) \leq G$. Moreover, for any $A \leq G$ we have $\{e_G N\} \subseteq A$, hence

$$N = \ker(\pi) = \pi^{-1}(\{e_G N\}) \subseteq \pi^{-1}(A) = \Phi(A).$$

Thus Ψ is well-defined.

To show that Ψ is bijective, we will show that Φ and Ψ are mutual inverses. First, note that since π is surjective, then $\pi(\pi^{-1}(A)) = A$ for all subgroups A of G/N , and thus

$$(\Psi \circ \Phi)(A) = \pi(\pi^{-1}(A)) = A.$$

Moreover,

$$\begin{aligned} x \in \pi^{-1}(H/N) &\iff \pi(x) \in H/N \\ &\iff xN = hN && \text{for some } h \in H \\ &\iff x \in hN && \text{for some } h \in H \\ &\iff x \in H && \text{since } N \subseteq H. \end{aligned}$$

Thus

$$(\Phi \circ \Psi)(H) = \pi^{-1}(\pi(H)) = \pi^{-1}(H/N) = H.$$

Thus, Ψ and Φ are well-defined and inverse to each other. Since π and π^{-1} both preserve containments, each of Ψ , Ψ^{-1} preserves containments as well.

Again by Lemma 3.8 and Exercise 13, images and preimages of subgroups by group homomorphisms are subgroups, which proves (1). Moreover, if $N \leq H \leq G$ and $H \trianglelefteq G$, then $ghg^{-1} \in H$ for all $g \in G$ and all $h \in H$, and thus

$$(gN)(hN)(gN)^{-1} = (ghg^{-1})N \in H/N.$$

Therefore, if $N \leq H \trianglelefteq G$, then $H/N \trianglelefteq G/N$. Finally, by Exercise 21, the preimage of a normal subgroup is normal. We have now shown (2).

We leave (3) as an exercise, and (4) is a consequence of the more general fact that lattice isomorphisms preserve suprema and infima. \square

We record here what is left to do.

Exercise 25. Let G be a group and N a normal subgroup of G . For all subgroups H of G with $N \leq H$, show that

$$[G : H] = [G/N : H/N] \quad \text{and} \quad [G : \pi^{-1}(A)] = [G/N : A].$$

Theorem 4.52 (Third Isomorphism Theorem). *Let G be a group, $M \leq N \leq G$, $M \trianglelefteq G$ and $N \trianglelefteq G$. Then*

$$M \trianglelefteq N, \quad N/M \trianglelefteq G/M,$$

and there is an isomorphism

$$\begin{aligned} \frac{(G/M)}{(N/M)} &\xrightarrow{\cong} G/N \\ gM &\longmapsto gN. \end{aligned}$$

Proof. By Remark 4.17, since M is a normal subgroup of G , then it is also a normal subgroup of N . Similarly, the fact that N is normal in G implies that it is normal in G/M , by Theorem 4.51.

The kernel of the canonical map $\pi : G \twoheadrightarrow G/N$ contains M , and so by Theorem 4.39 we get an induced homomorphism

$$\phi : G/M \rightarrow G/N$$

with $\phi(gM) = \pi(g) = gN$. Moreover, we know

$$\ker(\phi) = \ker(\pi)/M = N/M.$$

Finally, apply the First Isomorphism Theorem to ϕ . □

We can now prove the statement about indices in the Lattice Isomorphism Theorem in the case of normal subgroups.

Corollary 4.53. *Let G be a group and N a normal subgroup of G . For all normal subgroups H of G with $N \leq H$,*

$$[G : H] = [G/N : H/N] \quad \text{and} \quad [G : \pi^{-1}(A)] = [G/N : A].$$

Proof. By the Third Isomorphism Theorem,

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

and thus their orders are the same; in particular,

$$[G : H] = |G/H| = \left| \frac{(G/N)}{(H/N)} \right| = [G/N : H/N] = [G/N : H/N]. \quad \square$$

4.5 Presentations as quotient groups

We can finally define group presentations in a completely rigorous manner.

Definition 4.54. Let A be a set. Consider the new set of symbols

$$A^{-1} = \{a^{-1} \mid a \in A\}.$$

Consider the set of all finite words written using symbols in $A \cup A^{-1}$, including the empty word. If a word w contains consecutive symbols aa^{-1} or $a^{-1}a$, we can simplify w by erasing those two consecutive symbols, and we obtain a word that is equivalent to w . If a word cannot be simplified any further, we say that it is **reduced**. Given any $a \in A$, a^1 denotes a , to distinguish it from a^{-1} .

The **free group** on A , denoted $F(A)$, is the set of all reduced words in $A \cup A^{-1}$. In symbols,

$$F(A) = \{a_1^{i_1} a_2^{i_2} \cdots a_m^{i_m} \mid m \geq 0, a_j \in A, i_j \in \{-1, 1\}\}.$$

The set $F(A)$ is a group with the operation in which any two words are multiplied by concatenation.

Example 4.55. The free group on a singleton set $A = x$ is the infinite cyclic group C_∞ .

Theorem 4.56 (Universal mapping property for free groups). *Let A be a set, let $F(A)$ be the free group on A , and let H be any group. Given a function $g: A \rightarrow H$, there is a unique group homomorphism $f: F(A) \rightarrow H$ satisfying $f(a) = g(a)$ for all $a \in A$.*

Proof. Let $f: F(A) \rightarrow H$ be given by

$$f(a_1^{i_1} a_2^{i_2} \cdots a_m^{i_m}) = g(a_1)^{i_1} g(a_2)^{i_2} \cdots g(a_m)^{i_m}$$

for any $m \geq 0$, $a_j \in A$, and $i_j \in \{-1, 1\}$. To check that this is a well-defined function, note that

$$f(a_1^{i_1} a_2^{i_2} \cdots a a^{-1} \cdots a_m^{i_m}) = g(a_1)^{i_1} g(a_2)^{i_2} \cdots g(a)^{-1} \cdots g(a_m)^{i_m} = f(a_1^{i_1} a_2^{i_2} \cdots a_m^{i_m})$$

for any $a \in G$ and similarly for inserting $a^{-1}a$. The fact that f is a group homomorphism and its uniqueness are left as an exercise. \square

Definition 4.57. Let G be a group and let $R \subseteq G$ be a set. The *normal subgroup of G generated by R* , denoted $\langle R \rangle^N$, is the set of all products of conjugates of elements of R and inverses of elements of R . In symbols,

$$\langle R \rangle^N = \{g_1 r_1^{i_1} g_1^{-1} \cdots g_m r_m^{i_m} g_m^{-1} \mid m \geq 0, i_j \in \{1, -1\}, r_j \in R, g_j \in G\}.$$

Definition 4.58. Let A be a set and let R be a subset of the free group $F(A)$. The group with **presentation**

$$\langle A \mid R \rangle = \langle A \mid \{r = e \mid r \in R\} \rangle$$

is defined to be the quotient group $F(A)/\langle R \rangle^N$.

Example 4.59. Let $A = \{x\}$ and consider $R = \{x^n\}$. Then the group with presentation $\langle A \mid R \rangle$ is the cyclic group of order n :

$$C_n = \langle x \mid x^n = e \rangle = \frac{F(\{x\})}{\langle x^n \rangle^N} = C_\infty / \langle x^n \rangle.$$

Example 4.60. Taking $A = \{r, s\}$ and $R = \{s^2, r^n, srsr\}$, $\langle A \mid R \rangle$ is the usual presentation for D_n :

$$D_n = \langle r, s \mid s^2 = e, r^n = e, srsr = e \rangle = \frac{F(\{r, s\})}{\{s^2, r^n, srsr\}^N}.$$

Theorem 4.61 (Universal mapping property of a presentation). *Let A be a set, let $F(A)$ be the free group on A , let R be a subset of $F(A)$, and let H be a group. Let $g: A \rightarrow H$ be a function satisfying the property that whenever $r = a_1^{i_1} \cdots a_m^{i_m} \in R$, with each $a_j \in A, g_j \in G$ and $i_j \in \{1, -1\}$, then*

$$(g(a_1))^{i_1} \cdots (g(a_m))^{i_m} = e_H.$$

Then there is a unique homomorphism $\bar{f}: \langle A \mid R \rangle \rightarrow H$ satisfying

$$\bar{f}(a \langle R \rangle^N) = g(a) \quad \text{for all } a \in A.$$

Proof. By Theorem 4.56, there is a unique group homomorphism $\tilde{f}: F(A) \rightarrow H$ such that $\tilde{f}(a) = g(a)$ for all $a \in A$. Then for

$$r = a_1^{i_1} \cdots a_m^{i_m} \in R$$

we have

$$f(r) = (g(a_1))^{i_1} \cdots (g(a_m))^{i_m} = e_H,$$

showing that $R \subseteq \ker(f)$. Since $\ker(f) \trianglelefteq F(A)$ and $\langle R \rangle^N$ is the smallest normal subgroup containing R , it follows that $\langle R \rangle^N \subseteq \ker(f)$. By Theorem 4.39, f induces a group homomorphism $\bar{f}: G/\langle R \rangle^N \rightarrow H$. Moreover, for each $a \in A$ we have

$$g(a) = f(a) = \bar{f}(a \langle R \rangle^N).$$

□

Remark 4.62. The universal property of a presentation in Theorem 4.61 says that to give a group homomorphism from a group G with a given presentation to a group H is the same as picking images for each of the generators that satisfy the same relations in H as those given in the presentation.

Example 4.63. To find a group homomorphism $D_n \rightarrow \mathrm{GL}_2(\mathbb{R})$, it suffices to pick images for r and s , say $r \mapsto R, s \mapsto S$, and to verify that

$$S^2 = I_2, \quad R^n = I_2, \quad SRSR = I_2.$$

One can check that this does hold for the matrices

$$S = \begin{pmatrix} \cos 2\pi n & -\sin 2\pi n \\ \sin 2\pi n & \cos 2\pi n \end{pmatrix} \quad \text{and} \quad R = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

By the UMP of the presentation there is a unique group homomorphism $D_n \rightarrow \mathrm{GL}_2(\mathbb{R})$ that sends r to R and s to S .

Presentations of groups are remarkably complex mathematical constructions. What makes them so complicated is that $\langle R \rangle^N$ is very hard to calculate in general. The following theorem is a negative answer to what is known as the Word Problem, and illustrates how complicated the story can become:

Theorem 4.64 (Boone-Novikov). *There exists a finite set A and a finite subset R of $F(A)$ such that there exists no algorithm that determines whether a given element of $\langle A \mid R \rangle$ is equal to the trivial element.*

Chapter 5

Group actions... in action

It is time for some more group actions. We will start with some general facts about group actions, and then we will focus on some specific actions and use them to prove results about the structure of finite groups.

5.1 Orbits and Stabilizers

Let G be a group acting on a set S . Let us recall some notation and facts about group actions. The **orbit** of an element $s \in S$ is

$$\text{Orb}_G(s) = \{g \cdot s \mid g \in G\}.$$

A **permutation representation** of a group G is a group homomorphism $\rho: G \rightarrow \text{Perm}(S)$ for some set S . By Lemma 2.3, to give an action of G on a set S is equivalent to giving a permutation representation $\rho: G \rightarrow \text{Perm}(S)$, which is induced by the action via

$$\rho(g)(s) = g \cdot s.$$

An action is **faithful** if the only element $g \in G$ such that $g \cdot s = s$ for all $s \in S$ is $g = e_G$. Equivalently an action is faithful if $\ker(\rho) = \{e_G\}$. An action is **transitive** if for all $p, q \in S$ there is a $g \in G$ such that $q = g \cdot p$. Equivalently, an action is transitive if $\text{Orb}_G(p) = S$ for any $p \in S$.

Definition 5.1. Let G be a group acting on a set S . The **stabilizer** of an element s in S is the set of group elements that fix s under the action:

$$\text{Stab}_G(s) = \{g \in G \mid g \cdot s = s\}.$$

Definition 5.2. Let G be a group acting on a set S . An element $s \in S$ is a **fixed point** of the action if $g \cdot s = s$ for all $g \in G$.

Remark 5.3. Let G be a group acting on a set S . An element $s \in S$ is a fixed point if and only if $\text{Orb}_G(s) = \{s\}$. Moreover, s is a fixed point if and only if $\text{Stab}_G(s) = G$.

The stabilizer of any element is always a subgroup of G .

Lemma 5.4. Let G be a group acting on a set S , and let $s \in S$. The stabilizer $\text{Stab}_G(s)$ of s is a subgroup of G .

Proof. By definition of group action, $e \cdot s = s$, so $e \in \text{Stab}_G(s)$. If $x, y \in \text{Stab}_G(s)$, then $(xy)s = x(ys) = xs = s$ and thus $xy \in \text{Stab}_G(s)$. If $x \in \text{Stab}_G(s)$, then

$$xs = s \Rightarrow s = x^{-1}xs = x^{-1}s \Rightarrow x^{-1} \in \text{Stab}_G(s). \quad \square$$

Theorem 5.5 (Orbit-Stabilizer Theorem). Let G be a group that acts on a set S . For any $s \in S$ we have

$$|\text{Orb}_G(s)| = [G : \text{Stab}_G(s)].$$

Proof. Let \mathcal{L} be the collection of left cosets of $\text{Stab}_G(s)$ in G . Let $\alpha : \mathcal{L} \rightarrow \text{Orb}_G(s)$ be given by

$$\alpha(x\text{Stab}_G(s)) = x \cdot s.$$

This function is well-defined and injective:

$$x\text{Stab}_G(s) = y\text{Stab}_G(s) \iff x^{-1}y \in \text{Stab}_G(s) \iff x^{-1}y \cdot s = s \iff y \cdot s = x \cdot s.$$

The function α is surjective by definition of $\text{Orb}_G(s)$, and thus it is a bijection. Finally, we can now conclude that

$$[G : \text{Stab}_G(s)] = |\mathcal{L}| = |\text{Orb}_G(s)|. \quad \square$$

Corollary 5.6 (Orbit-Stabilizer Theorem part 2). Let G be a finite group acting on a set S . For any $s \in S$ we have

$$|G| = |\text{Orb}_G(s)| \cdot |\text{Stab}_G(s)|.$$

Proof. This is a direct consequence of the Orbit-Stabilizer Theorem, since by [Lagrange's Theorem](#)

$$[G : \text{Stab}_G(s)] = |G| / |\text{Stab}_G(s)|. \quad \square$$

Remark 5.7. Let G be a group acting on a finite set S . The orbits of the action form a partition of S . The one-element orbits correspond to the fixed points of the action. Pick one element s_1, \dots, s_m in each of the other orbits. This gives us the

$$\text{The Orbit Formula: } |S| = (\text{the number of fixed points}) + \sum_{i=1}^m |\text{Orb}_G(s_i)|.$$

By the Orbit-Stabilizer Theorem, we can rewrite this as

$$\text{The Stabilizer Formula: } |S| = (\text{the number of fixed points}) + \sum_{i=1}^m [G : \text{Stab}_G(s_i)].$$

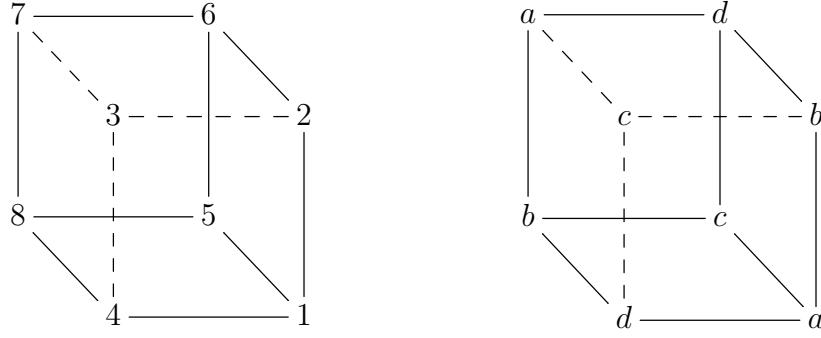
We will later see that these are very useful formulas.

We can now use these simple facts to do some explicit calculations with groups.

Example 5.8. Let G be the group of rotational (orientation-preserving) symmetries of the cube. To count the number of elements of G , think about an isometry as picking up a cube lying on a table, moving it, and placing it back in the same location. To do this, one must pick a face to place on the table. This can be chosen in 6 ways. Once that face is chosen, one needs to decide on where each vertex of that face goes and this can be done in 4 ways. Thus $|G| = 24$.

We can restrict the action of G to the four lines that join opposite vertices of the cube; the group of permutations of the four lines is S_4 , so the corresponding permutation representation associated to this action is a group homomorphism $\rho: G \rightarrow S_4$.

We claim that this homomorphism ρ is actually an isomorphism from G to S_4 . To see this, first label each vertex of the cube 1 through 8. Let a, b, c , and d denote each of the four lines, and let us also label the vertices of the cube a, b, c , or d according to which of the diagonal lines goes through that vertex.



Now note that each face corresponds to a unique order on a, b, c, d , read counterclockwise from the outside of the cube:

The face 1234	corresponds to	$adcb$
The face 1256	corresponds to	$abdc$
The face 1458	corresponds to	$adbc$
The face 5678	corresponds to	$abcd$
The face 2367	corresponds to	$adbc$
The face 3478	corresponds to	$acdb$.

So suppose that $g \in G$ fixes all of the four lines a, b, c, d . Then the face at the bottom must be $abcb$, which corresponds to 1234, and thus all the vertices of the cube in the bottom face must be fixed. We conclude that g must fix the entire cube, and thus g must be the identity.

Thus the action is faithful, and hence the permutation representation $\rho: G \rightarrow S_4$ is injective. Moreover, we showed above that $|G| = 24 = |S_4|$, and thus ρ is an injective function between two finite sets of the same size. We conclude that ρ must actually be a bijection, and thus an isomorphism.

The same group G also acts on the six faces of the cube. This action is transitive, since we can always pick up the cube and put it back on the table with any face on the top. Thus the one and only orbit for the action of G on the six faces of the cube has length 6. By the Orbit-Stabilizer Theorem, it follows that for any face f of the cube, its stabilizer has index

6 and, since we already know that $|G| = 24$, the Orbit-Stabilizer Theorem gives us

$$|\text{Stab}_G(f)| = \frac{|G|}{|\text{Orb}_G(f)|} = \frac{24}{6} = 4.$$

Thus, there are four symmetries that map f to itself. Indeed, they are the 4 rotations by 0, $\frac{\pi}{2}$, π or $\frac{3\pi}{2}$ about the line of symmetry passing through the midpoint of f and the midpoint of the opposite face.

Example 5.9. Let X be a regular dodecahedron, with 12 faces, centered at the origin in \mathbb{R}^3 .

Let G be the group of isometries of the dodecahedron that preserve orientation:

$$G := \{\alpha : \mathbb{R}^3 \rightarrow \mathbb{R}^3 \mid \alpha \text{ is an isometry, } \alpha \text{ preserves orientation, and } \alpha(X) = X\}.$$

This is a subgroup of the group of all bijections from \mathbb{R}^3 to \mathbb{R}^3 . Though not obvious, every element of G is given as rotation about a line of symmetry. There are three kinds of such lines: those joining midpoints of opposite face, those joining midpoints of opposite edges, and those joining opposite vertices. To count the number of elements of G informally, think about an isometry as picking up a dodecahedron that was lying on a table and replacing it in the same location. To do this, one must first pick one of the twelve faces to place on the table, and, for each possible face, there are five ways to orient it. Thus

$$|G| = 12 \cdot 5 = 60.$$

Let us use the Orbit-Stabilizer Theorem to do this more formally. Note that G acts on the collection S of the 12 faces of X . This action is transitive since it is possible to move one face to any other via an appropriate rotation. So, the one and only orbit has length 12. Letting F be any one of the faces, the orientation-preserving isometries of X that map F to itself are just the orientation-preserving elements of D_5 , of which there are 5. Indeed, these correspond to the five rotations of X by $\frac{2\pi nj}{5}$ radians for $j = 0, 1, 2, 4$ about the axis of symmetry passing through the midpoint of F and the midpoint of the opposite face. Applying the [Orbit-Stabilizer Theorem](#) gives

$$|G| = |\text{Orb}_G(F)| \cdot |\text{Stab}_G(F)| = 12 \cdot 5 = 60.$$

5.2 The class equation

The main goal of this subsection is to apply the Orbit-Stabilizer Formula to the action of G on itself by conjugation. Let G be a group. As we saw before, G acts on $S = G$ by conjugation: the action is defined by $g \cdot x = gxg^{-1}$.

Definition 5.10. Let G be a group. Two elements $g, g' \in G$ are **conjugate** if there exists $h \in G$ such that

$$g' = hgh^{-1}.$$

Equivalently, g and g' are conjugate if they are in the same orbit of the conjugation action. The **conjugacy class** of an element $g \in G$ is

$$[g]_c := \{hgh^{-1} \mid h \in G\}.$$

Equivalently, the conjugacy class of g is the orbit of g under the conjugation action.

Remark 5.11. Let G be any group. Then $geg^{-1} = e$ for all $g \in G$, and thus $[e]_c = e = \{e\}$.

Let us study the conjugacy classes of S_n . You proved in a problem set that two cycles in S_n are conjugate if and only if they have the same length:

Lemma 5.12. *For any $\sigma \in S_n$ and distinct integers i_1, \dots, i_p , we have*

$$\sigma(i_1 i_2 \cdots i_p) \sigma^{-1} = (\sigma(i_1) \cdots \sigma(i_p)).$$

Note that the right-hand cycle is a cycle since σ is injective. This generalizes to the following:

Theorem 5.13. *Two elements of S_n are conjugate if and only if they have the same cycle type.*

Proof. Consider two conjugate elements of S_n , say α and $\beta = \sigma\alpha\sigma^{-1}$. By Theorem 1.36, we may write α as a product of disjoint cycles $\alpha = \alpha_1 \cdots \alpha_m$. Then

$$\beta = \sigma\alpha\sigma^{-1} = (\sigma\alpha_1\sigma^{-1}) \cdots (\sigma\alpha_m\sigma^{-1}).$$

Since $\alpha_1, \dots, \alpha_m$ are disjoint cycles, then by Lemma 5.12 the elements $(\sigma\alpha_1\sigma^{-1}), \dots, (\sigma\alpha_m\sigma^{-1})$ are also disjoint cycles, and $\sigma\alpha_i\sigma^{-1}$ has the same length as α_i . We conclude that α and β must have the same cycle type.

Conversely, consider two elements α and β with the same cycle type. More precisely, assume $\alpha = \alpha_1 \cdots \alpha_k$ and $\beta = \beta_1 \cdots \beta_k$ are decompositions into disjoint cycles and that α_i, β_i both have length $p_i \geq 2$ for each i . We need to prove that α and β are conjugate. Let us start with the case $k = 1$. Given two cycles of the same length,

$$\alpha = (i_1 \dots i_p) \quad \text{and} \quad \beta = (j_1 \dots j_p).$$

By Lemma 5.12, any permutation σ such that $\sigma(i_m) = j_m$ for all $1 \leq m \leq p$ must satisfy $\sigma\alpha\sigma^{-1} = \beta$.

Note that such σ has no restrictions on what it does to the set $\{1, \dots, n\} \setminus \{i_1 \dots i_p\}$: it can map $\{1, \dots, n\} \setminus \{i_1 \dots i_p\}$ bijectively to $\{1, \dots, n\} \setminus \{j_1 \dots j_p\}$ in any way possible. From this observation, the general case follows: since the cycles are disjoint, we can find a single permutation σ such that $\sigma\alpha_i\sigma^{-1} = \beta_i$ for all i . \square

We can now classify all the conjugacy classes in S_n based on their cycle type.

Example 5.14. Given Theorem 5.13, we can now write a complete list of the conjugacy classes of S_4 :

- (1) The conjugacy class of the identity $\{e\}$.

- (2) The conjugacy class of (12) , which is the set of all two cycles and has $\binom{4}{2} = 6$ elements.
- (3) The conjugacy class of (123) , which is the set of all three cycles and has $4 \cdot 2 = 8$ elements.
- (4) The conjugacy class of (1234) , which is the set of all four cycles and has $3! = 6$ elements.
- (5) The conjugacy class of $(12)(34)$, which is the set of all products of two disjoint 2-cycles and has 3 elements.

We can check our work by recalling that the conjugacy classes partition S_4 , and indeed we counted 24 elements.

Example 5.15. Given Theorem 5.13, we can now write a complete list of the conjugacy classes of S_5 :

- (1) The conjugacy class of the identity $\{e\}$.
- (2) The conjugacy class of (12) , which is the set of all 2-cycles and has $\binom{5}{2} = 10$ elements.
- (3) The conjugacy class of (123) , containing all 3-cycles, of size $2! \cdot \binom{5}{3} = 20$ elements.
- (4) The conjugacy class of (1234) , containing all 4-cycles, of size $5 \cdot 3! = 30$ elements.
- (5) The conjugacy class of (12345) , which is the set of all 5-cycles, and has $4! = 24$ elements.
- (6) The conjugacy class of $(12)(34)$, which is the set of all products of two disjoint 2-cycles and has $5 \cdot 3 = 15$ elements.
- (7) The conjugacy class of $(12)(345)$, which is the set of all products of a 2-cycle by a 3-cycle, and has $\binom{5}{2} \cdot 2! = 20$ elements.

We can check our work by noting that indeed

$$1 + 10 + 20 + 30 + 24 + 15 + 20 = 120 = 5!.$$

Remark 5.16. For any nontrivial group G , since $[e]_c = \{e\}$ and the conjugacy classes partition G , then $[g]_c \neq G$ for all $g \in G$.

Definition 5.17. Let G be a group and $a \in G$. The **centralizer** of a is the set of elements of G that commute with a :

$$C_G(a) := \{x \in G \mid xa = ax\}.$$

More generally, given a subset $S \subseteq G$, the **centralizer** of S is the set

$$C_G(S) := \{x \in G \mid xs = sx \text{ for all } s \in S\}$$

Definition 5.18. Let G be a group and consider a subset $S \subseteq G$. The **normalizer** of S is the set

$$N_G(S) := \{g \in G \mid gSg^{-1} = S\}.$$

Exercise 26. Let G be a group and $S \subseteq G$. Prove that the centralizer and the normalizer of S are subgroups of G .

Lemma 5.19. Let $S \subseteq G$ be any subset of a group G . Then $C_G(S) \subseteq N_G(S)$.

Proof. Let G be a group and $S \subseteq G$. If $x \in C_G(S)$, then for all $s \in S$ we have

$$xs = sx \implies xsx^{-1} = s \in S \text{ and } x^{-1}sx = s.$$

Thus $xSx^{-1} \subseteq S$ and $x^{-1}Sx \subseteq S$. Now for any $s \in S$ we have $x^{-1}sx \in S$ and s can be written as

$$s = x(x^{-1}sx)x^{-1} \in xSx^{-1}.$$

This shows that $S \subseteq xSx^{-1}$. Thus $xSx^{-1} = S$, and therefore $x \in N_G(S)$. \square

Remark 5.20. If G is an abelian group, then for any $a \in G$ we have $C_G(a) = G = N_G(a)$.

Exercise 27. Let H be a subgroup of a group G , and S a subset of H . Then

$$C_H(S) = C_G(S) \cap H \quad \text{and} \quad N_H(S) = N_G(S) \cap H.$$

Exercise 28. Let G be a group and let H be a subgroup of G . Show that $N_G(H)/C_G(H)$ is isomorphic to a subgroup of the automorphism group $\text{Aut}(H)$ of H .

Exercise 29. Let G be a group and H a subgroup of G . Prove that if H is normal in G , then so is $C_G(H)$, and that $G/C_G(H)$ is isomorphic to a subgroup of the automorphism group of H .

Lemma 5.21. Let G be a group. Consider the action of G on G by conjugation, where $g \cdot h = ghg^{-1}$. For all $g \in G$,

$$\text{Orb}_G(g) = [g]_c \quad \text{and} \quad \text{Stab}_G(g) = C_G(g) \quad \text{and} \quad |[g]_c| = [G : C_G(g)].$$

Proof. The first statement is the definition of the conjugacy class of g : $\text{Orb}_G(g) = [g]_c$. Moreover, by simply following the definitions we see that

$$h \in \text{Stab}_G(g) \iff h \cdot g = g \iff hgh^{-1} = g \iff hg = gh \iff h \in C_G(G).$$

Thus, $\text{Stab}_G(g) = C_G(g)$, and by the [Orbit-Stabilizer Theorem](#),

$$|[g]_c| = |\text{Orb}_G(g)| = [G : C_G(g)]. \quad \square$$

Exercise 30. Let G be a group. Consider the action of G on the power set

$$P(G) = \{S \mid S \subseteq G\}$$

of G by conjugation, meaning $g \cdot S = gSg^{-1}$. For all $S \in P(G)$,

$$\text{Stab}_G(S) = N_G(S) \quad \text{and} \quad |\text{Orb}_G(S)| = [G : N_G(S)].$$

Corollary 5.22. For a finite group G , the size of any conjugacy class divides $|G|$.

Proof. Let $g \in G$. By Lemma 5.21, the order of the conjugacy class of g is the index of the centralizer:

$$|[g]_c| = [G : C_G(g)]$$

By Lagrange's Theorem, the index of any subgroup must divide $|G|$, and thus in particular $|[g]_c|$ divides $|G|$. \square

We will take the Orbit Equation and apply it to the special case of the conjugation action. In order to do that, all that remains is to identify the fixed points of the action.

Lemma 5.23. *Let G be a group acting on itself by conjugation. An element $g \in G$ is a fixed point of the conjugation action if and only if $g \in Z(G)$.*

Proof. (\Leftarrow) Suppose that $g \in Z(G)$. Then for all $h \in G$, g commutes with h , and thus

$$hgh^{-1} = (hg)h^{-1} = g(hh^{-1}) = g.$$

Thus g is conjugate to only itself, meaning it is a fixed point for the conjugation action.

(\Rightarrow) Conversely, suppose that g is a fixed point for the conjugation action. Then for all $h \in G$,

$$hgh^{-1} = h \cdot g = g \implies hg = gh.$$

Thus $g \in Z(G)$. \square

We can now write the Orbit Equation for the conjugation action; this turns out to be a very useful formula.

Theorem 5.24 (The Class Equation). *Let G be a finite group. For each conjugacy class of size greater than 1, pick a unique representative, and let $g_1, \dots, g_r \in G$ be the list of all the chosen representatives. Then*

$$|G| = |Z(G)| + \sum_i^r |G : C_G(g_i)|.$$

Proof. By Lemma 5.23, the elements of $Z(G)$ are precisely the fixed points of the conjugation action. In particular, $|Z(G)|$ counts the number of orbits that have only one element. Because the orbits of the conjugation action partition G , and the conjugacy classes are the orbits, then as noted in Remark 5.7

$$|G| = |Z(G)| + \sum_i^r |[g_i]_c|.$$

By the Orbit-Stabilizer Theorem, the index of the stabilizer is the order of the conjugacy class. Thus for each g_i as in the statement we have

$$|[g_i]_c| = [G : C_G(g_i)].$$

The class equation follows from substituting this into the equation above:

$$|G| = |Z(G)| + \sum_i^r |G : C_G(g_i)|. \quad \square$$

Remark 5.25. The class equation is not very interesting if G is abelian, since there is only one term on the right hand side: $|Z(G)|$.

But when G is nonabelian, the class equation can lead us to discover some very interesting facts, despite its simplicity.

Exercise 31. Prove that if G is a nonabelian group of order 21, then there is only one possible class equation for G , meaning that the numbers appearing in the class equation are uniquely determined up to permutation.

Corollary 5.26. *If p is a prime number and G is a finite group of order p^m for some $m > 0$, then $Z(G)$ is not the trivial group.*

Proof. Let $g_1, \dots, g_r \in G$ be a list of unique representatives of all of the conjugacy classes of G of size greater than 1, as in the class equation. By construction, each g_i is not a fixed point of the action, and thus $\text{Stab}_G(g_i) \neq G$. Then, $C_G(g_i) = \text{Stab}_G(g_i)$, so $C_G(g_i) \neq G$. In particular, $[G : C_G(g_i)] \neq 1$. Since $1 \neq [G : C_G(g_i)]$ and $[G : C_G(g_i)]$ divides $|G| = p^m$, we conclude that p divides $[G : C_G(g_i)]$ for each i . From the class equation, we can now conclude that p divides $|Z(G)|$, and in particular $|Z(G)| \neq 1$. \square

Exercise 32. Let p be prime and let G be a group of order p^m for some $m \geq 1$. Show that if N is a nontrivial normal subgroup of G , then $N \cap Z(G) \neq \{e\}$. In fact, show that $|N \cap Z(G)| = p^j$ for some $j \geq 1$.

Lemma 5.27. *Let G be a group and $N \trianglelefteq G$. The conjugation action of G on itself induces an action by conjugation of G on N . In particular, N is the disjoint union of some of the conjugacy classes in G .*

Proof. Define the conjugation action of G on N by $g \cdot n = gng^{-1}$ for all $g \in G$ and $n \in N$. Since $N \trianglelefteq G$, this always gives us back an element of N , and thus the action is well-defined. We can think of this action as a restriction of the action of G on itself by conjugation, and thus the two properties in the definition of an action hold for the action of G by conjugation on N . Therefore, this is indeed an action. The orbits of elements $n \in N$ under this action are the conjugacy classes $[n]_c$, and we have just shown that for all $n \in N$, $[n]_c \subseteq N$. But every element in N belongs to some conjugacy class, thus the conjugacy classes of the elements of N partition N . \square

Remark 5.28. Lemma 5.27 says that the orbits of the conjugation action of G on a normal subgroup N are just the orbits of the conjugation action of G on itself that contain elements of N (and must thus be completely contained in N). In contrast, if N is a normal subgroup of G , we can also consider the conjugation action of N on itself. If a and b are elements of N that are conjugate for the N -conjugation, then they must also be conjugate for the G -conjugation action, using the same element $n \in N$ such that $a = nbn^{-1}$. However, if a and b are conjugate for the G -conjugation, they might not necessarily be conjugate for the N -action, as all the elements $g \in G$ such that $a = gbg^{-1}$ could very well all be in $G \setminus N$.

We will see examples of this in the next section, where we will study the special case of the alternating group.

5.3 The alternating group

Since $A_n \leq S_n$, we know that *if* two elements of A_n are conjugate, *then* they have the same cycle type, as they are also conjugate elements of S_n , and thus we can apply our computation of conjugacy classes on in S_n . But as noted in Remark 5.28, there is no reason for the converse to hold: given $\alpha, \beta \in A_n$ of the same cycle type, the elements $\sigma \in S_n$ such that $\sigma\alpha\sigma^{-1} = \beta$ might all belong to $S_n \setminus A_n$. Indeed, we will see that this does happen in some cases.

Example 5.29. The two permutations (123) and (132) are not conjugates in A_3 , despite having the same cycle type and thus being conjugate in S_3 by our computation of conjugacy classes on in S_n . One can check this easily, for example, by conjugating (123) by the 3 elements in A_3 .

Lemma 5.30. Let σ be an m -cycle in S_n . Then

$$\sigma \in A_n \iff m \text{ is odd.}$$

Proof. Recall from earlier that,

$$(i_1 i_2 \cdots i_m) = (i_1 i_m)(i_1 i_{m-1})(i_1 i_3)(i_1 i_2)$$

is a product of $m - 1$ transpositions. Thus σ is even if and only if $m - 1$ is even. \square

Lemma 5.31 (Conjugacy classes of A_5). *The conjugacy classes of A_5 are given by the following list:*

- (1) *The singleton $\{e\}$ is a conjugacy class.*
- (2) *The conjugacy class of (12345) in A_5 has 12 elements.*
- (3) *The conjugacy class of (21345) in A_5 has 12 elements, and it is disjoint from the conjugacy class of (12345) .*
- (4) *The collection of all three cycles, of which there are 20, forms a conjugacy class in A_5 .*
- (5) *The collection of all products of two disjoint transpositions, of which there are 15, forms one conjugacy class in A_5 .*

As a reality check, note that $12 + 12 + 20 + 15 + 1 = 60 = |A_5|$.

Proof. By the work above, the cycle types of elements of A_5 are

- five cycles, of which there are $4! = 24$,
- three cycles, of which there are $\binom{5}{3}2 = 20$,
- products of two disjoint transpositions, of which there are $5 \cdot 3 = 15$, and
- the unique 1-cycle e , and indeed $[e]_c = \{e\}$.

By our computation of conjugacy classes on in S_n , we know that two permutations are conjugate in S_5 if and only if they have the same cycle type. It follows that the conjugacy classes in A_5 form a subset of the cycle types. The statement we are trying to prove asserts

that the set of five cycles breaks apart into two conjugacy classes in A_5 , whereas in all the other cases, the conjugacy classes remain whole.

Claim: Fix a 5-cycle σ . The conjugacy class of σ in A_5 has 12 elements.

By Lagrange's Theorem,

$$|C_{S_5}(\sigma)| = \frac{|S_5|}{[S_5 : C_{S_5}(\sigma)]}.$$

Thus,

$$[S_5 : C_{S_5}(\sigma)] = |[\sigma]_c|.$$

By our computation of conjugacy classes on in S_n , this is the number of 5-cycles in S_5 , which is $4!$. Thus

$$|C_{S_5}(\sigma)| = \frac{5!}{4!} = 5.$$

Since every power of σ commutes with σ , and there are 5 such elements, we conclude that

$$C_{S_5}(\sigma) = \{e, \sigma, \sigma^2, \sigma^3, \sigma^4\}.$$

But these are all in A_5 , and thus by an earlier lemma we conclude that

$$C_{A_5}(\sigma) = C_{S_5}(\sigma) \cap A_5 = \{e, \sigma, \sigma^2, \sigma^3, \sigma^4\}.$$

By the Orbit-Stabilizer Theorem and Lagrange's Theorem,

$$\text{the size of the conjugacy class of } \sigma \text{ in } A_5 = [A_5 : C_{A_5}(\sigma)] = \frac{|A_5|}{|C_{A_5}(\sigma)|} = \frac{60}{5} = 12.$$

This proves the claim.

We have now shown that the conjugacy class of each 5-cycle has 12 elements, and all twenty-four 5-cycles are in A_5 . Thus there are two conjugacy classes of 5-cycles in A_5 . This shows that σ is only conjugate in A_5 to half of the five cycles. If we pick two 5-cycles σ and τ that are not conjugate in A_5 , then τ is conjugate to exactly 12 elements, which must be exactly the other 5-cycles that σ is not conjugate to.

One can see that in fact (12345) and (21345) are not conjugate. While they *are* conjugate in S_5 , it is via the element (12) , which is *not* in A_5 . Suppose that $\alpha \in S_5$ is such that

$$\alpha(21345)\alpha^{-1} = (12345).$$

Note that $\tau = \alpha(12)$ satisfies

$$\begin{aligned} \tau(12345) &= \alpha(12)(12345) \\ &= \alpha(21345) \\ &= (21345)\alpha \\ &= (12345)(12)\alpha \\ &= (12345)\tau. \end{aligned}$$

Thus $\alpha(12) \in C_{S_5}(12345)$, or equivalently,

$$\alpha \in (12) \cdot C_{S_5}(21345).$$

But note that we just proved that every element in $C_{S_5}(2\ 1\ 3\ 4\ 5)$ is in A_5 , and thus even; this shows that every element in the coset

$$(1\ 2) \cdot C_{S_5}(2\ 1\ 3\ 4\ 5)$$

is odd (as we multiplied by *one* transposition), and thus there are no such α in A_5 . This proves (1) and (2).

Claim: All 20 three cycles are conjugate in A_5 .

Given two three cycles $(a\ b\ c)$ and $(d\ e\ f)$ in S_5 , we already know that they are both in A_5 and that there is a $\sigma \in S_5$ such that

$$\sigma(a\ b\ c)\sigma^{-1} = (d\ e\ f).$$

If $\sigma \notin A_5$, let $\{1, \dots, 5\} \setminus \{a, b, c\} = \{x, y\}$. Then σ is a product of an odd number of transpositions, so $\sigma \cdot (x\ y) \in A_5$. Moreover, since $(x\ y)$ and $(a\ b\ c)$ are disjoint cycles, so they must commute, so that

$$(x\ y)(a\ b\ c)(x\ y)^{-1} = (a\ b\ c).$$

Therefore,

$$(\sigma \cdot (x\ y))(a\ b\ c)(\sigma \cdot (x\ y))^{-1} = (d\ e\ f),$$

so $(a\ b\ c)$ and $(d\ e\ f)$ are still conjugate in S_5 . This proves the claim.

Claim: All products of two disjoint transpositions are conjugate in A_5 .

Set $\alpha = (1\ 2)(3\ 4)$. The conjugacy class of α in S_5 consists of all the products of two disjoint two-cycles, and there are 15 such elements. By the Orbit-Stabilizer Theorem,

$$15 = |\text{the conjugacy class of } \alpha \text{ in } S_5| = [S_5 : C_{S_5}(\alpha)] = \frac{120}{|C_{S_5}(\alpha)|}.$$

Thus

$$|C_{S_5}(\alpha)| = \frac{120}{15} = 8.$$

Since α commutes with e , α , $(1\ 3)(2\ 4)$ and $(1\ 4)(2\ 3)$ and each of these belongs to A_5 , we must have $|C_{A_5}(\alpha)| \geq 4$. Since

$$C_{A_5}(\alpha) = C_{S_5}(\alpha) \cap A_5,$$

it follows that $|C_{A_5}(\alpha)|$ must divide both 8 and 60, and so must be 1, 2 or 4. We conclude that $|C_{A_5}(\alpha)| = 4$. Thus α is conjugate in A_5 to $60/4 = 15$ elements. Since there are 15 products of disjoint two-cycles, they must all be conjugate to α , and thus the conjugacy class of α in A_5 is still the set of all 2-cycles. \square

Now that we have completely calculated all the conjugacy classes of A_5 , our hard work will pay off: we can now prove a very important result in group theory.

Definition 5.32. A nontrivial group G is **simple** if it has no proper nontrivial normal subgroups.

Exercise 33. Let p be prime. Show that \mathbb{Z}/p is a simple group.

Theorem 5.33. *The group A_5 is a simple group.*

Proof. Suppose $N \trianglelefteq A_5$. By Lagrange's Theorem, $|N|$ divides

$$|A_5| = \frac{5!}{2} = 60.$$

Then A_5 has only four nontrivial conjugacy classes, and they have order 12, 12, 15, and 20. Since N is normal, it is a union of conjugacy classes of A_5 . Thus

$$|N| = 1 + \text{the sum of a sublist of the list } 20, 12, 12, 15.$$

By checking the relatively small number of cases we see that $|N| = 1$ or $|N| = 60$ are the only possibilities, as the remaining options do not divide 60. \square

In fact, A_n is simple for all $n \geq 5$, but we will not prove this. In contrast, A_4 is not simple: we have seen in class that it has a normal subgroup with four elements consisting of the identity and the three products of two disjoint transpositions.

Example 5.34. The alternating group A_3 is simple and abelian since it has order 3.

Both A_1 and A_2 are the trivial group.

Thus the story goes:

Theorem 5.35. *Let $n \geq 3$. The alternating group A_n is simple if and only if $n \neq 4$.*

In fact, one can show that A_5 is the smallest nonabelian simple group, having 60 elements. This we will also not prove.

5.4 Other group actions with applications

Let's discuss a couple other group actions that often lead to useful information about the group doing the acting. The first one arises from the action of a group on the collection of left cosets of one of its subgroups. More precisely, let G be a group and H a subgroup, and let \mathcal{L} denote the collection of left cosets of H in G :

$$\mathcal{L} = \{xH \mid x \in G\}.$$

When H is normal, \mathcal{L} is the quotient group $\mathcal{L} = G/H$, but note that we are not assuming that H is normal. Then G acts on \mathcal{L} via the rule

$$g \cdot (xH) := (gx)H.$$

This action is transitive: for all x ,

$$xH = x \cdot (eH).$$

The stabilizer of the element $H \in \mathcal{L}$ is

$$\text{Stab}_G(H) = \{x \in G \mid xH = H\} = H,$$

which is consistent with the Orbit-Stabilizer Theorem, as indeed

$$\text{Orb}_G(H) = \mathcal{L}, \quad \text{so } |\text{Orb}(H)| = |\mathcal{L}| = [G : H],$$

while

$$\text{Stab}_G(H) = H, \quad \text{so } [G : \text{Stab}_G(H)] = [G : H].$$

As with any group action, this action induces a homomorphism

$$\rho: G \rightarrow \text{Perm}(\mathcal{L})$$

where for any g ,

$$\begin{aligned} \text{Perm}(\mathcal{L}) &\xrightarrow{\rho(g)} \text{Perm}(\mathcal{L}) \\ xH &\longmapsto (gx)H. \end{aligned}$$

If $n = [G : H] = |\text{Perm}(\mathcal{L})|$ is finite, then we have a homomorphism $\rho: G \rightarrow S_n$.

Lemma 5.36. *Let G be a group and H a subgroup of G . Consider the action of G on the set \mathcal{L} of left cosets of H , and the corresponding permutation representation $\rho: G \rightarrow \text{Perm}(\mathcal{L})$. Then*

$$\ker(\rho) = \bigcap_{x \in G} xHx^{-1}.$$

In particular, $\ker(\rho) \subseteq H$.

Note that $\bigcap_{x \in G} xHx^{-1}$ is the largest normal subgroup of G contained in H .

Proof. Note that

$$\begin{aligned} g \in \ker(\rho) &\iff (gx)H = xH \text{ for all } x \in G \\ &\iff x^{-1}gx \in H \text{ for all } x \in G \\ &\iff g \in xHx^{-1} \text{ for all } x \in G. \end{aligned}$$

Thus

$$\ker(\rho) = \bigcap_{x \in G} xHx^{-1}.$$

Since $eHe^{-1} = H$, we conclude that $\ker(\rho) \subseteq H$. \square

Remark 5.37. The action of G on the left cosets of H might be faithful or not. The Lemma above says that the action is faithful if and only if

$$\bigcap_{x \in G} xHx^{-1} = \{e\}.$$

If H is a normal subgroup of G , then in fact

$$\bigcap_{x \in G} xHx^{-1} = H,$$

and thus the action is not faithful unless $H = \{e\}$.

Remark 5.38. Consider the subgroup $H = \langle(12)\rangle$ of S_3 . The action of S_3 on the left cosets of H is faithful: for example, taking $\sigma = (13)$ we have

$$\sigma H\sigma^{-1} = \{e, (12)(13)\} = \{e, (23)\},$$

and thus the permutation representation $\rho: S_3 \rightarrow S_3$ associated with the action has

$$\ker \rho \subseteq \sigma H\sigma^{-1} \cap H = \{e\}.$$

Theorem 5.39. Let G be a finite group and H a subgroup of index p , where p is the smallest prime divisor of $|G|$. Then H is normal.

Proof. The action of G on the set of left cosets of H in G by left multiplication induces a homomorphism $\rho: G \rightarrow S_p$. By the Lemma above, its kernel $N := \ker(\rho)$ is contained in H . By the First Isomorphism Theorem,

$$[G : N] = |G/N| = |\text{im}(f)|.$$

By Lagrange's Theorem, since $\text{im}(f)$ is a subgroup of S_p then $[G : N] = |\text{im}(f)|$ divides $|S_p| = p!$. On the other hand, $[G : N]$ divides $|G|$ by Lagrange's Theorem. Since $[G : N]$ divides both $|G|$ and $p!$, it must divide $\gcd(|G|, p!)$. Since p is the smallest prime divisor of G , we must have

$$\gcd(|G|, p!) = p.$$

It follows that $[G : N]$ divides p , and hence $[G : N] = 1$ or $[G : N] = p$. But $N \subseteq H$, and H is a proper subgroup of G , so $N \neq G$, and thus $[G : N] \neq 1$. Therefore, we conclude that $[G : N] = p$. Since $N \subseteq H$ and $[G : H] = p = [G : N]$, we conclude that $H = N$. In particular, H must be a normal subgroup of G . \square

This generalizes an earlier exercise, which says that any subgroup of index 2 is normal. Another interesting action arises from the following: Let G be a group and let

$$\mathcal{S}(G) = \{H \mid H \leq G\}$$

be the collection of all subgroups of G . Then G acts on \mathcal{S} by

$$g \cdot H = gHg^{-1}.$$

Definition 5.40. Two subgroups A and B of a group G are **conjugate** if there exists $g \in G$ such that $A = gBg^{-1}$.

Equivalently, two subgroups are conjugate if they are in the same orbit by the following group action: the action of G on the set of its subgroups by conjugation.

Exercise 34. Let G be a group and let

$$\mathcal{S}(G) = \{H \mid H \leq G\}.$$

Check that the rule

$$g \cdot H = gHg^{-1}$$

defines an action of G on $\mathcal{S}(G)$. Moreover, prove that given any subgroup H of G , the stabilizer of H is given by $N_G(H)$.

The normalizer $N_G(H)$ is the largest subgroup of G that contains H as a normal subgroup, meaning that $H \trianglelefteq N_G(H)$.

Exercise 35. Let G be a group and H be a subgroup of G . Show that if K is any subgroup of G such that $H \trianglelefteq K$, then $K \leq N_G(H)$. In particular, $H \trianglelefteq G$ if and only if $N_G(H) = G$.

We can now show that the number of subgroups conjugate to a given subgroup is the index of its normalizer:

Lemma 5.41. *Let G be a group and H be a subgroup of G . The number of subgroups of G that are conjugate to H is equal to $[G : N_G(H)]$.*

Proof. The number of subgroups of G that are conjugate to H is just the size of the orbit of H under the action of G by conjugation on the set of subgroups of G . By the Orbit-Stabilizer Theorem, the number of elements in the orbit of H is the index of the stabilizer. Finally, by above, the stabilizer of H is $N_G(H)$. \square

Here is an application of this action:

Lemma 5.42. *If G is finite and H is a proper subgroup of G , then*

$$G \neq \bigcup_x xHx^{-1}.$$

Proof. First, suppose that H is a normal. Then $H = xHx^{-1}$ for all $x \in G$, so

$$\bigcup_x xHx^{-1} = H \neq G.$$

Now assume that H is not normal, so that $N_G(H) \neq G$ and $[G : N_G(H)] \geq 2$. By ??, we have $|H| = |xHx^{-1}|$ for all x . Since there are $[G : N_G(H)]$ conjugates of H (by an earlier Lemma), and since $e \in xHx^{-1}$ for all x , we get

$$\left| \bigcup_x xHx^{-1} \right| \leq [G : N_G(H)] \cdot |H|.$$

But in fact, this calculation can be improved, as there are at least two distinct conjugates of H and e is an element of all of them. This gives us

$$\left| \bigcup_x xHx^{-1} \right| \leq [G : N_G(H)] \cdot |H| - 1.$$

But $H \subseteq N_G(H)$ and so $[G : N_G(H)] \leq [G : H]$. We conclude that

$$\left| \bigcup_x xHx^{-1} \right| \leq [G : H] \cdot |H| - 1 = |G| - 1. \quad \square$$

Since $|H| = |xHx^{-1}|$ for all $x \in G$, we can fix a natural number n , set

$$\mathcal{S}_n(G) := \{H \mid H \leq G \text{ and } |H| = n\},$$

and consider the action of G on $\mathcal{S}_n(G)$ by conjugation. This idea will be exploited in the next section.

Exercise 36. Show that if G is a finite group acting transitively on a set S with at least two elements, then there exists $g \in G$ with no fixed points, meaning $g \cdot s \neq s$ for all $s \in S$.

Chapter 6

Sylow Theory

Sylow Theory is a very powerful technique for analyzing finite groups of relatively small order. One aspect of Sylow theory is that it allows us to deduce, in certain special cases, the existence of a unique subgroup of a given order, and thus it allows one to construct a normal subgroup.

6.1 Cauchy's Theorem

We start by proving a very powerful statement: that every finite group whose order is divisible by p must have an element of order p .

Theorem 6.1 (Cauchy's Theorem). *If G is a finite group and p is a prime number dividing $|G|$, then G has an element of order p . In fact, there are at least $p - 1$ elements of order p .*

Proof. Let S denote the set of ordered p -tuples of elements of G whose product is e :

$$S = \{(x_1, \dots, x_p) \mid x_i \in G \text{ and } x_1 x_2 \cdots x_p = e\}.$$

Consider

$$G^{p-1} := \underbrace{G \times \cdots \times G}_{p-1 \text{ factors}}$$

and the map

$$\begin{array}{ccc} G^{p-1} & \xrightarrow{\phi} & S \\ (x_1, \dots, x_{p-1}) & \longmapsto & (x_1, \dots, x_{p-1}, x_{p-1}^{-1} \cdots x_1^{-1}). \end{array}$$

Given the definition of S , the map ϕ does indeed land in S . Moreover, ϕ is bijective since the map $\psi: S \rightarrow G^{p-1}$ given by

$$\psi(x_1, \dots, x_p) = (x_1, \dots, x_{p-1})$$

is a two-sided inverse of the map above. Therefore, $|S| = |G^{p-1}| = |G|^{p-1}$.

Let C_p denote cyclic subgroup of S_p of order p generated by the p -cycle

$$\sigma = (1 \ 2 \ \cdots \ p).$$

The following rule gives an action of C_p on S :

$$\sigma^i \cdot (x_1, \dots, x_p) := (x_{\sigma^i(1)}, \dots, x_{\sigma^i(p)}) = (x_{1+i}, x_{2+i}, \dots, x_{p+i}),$$

where the indices are taken modulo p . We should check that this is indeed an action. On the one hand, σ^0 is the identity map, so

$$e \cdot (x_1, \dots, x_p) = \sigma^0 \cdot (x_1, \dots, x_p) = (x_{\sigma^0(1)}, \dots, x_{\sigma^0(p)}) = (x_1, \dots, x_p).$$

Moreover,

$$\sigma^i \cdot (\sigma^j \cdot (x_1, \dots, x_p)) = \sigma^i \cdot (x_{1+j}, x_{2+j}, \dots, x_{p+j}) = (x_{1+j+i}, x_{2+j+i}, \dots, x_{p+j+i}),$$

while

$$(\sigma^i \sigma^j) \cdot (x_1, \dots, x_p) = \sigma^{i+j} \cdot (x_1, \dots, x_p) = (x_{1+i+j}, x_{2+i+j}, \dots, x_{p+i+j}).$$

Thus

$$\sigma^i \cdot (\sigma^j \cdot (x_1, \dots, x_p)) = (\sigma^i \sigma^j) \cdot (x_1, \dots, x_p),$$

and we have shown that this is indeed an action.

Now let us consider the fixed points of this action. If

$$\sigma \cdot (x_1, \dots, x_p) = (x_1, \dots, x_p),$$

then $x_{i+1} = x_i$ for $1 \leq i \leq p$, so it follows that

$$x_1 = x_2 = \dots = x_p.$$

Thus if $\sigma \cdot (x_1, \dots, x_p) = (x_1, \dots, x_p)$, then (x_1, \dots, x_p) corresponds to an element x such that $x^p = x_1 \cdots x_p = e$. On the other hand, if σ fixes (x_1, \dots, x_p) , then so does any element of C_p . Therefore, a fixed point for this action corresponds to an element x such that $x^p = e$. The element (e, e, \dots, e) is a fixed point. Any other fixed point, meaning an orbit of size one, corresponds to an element of G order p , thus we wish to show that there is at least one fixed point besides (e, \dots, e) .

By the Orbit-Stabilizer Theorem, the size of every orbit divides $|C_p| = p$. Since p is prime, every orbit for this action has size 1 or p . By the Orbit Equation,

$$|S| = \# \text{ fixed points} + p \cdot \# \text{ orbits of size } p$$

Since p divides $|S|$, we conclude that p divides the number of fixed points. We already know that there is at least one fixed point, (e, \dots, e) . Thus there must be at least one other fixed point; in fact, at least $p - 1$ others, since the number of fixed points must then be at least p . \square

We now know that if p divides $|G|$, then G has an element of order p . However, this is not true if n divides $|G|$ but n is not prime. In fact, G may not even have any subgroup of order n .

Exercise 37. Prove that the converse to Lagrange's Theorem is false: find a group G and an integer $d > 0$ such that d divides the order of G but G does not have any subgroup of order d .

6.2 The Main Theorem of Sylow Theory

Definition 6.2. Let G be a finite group and p a prime. Write the order of G as $|G| = p^e m$ where $p \nmid m$. A **p -subgroup** of G is a subgroup of G of order p^k for some k . A **Sylow p -subgroup** of G is a subgroup $H \leq G$ such that $|H| = p^e$.

Thus a Sylow p -subgroup of G is a subgroup whose order is the highest conceivable power of p according to Lagrange's Theorem.

Definition 6.3. We will denote the collection of all Sylow p -subgroups of G by $\text{Syl}_p(G)$.

This is, of course, not very interesting unless $e > 0$. Nevertheless, we allow that case.

Remark 6.4. When p does not divide $|G|$, we have $e = 0$ and G has a unique Sylow p -subgroup, namely $\{e\}$, which indeed has order $p^0 = 1$.

Note that even if p does divide $|G|$, it is a priori possible that $n_p = 0$ for some groups G and primes p . We will prove this is not possible, and that is actually one of the hardest things to prove to establish Sylow theory.

Example 6.5. Let $p > 2$ be a prime and consider the group D_p . The subgroup $\langle r \rangle$ is a Sylow p -subgroup, as it has order p and $|D_p| = 2p$. In fact, this is the only Sylow p -subgroup of D_p , as by Exercise, every group of order p is cyclic, and the only elements of order p in D_p are r and its powers.

In D_n for n odd, each of the subgroups $\langle sr^j \rangle$, for $j = 0, \dots, n-1$ is a Sylow 2-subgroup. Since n is odd, only the reflections have order 2, and we have listed all the subgroups generated by reflections, so we conclude that the number of Sylow 2-subgroups is n .

Example 6.6. If G is cyclic of finite order, there is a unique Sylow p -subgroup for each p , since by the structure theorem for cyclic groups there is a unique subgroup of each order that divides $|G|$: if $G = \langle x \rangle$ and $|x| = p^e m$ with $p \nmid m$, then the unique Sylow p -subgroup of G is $\langle x^m \rangle$.

Let G be a finite group and p is a prime that divides $|G|$. Then G acts on its Sylow p -subgroups of G via conjugation. As of now, for all we know, this might be the action on the empty set. Sylow Theory is all about understanding this action very well. Before we can prove the main theorem, we need a technical lemma.

Lemma 6.7. Let G be a finite group, p a prime, P a Sylow p -subgroup of G , and Q any p -subgroup of G . Then $Q \cap N_G(P) = Q \cap P$.

Proof. (\subseteq) Since $P \leq N_G(P)$, then $Q \cap P \leq Q \cap N_G(P)$.

(\supseteq) Let $H := Q \cap N_G(P)$. Since $H \subseteq N_G(P)$, then $PH = HP$, so we get that PH is a subgroup of G . By the Diamond Isomorphism Theorem, we have

$$|PH| = \frac{|P| \cdot |H|}{|P \cap H|}$$

and since each of $|P|$, $|H|$, and $|P \cap H|$ is a power of p , we conclude that the order of PH is also a power of p . In particular, PH is a p -subgroup of G . On the other hand, $P \leq PH$ and P is already a p -subgroup of the largest possible order, so we must have $P = PH$. Note that $H \leq PH$ always holds. We conclude that $H \leq P$ and thus $H \leq Q \cap P$. \square

Theorem 6.8 (Main Theorem of Sylow Theory). *Let p be prime. Assume G is a group of order $p^e m$, where p is prime, $e \geq 0$, and $\gcd(p, m) = 1$.*

- (1) *There exists at least one Sylow p -subgroup of G . In short, $\text{Syl}_p(G) \neq \emptyset$.*
- (2) *If P is a Sylow p -subgroup of G and $Q \leq G$ is any p -subgroup of G , then $Q \leq gPg^{-1}$ for some $g \in G$. Moreover, any two Sylow p -subgroups are conjugate and the action of G on $\text{Syl}_p(G)$ by conjugation is transitive.*
- (3) *We have*

$$|\text{Syl}_p(G)| \equiv 1 \pmod{p}.$$

- (4) *For any $P \in \text{Syl}_p(G)$,*

$$|\text{Syl}_p(G)| = [G : N_G(P)],$$

and hence

$$|\text{Syl}_p(G)| \text{ divides } m.$$

Proof. First we will prove G contains a subgroup of order p^e by induction on $|G| = p^e m$.

When $|G| = 1$, $\{e\}$ is a Sylow p -subgroup by our convention. In fact, this argument applies for whenever $e = 0$, so we may thus assume through the rest of the proof that p does divide $|G|$. So suppose that p divides $|G|$ and every group of order $n < |G|$ has a Sylow p -subgroup. We will consider two cases, depending on whether p divides $|\text{Z}(G)|$.

If p divides $|\text{Z}(G)|$, then by Cauchy's Theorem there is an element $z \in \text{Z}(G)$ of order p . Set $N := \langle z \rangle$. Since $z \in \text{Z}(G)$, then for all $g \in G$ we have

$$gz^i g^{-1} = z^i \in N,$$

and thus $N \trianglelefteq G$. Since

$$|G/N| = \frac{|G|}{|N|} = \frac{p^e m}{p} = p^{e-1} m,$$

by induction hypothesis G/N has a subgroup of order p^{e-1} , which must then have index m . By the Lattice Isomorphism Theorem, this subgroup corresponds to a subgroup of G of index m , hence of order p^e .

Now assume p does not divide $|\text{Z}(G)|$, and consider the class equation for G : g_1, \dots, g_k are a complete list of noncentral conjugacy class representatives, without repetition of any class, we have

$$|G| = |\text{Z}(G)| + \sum_{i=1}^k [G : C_G(g_i)].$$

Suppose that p divides $[G : C_G(g_i)]$ for all i . Since p also divides $|G|$, then this would imply that p divides $|\text{Z}(G)|$, but we assumed that p does not divide $|\text{Z}(G)|$. We conclude that p does not divide $[G : C_G(g_i)]$ for some i .

Note that $[G : C_G(g_i)]$ divides $|G|$ by Lagrange's Theorem, and thus it must divide m . Set

$$d := \frac{m}{[G : C_G(g_i)]}.$$

Then

$$|C_G(g_i)| = \frac{|G|}{[G : C_G(g_i)]} = \frac{p^e m}{[G : C_G(g_i)]} = p^e d,$$

and note that p does not divide d since it does not divide m . Since g_i is not central, then $e \notin C_G(g_i)$, and in particular $|C_G(g_i)| < |G|$. By induction hypothesis, $C_G(g_i)$ contains a subgroup S of order p^e . But S is also a subgroup of G , and it has order p^e , as desired. This completes the proof of (1): we have shown that G contains a subgroup of order p^e .

To prove (2) and (3), let P be a Sylow p -subgroup and let Q be any p -subgroup. Let \mathcal{S}_P denote the collection of all conjugates of P :

$$\mathcal{S}_P := \{gPg^{-1} \mid g \in G\}.$$

By definition, G acts transitively on \mathcal{S}_P by conjugation. Restricting that action to Q , we get an action of Q on \mathcal{S}_P , though note that we do not know if that action is transitive. The key to proving parts (2) and (3) of the Sylow Theorem is to analyze the action of Q on \mathcal{S}_P .

Let $\mathcal{O}_1, \dots, \mathcal{O}_s$ be the distinct orbits of the action of Q on \mathcal{S}_P , and for each i pick a representative $P_i \in \mathcal{O}_i$. Note that

$$\begin{aligned} \text{Stab}_Q(P_i) &= \{q \in Q \mid qP_iq^{-1} = P_i\} && \text{by the definition of the action} \\ &= N_Q(P_i) && \text{by definition of normalizer} \\ &= Q \cap N_G(P_i) \\ &= Q \cap P_i && \text{by the Sylow normalizer Lemma.} \end{aligned}$$

By the Orbit-Stabilizer Theorem, we have $|\mathcal{O}_i| = [Q : Q \cap P_i]$, and thus, collecting the orbits,

$$|\mathcal{S}_P| = \sum_{i=1}^s [Q : Q \cap P_i]. \quad (6.2.1)$$

This equation 6.2.1 holds for any p -subgroup Q of G . In particular, we can take $Q = P_1$. In this case, the first term in the sum is $[Q : Q \cap P_1] = 1$ and, for all $i \neq 1$ we have

$$Q \cap P_i = P_1 \cap P_i \neq P_1 = Q \implies [Q : Q \cap P_i] > 1.$$

But $|Q|$ is a power of p , so $[Q : Q \cap P_i]$ must be divisible by p for all i . We conclude that

$$|\mathcal{S}_P| \equiv 1 \pmod{p}. \quad (6.2.2)$$

Note, however, that this does not yet prove part (3), since we do not yet know that \mathcal{S}_P consists of *all* the Sylow p -subgroups. But we do have all the pieces we need to prove part (2). Suppose, by way of contradiction, that Q is a p -subgroup of G that is not contained in any of the subgroups in \mathcal{S}_P . Then $Q \cap P_i \neq Q$ for all i , and thus every term on the right-hand side of

$$|\mathcal{S}_P| = \sum_{i=1}^s [Q : Q \cap P_i]$$

is divisible by p , contrary to (6.2.2). We conclude that Q must be contained in at least one of the subgroups in \mathcal{S}_P . This proves the first part of (2).

Moreover, if we take Q to be a Sylow p -subgroup of G , then $Q \leq gPg^{-1}$ for some g , but Q and P are both Sylow p -subgroups of G , so

$$|Q| = |P| = |gPg^{-1}|.$$

We conclude that $Q = gPg^{-1}$ is conjugate to P . In particular, the conjugation action of G on $\text{Syl}_p(G)$ is transitive, and this finishes the proof of (2).

This proves, in particular, that \mathcal{S}_P in fact does consist of all Sylow p -subgroups, we can now also conclude part (3) from (6.2.2).

Finally, for any $P \in \text{Syl}_p(G)$, the stabilizer of P for the action of G on $\text{Syl}_p(G)$ by conjugation is $N_G(P)$. Since we now know the action is transitive, the Orbit-Stabilizer Theorem says that

$$|\text{Syl}_p(G)| = [G : N_G(P)].$$

Moreover, since $P \leq N_G(P)$ and $|P| = p^e$, it follows that p divides $|N_G(P)|$, so

$$|N_G(P)| = p^e d$$

for some d that divides m . We conclude that

$$[G : N_G(P)] = \frac{|G|}{|N_G(P)|} = \frac{p^e m}{p^e d} = \frac{m}{d},$$

so $[G : N_G(P)]$ divides m . □

Remark 6.9. In general, Cauchy's Theorem can be deduced from part one of the Sylow Theorem. However, we used Cauchy's Theorem to prove the Sylow Theorem, so it is important to see that Cauchy's Theorem can be proven independently of Sylow theory.

To see how Cauchy's Theorem follows from the Sylow Theorem, suppose that the prime p divides $|G|$. Then by the Sylow Theorem there exists a Sylow p -subgroup P of G . Pick any nontrivial element $x \in P$. Then $|x| = p^j$ for some $j \geq 1$, since by Lagrange's Theorem $|x|$ must divide $|P| = p^e$. Then $y = x^{p^{j-1}}$ has order p :

$$y^p = (x^{p^{j-1}})^p = x^{p \cdot p^{j-1}} = x^{p^j} = e,$$

Moreover, $y^i \neq e$ for $2 \leq i < p$, as otherwise

$$|x| \leq ip^{j-1} < p^j.$$

Remark 6.10. Let G be a group. We saw that if H is the unique subgroup of finite order n , then H must be a normal subgroup of G . One consequence of the Main Theorem of Sylow Theory is a sort of converse to this: if G has multiple Sylow p -subgroups, then G has no normal Sylow p -subgroups, since any two Sylow p -subgroups must be conjugate to each other.

6.3 Using Sylow Theory

Using the Main Theorem of Sylow Theory, we can often find the exact number of Sylow p -subgroups, sometimes leading us to find normal subgroups. In particular, these techniques can be used to show that there are no normal subgroups of a particular order, as the next example will illustrate.

Example 6.11 (No simple groups of order 12). Let us prove that there are no simple groups of order 12. To do that, let G be any group of order $12 = 2^2 \cdot 3$. We will prove that G must have either a normal subgroup of order 3 or a normal subgroup of order 4.

First, consider $n_2 = |\text{Syl}_2(G)|$. By the Main Theorem of Sylow Theory, $n_2 \equiv 1 \pmod{2}$ and n_2 divides 3. This gives us $n_2 \in \{1, 3\}$. Similarly, $n_3 = |\text{Syl}_3(G)|$ satisfies

$$n_3 \equiv 1 \pmod{3} \quad \text{and} \quad n_3 \mid 4,$$

so $n_3 \in \{1, 4\}$. If either of these numbers is 1, we have a unique subgroup of order 4 or of order 3, and such a subgroup must be normal.

Suppose that $n_3 \neq 1$, which leaves us with $n_3 = 4$. Let P_1, P_2, P_3 , and P_4 be the Sylow 3-subgroups of G . Consider any $i \neq j$. Since $P_i \cap P_j$ is a subgroup of P_i , its order must divide 3. On the other hand, P_i and P_j are distinct groups of order 3, so $|P_i \cap P_j| < 3$, and we conclude that $|P_i \cap P_j| = 1$. Therefore, $P_i \cap P_j = \{e\}$ for all $i \neq j$. Thus the set

$$T := \bigcup_{i=1}^4 P_i$$

has 9 elements: the identity e and 8 other distinct elements. Since each P_i has order 3, those 8 elements must all have order 3. Note, moreover, that any other potential element of order 3 would generate its own Sylow 3-subgroup, so this is a complete count of all the elements of order 3. We conclude that there are 8 elements of order 3 in G .

In particular, there are 9 elements in G that are either the identity or have order 3, and thus there are only $12 - 9 = 3$ elements in G of order not 3, say a, b, c .

Now consider any Sylow 2-subgroup Q , which has 4 elements. None of its elements has order 3, so we must have $Q = \{e, a, b, c\}$. In particular, this shows that there is a unique Sylow 2-subgroup, which must then be normal.

Remark 6.12 (Warning!). In Example 6.11, it would not be so easy to count the elements of order 2 and 4. We do know that every element in

$$S := \bigcup_i Q_i$$

has order 1, 2, or 4, but the size of this set is harder to calculate. The issue is that $Q_i \cap Q_j$ might have order 2 for distinct i and j . The best we can say for sure is that S has at least $4 + 4 - 2 = 6$ elements.

More generally, if P and Q are both subgroups of G of prime order p , we can say that $P \cap Q = \{e\}$ using the same argument we employed in Example 6.11. However, if P and Q are two subgroups of order p^e with $e \geq 2$, we can no longer guarantee that $P \cap Q = \{e\}$.

Example 6.13 (No simple groups of order 80). Let G be a group of order $80 = 5 \cdot 16$, and let $n_2 = |\text{Syl}_2(G)|$ and $n_5 = |\text{Syl}_5(G)|$. By the Main Theorem of Sylow Theory,

$$n_2 \equiv 1 \pmod{2} \quad \text{and} \quad n_2 \mid 5 \implies n_2 \in \{1, 5\}$$

and

$$n_5 \equiv 1 \pmod{5} \quad \text{and} \quad n_5 \mid 16 \implies n_5 \in \{1, 16\}.$$

If either $n_2 = 1$ or $n_5 = 1$, then the unique Sylow 2-subgroup or 5-subgroup would be normal. If G is a simple group, then we must have

$$n_2 = 5 \quad \text{and} \quad n_5 = 16.$$

While the counting trick we used in Example 6.11 would work, let us try on a different tactic here.

Consider the action of G on $\text{Syl}_2(G)$ by conjugation, and let

$$\rho: G \rightarrow S_5$$

be the associated permutation representation. The action is transitive by the Main Theorem of Sylow Theory, so the map ρ is nontrivial. By Lemma 3.8, $\text{im}(\rho)$ is a subgroup of S_5 , and thus by Lagrange's Theorem the order of $\text{im}(\rho)$ divides $|S_5|$. However, $|G| = 80$ does not divide $120 = |S_5|$, so the image of ρ cannot have 80 elements, and in particular ρ cannot be injective. It follows that $\ker(\rho)$ is a nontrivial, proper normal subgroup of G , a contradiction.

Chapter 7

Products and finitely generated abelian groups

In this chapter, we will discuss how to build new groups from old ones, and completely classify all finitely generated abelian groups.

7.1 Direct products of groups

Definition 7.1. Let I be a set and consider a group G_i for each $i \in I$. The **direct product** of the groups $\{G_i\}_{i \in I}$, denoted by

$$\prod_{i \in I} G_i,$$

is the group with underlying set the Cartesian product

$$\prod_{i \in I} G_i$$

equipped with the operation defined by

$$(g_i)_{i \in I} (h_i)_{i \in I} = (g_i h_i)_{i \in I}.$$

The **direct sum** of the groups G_i is the subgroup of the direct product of $\{G_i\}_{i \in I}$ given by

$$\bigoplus_{i \in I} G_i := \{(g_i)_{i \in I} \in \prod_{i \in I} G_i \mid g_i = e_{G_i} \text{ for all but finitely many } i \in I\}.$$

In particular, the direct sum of $\{G_i\}_{i \in I}$ has the same operation as the direct product.

When I is finite, say $I = \{1, \dots, n\}$, we write

$$G_1 \times \cdots \times G_n := \prod_{i=1}^n G_i.$$

Remark 7.2. When I is finite, the direct sum and the direct product of $\{G_i\}_{i \in I}$ coincide. This is the case we will be most interested in.

Exercise 38. The direct product of a collection of groups is a group, and the direct sum is a subgroup of the direct product.

Remark 7.3. If G_1, \dots, G_n are all finite groups, then

$$|G_1 \times \cdots \times G_n| = |G_1| \cdots |G_n|.$$

Exercise 39. Let $\{G_i\}_{i \in I}$ be a collection of abelian groups. Show that

$$\prod_{i \in I} G_i$$

is an abelian group.

Exercise 40. Let G and H be groups, and $g \in G$ and $h \in H$.

- (a) Show that if $|g|$ and $|h|$ are both finite, then $|(g, h)| = \text{lcm}(|g|, |h|)$.
- (b) Show that if at least one of g or h has infinite order, then (g, h) also has infinite order.

Lemma 7.4 (CRT). *If $\gcd(m, n) = 1$, then $\mathbb{Z}/m \times \mathbb{Z}/n \cong \mathbb{Z}/mn$.*

Proof. By Exercise 40,

$$|(1, 1)| = \text{lcm}(m, n) = mn.$$

But $\mathbb{Z}/m \times \mathbb{Z}/n \cong \mathbb{Z}/mn$ has order mn , so $(1, 1)$ is a generator for the group, which must then be cyclic. By Theorem 3.41, all cyclic groups of order mn are isomorphic to \mathbb{Z}/mn , so

$$\mathbb{Z}/m \times \mathbb{Z}/n \cong \mathbb{Z}/mn. \quad \square$$

Exercise 41. Show that the converse holds: for all integers $m, n > 1$, if

$$\mathbb{Z}/m \times \mathbb{Z}/n \cong \mathbb{Z}/mn,$$

then $\gcd(m, n) = 1$.

Sometimes it is convenient to write the **CRT** in terms of prime factorization, as follows:

Theorem 7.5 (CRT). *Suppose $m = p_1^{e_1} \cdots p_l^{e_l}$ for distinct primes p_1, \dots, p_l . Then there is an isomorphism*

$$\mathbb{Z}/m \cong \mathbb{Z}/(p_1^{e_1}) \times \cdots \times \mathbb{Z}/(p_l^{e_l}).$$

Recall that we saw in Exercise 24 that given a group G and subgroups H and K , if H is normal then HK is a subgroup of G . In fact, we can say more:

Theorem 7.6 (Recognition theorem for direct products). *Suppose G is a group with normal subgroups $H \trianglelefteq G$ and $K \trianglelefteq G$ such that $H \cap K = \{e\}$. Then $HK \cong H \times K$ via the isomorphism $\theta: H \times K \rightarrow HK$ given by*

$$\theta(h, k) = hk.$$

Moreover,

$$H \cong \{(h, e) \mid h \in H\} \leq H \times K$$

and

$$K \cong \{(e, k) \mid k \in K\} \leq H \times K.$$

Proof. By Exercise 24, the hypothesis implies $HK \leq G$. Moreover, consider any $h \in H$ and any $k \in K$. Since H is a normal subgroup,

$$khk^{-1} \in H, \text{ say}$$

so also

$$[k, h] = khk^{-1}h^{-1} \in H.$$

But K is also a normal subgroup, so similarly we obtain

$$[k, h] \in K.$$

Therefore,

$$[k, h] \in H \cap K = \{e\},$$

so $[k, h] = e$. We conclude that

$$hk = kh \quad \text{for all } h \in H, k \in K.$$

The function θ defined above must then satisfy

$$\begin{aligned} \theta((h_1, k_1)(h_2, k_2)) &= \theta(h_1h_2, k_1k_2) \\ &= (h_1h_2)(k_1k_2) \quad \text{by definition of } \theta \\ &= h_1(h_2k_1)k_2 \\ &= (h_1k_1)(h_2k_2) \quad \text{since } h_2k_1 = k_1h_2 \\ &= \theta(h_1, k_1)\theta(h_2, k_2) \quad \text{by definition of } \theta \end{aligned}$$

and thus θ is a homomorphism. Its kernel is

$$\ker(\theta) = \{(k, h) \mid k = h^{-1}\} = \{(e, e)\}$$

since $H \cap K = \{e\}$. Moreover, θ is surjective, as any element in HK is of the form $hk \in HK$, and

$$\theta(h, k) = hk.$$

This proves θ is an isomorphism. Finally, restricting the codomain to any subgroup L of G and the domain to $\theta^{-1}(L)$ gives an isomorphism between L and $\theta^{-1}(L)$, so in particular

$$H \cong \theta^{-1}(H) = \{(h, e) \mid h \in H\} \leq H \times K$$

and

$$K \cong \theta^{-1}(K) = \{(e, k) \mid k \in K\} \leq H \times K. \quad \square$$

Remark 7.7. If $H \trianglelefteq G$ and $K \trianglelefteq G$ are such that $H \cap K = \{e\}$, then each elements of HK is uniquely of the form hk . This is a consequence of the fact that the map θ is a bijection.

Definition 7.8. Let G be a group. If $H \trianglelefteq G$ and $K \trianglelefteq G$ are such that $H \cap K = \{e\}$, then the subgroup HK of G is called the **internal direct product** of H and K , while the group $H \times K$ is called the **external direct product** of H and K .

Example 7.9. Let $G = D_n$, $H = \langle r \rangle$ and $K = \langle s \rangle$. Then $H \cap K = \{e\}$, $HK = G$, and $H \trianglelefteq G$, but K is not normal in G . So Theorem 7.6 does not apply to say that G is isomorphic to $H \times K$. In fact, G is not isomorphic to $H \times K$, since $H \times K$ is abelian, while G is not. As we shall see, G is the semidirect product of H and K .

7.2 Semidirect products

Remark 7.10. Let G be a group. Suppose we are given subgroups $H \trianglelefteq G$ and $K \leq G$ such that $H \cap K = \{e\}$ but K is not normal. Then we still have $HK \leq G$, but it is not necessarily true that the map $\theta : H \times K \rightarrow HK$ defined by $\theta(h, k) = hk$ is a group homomorphism. The issue is that given $h \in H$ and $k \in K$, while

$$khk^{-1} \in H \implies kh = h'k \text{ for some } h' \in H,$$

we can no longer guarantee that $kh = hk$. So given $h_1, h_2 \in H$ and $k_1, k_2 \in K$, suppose that $k_1h_1 = h'_2k_1$. For θ to be a homomorphism, we would need the following:

$$\theta(h_1, k_1)\theta(h_2, k_2) = (h_1k_1)(h_2k_2) = h_1h'_2k_1k_2 = \theta(h_1h'_2, k_1k_2).$$

This we would need

$$(h_1, k_1)(h_2, k_2) = (h_1h'_2, k_1k_2).$$

This motivates the following definition:

Definition 7.11. Let H and K be groups and let $\rho : K \rightarrow \text{Aut}(H)$ be a homomorphism. The (external) **semidirect product** induced by ρ is the set $H \times K$ equipped with the binary operation defined by

$$(h_1, k_1)(h_2, k_2) := (h_1\rho(k_1)(h_2), k_1k_2).$$

This group is denoted by $H \rtimes_{\rho} K$.

The underlying set of $H \rtimes_{\rho} K$ is the same as the direct product, but it is the operation that differs.

Remark 7.12. Note in particular that if H and K are finite, then $|H \rtimes_{\rho} K| = |H| \cdot |K|$.

The proof that the semidirect product is indeed a group is straightforward but a bit messy, as we need to check all the group axioms.

Theorem 7.13. If H and K are groups and $\rho : K \rightarrow \text{Aut}(H)$ is a homomorphism, then $H \rtimes_{\rho} K$ is a group.

Proof. First, we show that the operation is associative. Indeed,

$$\begin{aligned} (y_1, x_1)((y_2, x_2)(y_3, x_3)) &= (y_1, x_1)(y_2\rho(x_2)(y_3), x_2x_3) \\ &= (y_1\rho(x_1)(y_2\rho(x_2)(y_3)), x_1x_2x_3) \\ &= (y_1\rho(x_1)(y_2)(\rho(x_1) \circ \rho(x_2))(y_3), x_1x_2x_3) \\ &= (y_1\rho(x_1)(y_2)\rho(x_1x_2)(y_3), x_1x_2x_3) \\ &= (y_1\rho(x_1)(y_2), x_1x_2)(y_3, x_3) \\ &= ((y_1, x_1)(y_2, x_2))(y_3, x_3). \end{aligned}$$

To show that (e, e) is a two-sided identity, consider any $h \in H$ and $k \in K$. Since $\rho(k)$ is a homomorphism, then $\rho(k)(e) = e$, and thus

$$(h, k)(e, e) = (h\rho(k)(e), ke) = (he, ke) = (h, k).$$

Moreover, since ρ is a homomorphism, $\rho(e) = \text{id}_H$, and thus $\rho(e)(y) = \text{id}_H(y) = y$ for any $y \in K$, so that

$$(e, e)(h, k) = (e\rho(e)(h), ek) = (eh, ek) = (h, k).$$

Finally, for any $x \in H$ and $y \in K$ we have

$$\begin{aligned} (x, y)(\rho(y^{-1})(x^{-1}), y^{-1}) &= (x\rho(y)(\rho(y^{-1})(x^{-1})), yy^{-1}) \\ &= (x(\rho(y) \circ \rho(y^{-1}))(x^{-1}), e) \\ &= (x\rho(e)(x^{-1}), e) \quad \text{since } \rho \text{ is a homomorphism} \\ &= (xx^{-1}, e) \quad \text{since } \rho(e) = \text{id}_H \\ &= (e, e), \end{aligned}$$

and similarly,

$$\begin{aligned} (\rho(y^{-1})(x^{-1}), y^{-1})(x, y) &= (\rho(y^{-1})(x^{-1})\rho(y^{-1})(x), y^{-1}y) \\ &= (\rho(y^{-1})(x^{-1}x), e) \quad \text{since } \rho(y^{-1}) \text{ is a homomorphism} \\ &= (\rho(y^{-1})(e), e) \\ &= (e, e) \quad \text{since } \rho(y^{-1}) \text{ is a homomorphism}. \end{aligned}$$

Thus (x, y) has an inverse, given by

$$(x, y)^{-1} = (\rho(x^{-1})(y^{-1}), x^{-1}).$$

This completes the proof that the semidirect product is a group. \square

Example 7.14. Given any two groups H and K , we can always take ρ to be the trivial homomorphism. In that case, $H \rtimes_{\rho} K$ is just the usual direct product: for all $h \in H$ and all $k \in K$, $\rho(k) = \text{id}_H$, so

$$(h, k)(h', k') = (h\rho(k)(h'), kk') = (hh', kk').$$

Theorem 7.15. Given groups H and K are groups and a homomorphism $\rho: K \rightarrow \text{Aut}(H)$, H and K are isomorphic to subgroups of $H \rtimes_{\rho} K$, as follows:

$$H \cong \{(h, e) \mid h \in H\} \trianglelefteq H \rtimes_{\rho} K \text{ and } K \cong \{(e, k) \mid k \in K\} \leq H \rtimes_{\rho} K.$$

Moreover,

$$\frac{(H \rtimes_{\rho} K)}{\{(h, e) \mid h \in H\}} \cong K.$$

Proof. Consider the function $i: H \rightarrow H \rtimes_{\rho} K$ given by

$$i(y) = (y, e).$$

Then i is a homomorphism:

$$i(y_1)i(y_2) = (y_1, e)(y_2, e) = (y_1\rho(e)(y_2), ee) = (y_1y_2, e) = i(y_1y_2).$$

Moreover, i is injective by construction, and hence its image is isomorphic to H by the [First Isomorphism Theorem](#). We can describe $\text{im}(i)$ as the set of all elements whose second component is e . The image $\text{im}(i)$ is normal since the second component of

$$(h, k)(a, e)(h, k)^{-1} = (h, k)(a, e)(\rho(k^{-1})(h^{-1}), h^{-1})$$

is

$$kek^{-1} = e,$$

which shows that any for any $(a, e) \in \text{im}(H)$ and any $(h, k) \in H \rtimes_{\rho} K$,

$$(h, k)(a, e)(h, k)^{-1} \in \text{im}(i).$$

Let us write the image of i , which we now know is a normal subgroup of $H \rtimes_{\rho} K$, as

$$H' := \text{im}(i) = \{(y, e) \mid y \in H\} \trianglelefteq H \rtimes_{\rho} K.$$

Similarly, the function

$$j: K \rightarrow H \rtimes_{\rho} K \quad \text{given by } j(x) = (e, x)$$

is also an injective homomorphism (exercise!), and thus its image

$$K' := \{(e, x) \mid x \in H\} \leq H \rtimes_{\rho} K$$

is isomorphic to K . Finally, given any $(h, k) \in H \rtimes_{\rho} K$, we can write

$$(h, k) = (h\rho(e)(e), k) = (h, e)(e, k) \in H'K',$$

so $H'K' = H \rtimes_{\rho} K$.

Consider the projection onto the second factor

$$\pi_2: H \rtimes_{\rho} K \rightarrow K,$$

which is the map given by

$$\pi_2(x, y) = y.$$

This is a group homomorphism, since the second component of $(x_1, y_1)(x_2, y_2)$ is y_1y_2 , and thus

$$\pi_2((x_1, y_1)(x_2, y_2)) = y_1y_2 = \pi_2(y_1)\pi_2(y_2).$$

Moreover, π_2 is surjective by definition. Finally,

$$\ker(\pi_2) = \{(y, e_K) \mid y \in H\} = H' \cong H.$$

By the [First Isomorphism Theorem](#), we conclude that

$$(H \rtimes_{\rho} K)/H' \cong K. \quad \square$$

In Theorem 7.15, we showed that $\{(h, e) \mid h \in H\}$ is a normal subgroup of $H \rtimes_{\rho} K$. However, $\{(e, k) \mid k \in K\}$ is typically *not* a normal subgroup of $H \rtimes_{\rho} K$. We will see a concrete example of this below in Example 7.22.

Studying semidirect products is a great motivation to studying automorphism groups.

Exercise 42. Let C_n denote the cyclic group of order $n \geq 2$, and consider the group

$$(\mathbb{Z}/n)^{\times} = \{[j]_n \mid \gcd(j, n) = 1\}$$

with the binary operation given by the usual multiplication. Prove that

$$\text{Aut}(C_n) \cong (\mathbb{Z}/n)^{\times}.$$

Remark 7.16. We can now count the number of elements in $\text{Aut}(C_n)$, since it is the number of integers $1 \leq i < n$ that are coprime with n . This number is given by what is known as the **Euler φ function**,

$$\varphi(n) = n \prod_{p|n} \left(1 - \frac{1}{p}\right).$$

Equivalently, if $n = p_1^{a_1} \cdots p_k^{a_k}$, where p_1, \dots, p_k are distinct primes and each $a_i \geq 1$, then

$$\varphi(n) = \prod_{i=1}^k (p_i^{a_i-1}(p_i - 1)).$$

In particular, if p is prime then $|\text{Aut}(\mathbb{Z}/p)| = p - 1$.

The next fact is very useful, but we will hold off until next semester to prove it. For now, we record this fact so we can use it to construct nonabelian groups of a given order.

Exercise 43. If p is prime, then $\text{Aut}(C_p) \cong \mathbb{Z}/p^{\times}$ is cyclic of order $p - 1$.

Exercise 44. Let p be a prime integer. Show that

$$\text{Aut}(\underbrace{\mathbb{Z}/p \times \cdots \times \mathbb{Z}/p}_{n \text{ factors}}) \cong \text{GL}_n(\mathbb{Z}/p)$$

and that these groups have order $(p^n - 1)(p^n - p)(p^n - p^2) \cdots (p^n - p^{n-1})$.

To better understand semidirect products, we should also better understand what it means to have a homomorphism $K \rightarrow \text{Aut}(H)$.

Definition 7.17. Let G and H be groups. A (left) **action of G on H via automorphisms** is a pairing $G \times H \rightarrow H$, written as $(g, h) \mapsto g \cdot h$, such that

- For all $g_1, g_2 \in G$ and $h \in H$, $g_1 \cdot (g_2 \cdot h) = (g_1 \cdot_G g_2) \cdot h$.
- For all $h \in H$, $e_G \cdot h = h$.
- For all $g \in G$ and all $h_1, h_2 \in H$, $g \cdot (h_1 \cdot_H h_2) = (g \cdot h_1) \cdot_H (g \cdot h_2)$.

Remark 7.18. Note that the first two axioms are just the axioms for a group action. So given a group action of G on H , let $\rho: G \rightarrow \text{Perm}(H)$ be the corresponding permutation representation. If the action satisfies the third axiom in Definition 7.17, then that means that for each $g \in G$, $\rho(g)$ satisfies

$$\rho(g)(h_1 \cdot_H h_2) = \rho(g)(h_1) \rho(g)(h_2).$$

This condition simply says that $\rho(g)$ must be a homomorphism. Since $\rho(g)$ is already a bijection, we conclude that $\rho(g)$ must be an automorphism of H . Conversely, given any homomorphism $\rho: K \rightarrow \text{Aut}(H)$, we can define a group action of K on H via automorphisms by setting

$$k \cdot h := \rho(k)(h).$$

Since $\text{Aut}(H) \subseteq \text{Perm}(H)$, we can extend ρ to a homomorphism $K \rightarrow \text{Perm}(H)$, which we saw in Lemma 2.3 is equivalent to the action of K on H we just defined. That action satisfies

$$\begin{aligned} k \cdot (h_1 \cdot_H h_2) &= \rho(k)(h_1 \cdot_H h_2) \\ &= \rho(k)(h_1) \cdot_H \rho(k)(h_2) \quad \text{since } \rho \text{ is a homomorphism} \\ &= (k \cdot h_1) \cdot_H (k \cdot h_2) \end{aligned}$$

In conclusion, we can now say that to give an action of G on H via automorphisms is to give a group homomorphism

$$\rho: G \rightarrow \text{Aut}(H).$$

Moreover, given a group K acting on a group H by automorphisms, we get an induced semidirect product $H \rtimes_{\rho} K$, where $\rho: K \rightarrow \text{Aut}(H)$ is the corresponding homomorphism.

Here is an important example of an action by automorphisms.

Exercise 45 (Conjugation action by automorphisms). Fix a group G , a normal subgroup $H \trianglelefteq G$ and a subgroup $K \leq G$. Show that the rule

$$k \cdot h = khk^{-1}$$

for $k \in K$ and $h \in H$ determines an action of K on H via automorphisms, and the associated homomorphism $\rho: K \rightarrow \text{Aut}(H)$ is given by

$$\rho(k)(h) = khk^{-1}.$$

So now that we have a bit more context, let us now look at some examples of semidirect products.

Example 7.19. Let $K = \langle x \rangle$ be the cyclic of order 2 and $H = \langle y \rangle$ be the cyclic of order n for some $n \geq 2$. By the [UMP for cyclic groups](#), to give a homomorphism out of K is to pick the image i of the generator x , which must satisfy $i^2 = e$. In particular, i must be either the identity or an element of order 2.

Since H is abelian, the inverse map $f: H \rightarrow H$ given by $f(a) = a^{-1}$ is an automorphism of H ; we showed this in Problem Set 2.¹ This automorphism f is not the identity but it is its own inverse, so it has order 2. Therefore, by the [UMP for cyclic groups](#), there is a homomorphism

$$\rho: K \rightarrow \text{Aut}(H) \quad \text{with} \quad \rho(x)(y) = y^{-1}.$$

Consider the semidirect product $H \rtimes_{\rho} K$. The elements of $H \rtimes_{\rho} K$ are the tuples (y^i, x^j) for $0 \leq i \leq n-1$ and $0 \leq j \leq 1$. In particular, $|H \rtimes_{\rho} K| = 2n$. Set

$$\tilde{y} = (y, e_K) \in G \quad \text{and} \quad \tilde{x} = (e_H, x) \in G.$$

Then $\tilde{y}^n = (y, e_K)^n = (y^n, e_K) = (e_H, e_K) = e_G$ and $\tilde{x}^2 = (e_H, x)^2 = (e_H, x^2) = (e_H, e_K) = e_G$. Moreover,

$$\tilde{x}\tilde{y}\tilde{x}\tilde{y} = (e_H, x)(y, e_K)(e_H, x)(y, e_K) = (\rho(x)(y), x)(\rho(x)(y), x) = (y^{-1}, x)(y^{-1}, x) = (y^{-1}y, e) = e_G.$$

Looks familiar? Indeed, using our presentation for D_n from [Theorem 1.66](#) and the UMP for presentations from [Theorem 4.61](#), we have a homomorphism

$$\theta: D_n \longrightarrow G \quad \text{given by} \quad \theta(r) = (y, e_K) \text{ and } \theta(s) = (x, e_H).$$

Moreover, θ is surjective since

$$\theta(r^i s^j) = (y^i, x^j) \text{ for all } 0 \leq i \leq n-1, 0 \leq j \leq 1.$$

Since $|D_n| = |G| = 2n$, this surjection must also be a bijection, and we conclude that θ is an isomorphism. So the dihedral group is a semidirect product of the cyclic of order n and the cyclic group of order 2 respectively:

$$D_n \cong \langle y \rangle \rtimes_{\rho} \langle x \rangle$$

where ρ is the inverse map as described above.

So given any group, how can we recognize it is in fact a semidirect product?

Theorem 7.20 (Recognition theorem for internal semidirect products). *Let G be a group. Suppose we are given subgroups H and K of G such that*

$$H \trianglelefteq G \quad HK = G \quad \text{and} \quad H \cap K = \{e\}.$$

Let $\rho: K \rightarrow \text{Aut}(H)$ be the permutation representation of the action of K on H via automorphisms given by conjugation in G , meaning that

$$\rho(k)(h) = khk^{-1}.$$

Then

$$G \cong H \rtimes_{\rho} K$$

via the isomorphism $\theta: H \rtimes_{\rho} K \rightarrow G$ given by $\theta(x, y) = xy$. Moreover,

$$H \cong \{(h, e) \in H \rtimes_{\rho} K \mid h \in H\} \quad \text{and} \quad K \cong \{(e, k) \in H \rtimes_{\rho} K \mid k \in K\}.$$

¹In fact, we can say more: By [Exercise 42](#), $\text{Aut}(H) \cong (\mathbb{Z}/n)^{\times}$. In particular, -1 is an element of $(\mathbb{Z}/n)^{\times}$, and the associated automorphism sends y to y^{-1} .

Proof. First, we show that θ is a group homomorphism. Indeed,

$$\begin{aligned}\theta((y_1, x_1)(y_2, x_2)) &= \theta(y_1\rho(x_1)(y_2), x_1x_2) \\ &= y_1(x_1y_2x_1^{-1})x_1x_2 \\ &= y_1x_1y_2x_2 \\ &= \theta(y_1, x_1)\theta(y_2, x_2).\end{aligned}$$

Since $H \cap K = \{e\}$, the kernel of θ is

$$\ker(\theta) = \{(y, x) \in H \rtimes_{\rho} K \mid y = x^{-1}\} = \{e\}.$$

By construction, the image of θ is $KH = G$. Therefore, θ is an isomorphism. Finally,

$$\theta^{-1}(H) = \{(h, e) \mid h \in H\} \quad \text{and} \quad \theta^{-1}(K) = \{(e, k) \mid k \in K\}. \quad \square$$

Definition 7.21. Given subgroups H and K of G such that $H \trianglelefteq G$, $HK = G$, and $H \cap K = \{e\}$, we say that G is the **internal semidirect product** of H and K .

Example 7.22. Consider $G = D_n$ and its subgroups $H = \langle r \rangle$ and $K = \langle s \rangle$. Then $H \trianglelefteq G$, $K \leq G$, $HK = G$ and $H \cap K = \{e\}$. By Theorem 7.20, $G \cong H \rtimes_{\rho} K$, where $\rho: K \rightarrow \text{Aut}(H)$

$$\rho(s)(r^i) = sr^i s^{-1} = r^{n-i}.$$

The last equality is Exercise 10. Note in particular that K is *not* a normal subgroup of G . We had already seen in Example 7.9 that G is not the internal direct product of H and K , but now know it is their internal semidirect product. We also already knew that D_n is a semidirect product by Example 7.19.

For a fixed pair of groups H and K , different actions of K on H via automorphisms can result in isomorphic semidirect products. Indeed, determining when $K \rtimes_{\rho} H \cong K \rtimes_{\rho'} H$ is in general a tricky business. Here is an example of this:

Example 7.23. Let $n \geq 3$ and consider $G = S_n$, $H = A_n$, and $K = \langle (1 2) \rangle$. Then $H \trianglelefteq G$, $K \leq G$, $HK = G$ and $K \cap H = \{e\}$. Note that $H \cong C_2$ is the cyclic group with 2 elements. By Theorem 7.20,

$$S_n \cong A_n \rtimes_{\rho} C_2$$

where $\rho: C_2 \rightarrow \text{Aut}(A_n)$ sends x to conjugation by $(1 2)$. Similarly, we can also consider the subgroup $H' = \langle (1 3) \rangle = (1 2 3)\langle (1 2) \rangle(1 2 3)^{-1}$ of S_n , and we also have

$$S_n \cong A_n \rtimes_{\rho'} C_2$$

where $\rho': C_2 \rightarrow \text{Aut}(A_n)$ sends x to conjugation by $(1 3)$.

However, the actions determined by ρ and ρ' are not identical. For example,

$$\rho(x)(1 2 3) = (1 2 3) \quad \text{and} \quad \rho'(x)(1 2 3) = (2 1 3).$$

Yet

$$A_n \rtimes_{\rho} H \cong S_n \cong A_n \rtimes_{\rho'} H'.$$

One good reason why this happened in this case is that H and H' are conjugate in S_n .

Exercise 46. Let K be a finite cyclic group and let H be an arbitrary group. Suppose $\phi: K \rightarrow \text{Aut}(H)$ and $\theta: K \rightarrow \text{Aut}(H)$ are homomorphisms whose images are conjugate subgroups of $\text{Aut}(H)$; that is, suppose there is $\sigma \in \text{Aut}(H)$ such that $\sigma\phi(K)\sigma^{-1} = \theta(K)$. Then $H \rtimes_{\phi} K \cong H \rtimes_{\theta} K$.

Example 7.24. Let K be a cyclic group of prime order p and H be a group such that $\text{Aut}(H)$ has a unique subgroup of order p . Suppose $\phi: K \rightarrow \text{Aut}(H)$ and $\theta: K \rightarrow \text{Aut}(H)$ are any two *nontrivial* maps. Then ϕ and θ are injective, since K is simple and the kernel would be a proper normal subgroup. Hence, the images of ϕ and θ are both the unique subgroup of $\text{Aut}(H)$ of order p , and in particular they must be equal. Thus Exercise 46 applies to give $H \rtimes_{\phi} K \cong H \rtimes_{\theta} K$.

Remark 7.25. If $\rho: K \rightarrow \text{Aut}(H)$ is a nontrivial homomorphism, then the semidirect product $H \rtimes_{\rho} K$ is *never* abelian. Indeed, all we need is to consider any $k \in K$ such that $\rho(k) \neq \text{id}_H$, so that $\rho(k)(h) \neq h$ for some $h \in H$, and note that

$$(e, k)(h, e) = (\rho(k)(h), k) \quad \text{while} \quad (h, e)(e, k) = (h\rho(e)(e), k) = (h, k).$$

Thus we can use semidirect products to construct nonabelian groups. Given an integer $n \geq 2$, to construct a nonabelian group we might set out to find groups K and H such that

$$|K||H| = n$$

and such that there exists a nontrivial homomorphism

$$\rho: K \rightarrow \text{Aut}(H).$$

7.3 Finitely generated groups

Recall that a group G is finitely generated if it $G = \langle A \rangle$, where A is a finite set.

Remark 7.26. Any finite group G is finitely generated, since we can take $A = G$. However, a finitely generated group need not be finite: for example \mathbb{Z} is even cyclic but infinite.

The main theorem of this section is a special case of a much more general theorem we will prove in the Spring: the classification of finitely generated modules over PIDs. Thus we leave the proof for next semester.

Theorem 7.27 (Fundamental Theorem of Finitely Generated Abelian Groups: Invariant Factor Form). *Let G be a finitely generated abelian group. There exist integers $r \geq 0$, $t \geq 0$, and $n_i \geq 2$ for $1 \leq i \leq t$, satisfying $n_1 | n_2 | \cdots | n_t$ such that*

$$G \cong \mathbb{Z}^r \times \mathbb{Z}/n_1 \times \cdots \times \mathbb{Z}/n_t.$$

Moreover, the list r, s, n_1, \dots, n_t is uniquely determined by G .

Definition 7.28. In Theorem 7.27, the number r is the **rank** of G , the numbers n_1, \dots, n_t are the **invariant factors** of G , and the decomposition of G in this form is the **invariant factor decomposition** of G .

Remark 7.29. A finitely generated abelian group is finite if and only if its rank is 0. A special case of the [classification theorem](#) is that if G is a finite abelian group then

$$G \cong \mathbb{Z}/n_1 \times \cdots \times \mathbb{Z}/n_t$$

for a unique list of integers $n_i \geq 2$ such that $n_1 | n_2 | \cdots | n_t$.

Here is another version of the [classification theorem](#):

Theorem 7.30 (Fundamental Theorem of Finitely Generated Abelian Groups: Elementary Divisor Form). *Let G be a finitely generated abelian group. Then there exist integers $r \geq 0$ and $s \geq 0$, not necessarily distinct positive prime integers p_1, \dots, p_s , and integers $a_i \geq 1$ for $1 \leq i \leq s$ such that*

$$G \cong \mathbb{Z}^r \times \mathbb{Z}/p_1^{a_1} \times \cdots \times \mathbb{Z}/p_s^{a_s}.$$

Moreover, r and s are uniquely determined by G , and the list of prime powers $p_1^{a_1}, \dots, p_s^{a_s}$ is unique up to the ordering.

Definition 7.31. In Theorem 7.30, the number r is the **rank** of G , the $p_i^{a_i}$ are the **elementary divisors** of G , and the decomposition of G is called the **elementary divisor decomposition** of G .

The two forms of the classification theorem are equivalent, which we can prove using the [CRT](#). Rather than a careful proof that the two versions of the classification theorem are equivalent, we will now see in examples how the [CRT](#) allows us to go between invariant factors and elementary divisors.

Example 7.32 (Converting elementary divisors to invariant factors). Suppose G is a finitely generated abelian group of rank 3 with elementary divisors 4, 8, 9, 27, 25. This means that

$$G \cong \mathbb{Z}^3 \times \mathbb{Z}/4 \times \mathbb{Z}/8 \times \mathbb{Z}/9 \times \mathbb{Z}/27 \times \mathbb{Z}/25.$$

By the [CRT](#),

$$\mathbb{Z}/8 \times \mathbb{Z}/27 \times \mathbb{Z}/25 \cong \mathbb{Z}/(8 \cdot 27 \cdot 25) \quad \text{and} \quad \mathbb{Z}/4 \times \mathbb{Z}/9 \cong \mathbb{Z}/(4 \cdot 9),$$

so that

$$G \cong \mathbb{Z}^3 \times \mathbb{Z}/(8 \cdot 27 \cdot 25) \times \mathbb{Z}/(4 \cdot 9) = \mathbb{Z}^3 \times \mathbb{Z}/5400 \times \mathbb{Z}/36.$$

Since $36 \mid 5400$, we conclude that G has rank 3 and invariant factors 5400 and 36.

Example 7.33 (Converting invariant factors to elementary divisors). Let

$$G \cong \mathbb{Z}^4 \times \mathbb{Z}/6 \times \mathbb{Z}/36 \times \mathbb{Z}/180.$$

Then by the [CRT](#),

$$G \cong \mathbb{Z}^4 \times \mathbb{Z}/2 \times \mathbb{Z}/3 \times \mathbb{Z}/4 \times \mathbb{Z}/9 \times \mathbb{Z}/4 \times \mathbb{Z}/5 \times \mathbb{Z}/9,$$

is the elementary divisor form for G .

Example 7.34. Let $G = \mathbb{Z}/60 \times \mathbb{Z}/50$. This group is finite and abelian, and thus $r = 0$, but not in either invariant factor nor elementary divisor factorization.

Applying the [CRT](#) to $60 = 12 \cdot 5 = 2^2 \cdot 3 \cdot 5$ and $50 = 2 \cdot 5^2$, we have

$$\mathbb{Z}/60 \cong \mathbb{Z}/4 \times \mathbb{Z}/3 \times \mathbb{Z}/5 \quad \text{and} \quad \mathbb{Z}/50 \cong \mathbb{Z}/2 \times \mathbb{Z}/25$$

so

$$G \cong \mathbb{Z}/2 \times \mathbb{Z}/4 \times \mathbb{Z}/3 \times \mathbb{Z}/5 \times \mathbb{Z}/25.$$

This gives the elementary divisor decomposition: G has rank 0 and elementary divisors 2, 4, 3, 5, and 25. Applying the [CRT](#) again, in a different way, gives

$$G \cong \mathbb{Z}/(4 \cdot 3 \cdot 25) \times \mathbb{Z}/(2 \cdot 5) = \mathbb{Z}/300 \times \mathbb{Z}/10.$$

This is the invariant factor decomposition: G has rank 0 and invariant factors 10 and 300.

This classification makes the classification of finite abelian groups a very quick matter.

Example 7.35. Let us classify the abelian groups of order 75. First, note that $75 = 5^2 \cdot 3$. The two possible elementary divisor decompositions are

$$\mathbb{Z}/25 \times \mathbb{Z}/3 \quad \text{and} \quad \mathbb{Z}/5 \times \mathbb{Z}/5 \times \mathbb{Z}/3.$$

Note that the two groups above are not isomorphic. This is part of the theorem, but to see this directly, note that there is an element of order 25 in $\mathbb{Z}/25 \times \mathbb{Z}/3$, namely $([1]_{25}, [0]_3)$ whereas every element $(a, b, c) \in \mathbb{Z}/5 \times \mathbb{Z}/5 \times \mathbb{Z}/3$ has order

$$|(a, b, c)| = \text{lcm}(|a|, |b|, |c|) \leq 3 \cdot 5 = 15,$$

since $|a|, |b| \in \{1, 5\}$ and $|c| \in \{1, 3\}$.

Alternatively, the two possible invariant factor decompositions are

$$\mathbb{Z}/75 \quad \text{or} \quad \mathbb{Z}/15 \times \mathbb{Z}/5.$$

They are also not isomorphic, as the second option has no elements of order 75.

Remark 7.36. Let $n = p_1^{e_1} \cdots p_k^{e_k}$ for *distinct* positive prime integers p_1, \dots, p_k and integers $e_i \geq 1$. The classification of finitely generated abelian groups implies that there are $p(e_1) \cdots p(e_k)$ isomorphism classes of abelian groups of order n , where $p(m)$ is the number of partitions of m . For example, for $n = 2^4 \cdot 3^5 \cdot 5^2$ there are

$$p(4)p(6)p(2) = 5 \cdot 7 \cdot 2 = 70$$

abelian groups of order n up to isomorphism.

7.4 Classifying finite groups of a given order

We can now combine the ideas from Sylow theory, (semi)direct products and the classification theorem for finitely generated abelian groups to classify the isomorphism classes of groups of a given order. You have already done some examples of this kind, such as the following problem set question:

Exercise 47. Show that any group of order 6 is isomorphic either to $\mathbb{Z}/6$ or to D_3 .

Here is an example of the type of classification theorem we can prove.

Theorem 7.37. *Let $p < q$ be primes.*

- (1) *If p does not divide $q - 1$, there is a unique group of order pq up to isomorphism, the cyclic group C_{pq} .*
- (2) *If p divides $q - 1$, there are exactly two groups of order pq up to isomorphism, the cyclic group C_{pq} and a nonabelian group.*

Proof. Let G be a group of order pq and let $n_q = |\text{Syl}_q(G)|$. Since $n_q \equiv 1 \pmod{q}$, $n_q \mid p$, p is prime, and $q > p$, we must have $n_q = 1$. Thus by ??, the unique Sylow q -subgroup H is a normal subgroup.²

Now let K be a Sylow subgroup of order p . Since H is normal, by Corollary 4.49 we know that HK is a subgroup of G . By Lagrange's Theorem, $|H \cap K|$ divides $|H|$ and $|H \cap K|$ divides $|K|$. Therefore, $H \cap K = \{e_G\}$. By Exercise 24,

$$|HK| = \frac{|H||K|}{|H \cap K|} = \frac{q \cdot p}{1} = pq = |G|$$

and so $HK = G$. The [recognition theorem for semidirect products](#) thus yields that

$$G \cong H \rtimes_\rho K$$

for some homomorphism $\rho: K \rightarrow \text{Aut}(H)$. Note that H and K are cyclic, since they have prime order (see Exercise 18). Let us identify H with $C_q = \langle x \mid x^q \rangle$ and K with $C_p = \langle y \mid y^p \rangle$. Then

$$G \cong C_q \rtimes_\rho C_p \quad \text{for some homomorphism } \rho: C_p \rightarrow \text{Aut}(C_q).$$

We just need to classify all such semidirect products up to isomorphism. By the [UMP of cyclic groups](#), the homomorphism $\rho: C_p \rightarrow \text{Aut}(C_q)$ is uniquely determined by the image of the generator x , which must be an element $\alpha \in \text{Aut}(C_q)$ with $\alpha^p = \text{id}$. Given such an α , we have $\rho(y) = \alpha$ and more generally $\rho(y^i) = \alpha^i$.

By Exercise 43, $\text{Aut}(C_q)$ is cyclic of order $q - 1$. On the other hand, $\text{im}(\rho)$ is a subgroup of both C_p and $\text{Aut}(C_q)$, so its order must divide both p and $q - 1$. In particular, there is a nontrivial automorphism ρ if and only if $p \mid q - 1$.

If p does not divide $q - 1$, then ρ is trivial, and by Example 7.14 and the [CRT](#) we have

$$G \cong C_q \times C_q \cong C_{pq}.$$

²Alternatively, H is normal since $[G : H] = p$ is the smallest prime that divides $|G|$.

If p does divide $q - 1$, there is at least one nontrivial ρ . We still have $G \cong C_{pq}$ if ρ is trivial. When ρ is nontrivial, G is not abelian, giving us at least two isomorphism classes. It remains to show that if ρ_1 and ρ_2 are any two nontrivial homomorphisms from C_p to $\text{Aut}(C_q)$, then the resulting semidirect products are isomorphic.

Since $\text{Aut}(C_q)$ is a cyclic group and p divides its order, it has a unique subgroup of order p . Thus, we conclude that $\text{im}(\rho_1) = \text{im}(\rho_2)$, so that by Exercise 46 we have

$$C_q \rtimes_{\rho_1} C_p \cong C_q \rtimes_{\rho_2} C_p.$$

□

Example 7.38. If $p = 2$ and q is any odd prime, then there are two groups of order $2q$ up to isomorphism: C_{2q} and D_q .

Part II

Rings

Chapter 8

An introduction to ring theory

8.1 Definitions and examples

Definition 8.1. A **ring** is a set R equipped with two binary operations, $+$ and \cdot , satisfying:

- $(R, +)$ is an abelian group. We use additive notation: the identity element for $+$ is denoted by 0 and the inverse of an element r for $+$ is written as $-r$.
- The operation \cdot is associative, making (R, \cdot) a semigroup.
- There is a multiplicative identity element, written as 1 , such that

$$1 \cdot a = a = a \cdot 1$$

for all $a \in R$, and thus (R, \cdot) is a monoid.

- Distributivity: For all $a, b, c \in R$, we have

$$a \cdot (b + c) = a \cdot b + a \cdot c \quad \text{and} \quad (a + b) \cdot c = a \cdot c + b \cdot c.$$

- We also require $0 \neq 1$.

We sometimes write 0_R and 1_R if we need to emphasize what ring these elements live in.

Definition 8.2. An object satisfying just the first three conditions, but without a multiplicative identity, is a **nonunital ring** or a **rng**. To emphasize that R has a multiplicative identity, one might say that a ring is **unital**.

While some authors consider nonunital rings, in this class all our rings will be unital.

Remark 8.3. If we drop the requirement that $0 \neq 1$, we may consider the **zero ring**, which is the set $\{0\}$ together with the only possible operations on it. Conversely, if $1 = 0$ in a ring, then $R = \{0\}$, since in this case all $a \in R$ satisfy $a \cdot 0 = 0$ and hence $a = a \cdot 1 = a \cdot 0 = 0$.

Example 8.4. The integers with the usual addition and multiplication form a ring $(\mathbb{Z}, +, \cdot)$.

Remark 8.5. The last condition, asking that $1 \neq 0$, is not universal: some authors allow the *zero ring*, which is the ring with only one element. Requiring $0 \neq 1$ is really asking that R should have at least two elements.

Lemma 8.6 (Ring arithmetic). *The following hold for any ring R and all $a, b \in R$:*

- (1) $a \cdot 0 = 0 = 0 \cdot a$,
- (2) $(-a)b = -(ab) = a(-b)$,
- (3) $(-a)(-b) = ab$.
- (4) 1 is unique, and
- (5) $(-1)a = -a$.

Proof. (1) Note that

$$a \cdot 0 = a \cdot (0 + 0) = a \cdot 0 + a \cdot 0.$$

By subtracting $a \cdot 0$ on both sides, we conclude that

$$a \cdot 0 = a \cdot (0 + 0) = 0.$$

Analogously, $0 \cdot a = 0$.

- (2) By distributivity,

$$ab + (-a)b = (a - a)b = 0 \cdot b = 0.$$

Thus $(-a)b = -ab$. Analogously, $a(-b) = -ab$.

- (3) Applying the previous property twice, and noting that $-(-x) = x$ by Exercise 2 (3), we get

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

- (4) Note that (R, \cdot) is a monoid, and thus the identity 1 is unique by Lemma 1.7.

- (5) We have $(-1)a = -1 \cdot a = -a$. □

There are some additional conditions we might ask for a ring to satisfy, and that are so important they have their own names:

Definition 8.7. A ring R is

- a **commutative ring** if \cdot is commutative, meaning that for all $a, b \in R$ ¹

$$a \cdot b = b \cdot a.$$

- a **noncommutative ring** if it is not commutative.
- a **division ring** if $(R - \{0\}, \cdot)$ is a group, meaning that every nonzero element has a multiplicative inverse.
- a **field** if it is a commutative division ring.

We are now ready to see many examples of rings.

¹The word *abelian* is never used in the context of rings, except to say things like “the additive group $(R, +)$ is abelian”.

- Example 8.8.** (1) The ring \mathbb{Z} is a commutative ring.
- (2) Let $n \geq 2$. The set \mathbb{Z}/n of integers modulo n is a commutative ring under addition and multiplication modulo n . Note that \mathbb{Z}/n is a field if and only if n is prime.
- (3) The familiar sets of numbers $\mathbb{Q}, \mathbb{R}, \mathbb{C}$ are fields.
- (4) (**Matrix ring**) If R is any ring, not necessarily commutative, then the set $\text{Mat}_n(R)$ of $n \times n$ matrices with entries in R is a ring with the usual rules for addition and multiplication of square matrices.
- (5) (**The endomorphism ring of an abelian group**) Let $A = (A, +)$ be any abelian group, and set $\text{End}_{Ab}(A)$ to be the collection of endomorphisms of A — that is, the set of group homomorphisms $f: A \rightarrow A$ from A to itself. This set of endomorphisms $\text{End}_{Ab}(A)$ is a ring with pointwise addition

$$(f + g)(a) := f(a) + g(a)$$

and multiplication given by composition of functions

$$f \cdot g := f \circ g.$$

The additive identity is the 0-map and the multiplicative identity is the identity map. This is almost always a noncommutative ring.

- (6) (**The real Hamiltonian quaternion ring**) Let i, j, k be formal symbols and set \mathcal{H} to be the four dimensional \mathbb{R} -vector space consisting of all expressions of the form $a + bi + cj + dk$ with $a, b, c, d \in \mathbb{R}$. We claim that this can be given a ring structure, as follows. Addition is vector space addition:

$$(a + bi + cj + dk) + (a' + b'i + c'j + d'k) = (a + a') + (b + b')i + (c + c')j + (d + d')k.$$

Moreover, multiplication is uniquely determined by the axioms of a ring together with the rules

$$i^2 = j^2 = k^2 = -1, -ji = ij = k, -kj = jk = i, -ik = ki = j.$$

and the fact that the real coefficients commute with each other and i, j, k .

It is not obvious that the multiplication defined in this way satisfies associativity, but in fact it does, and this amounts to conditions very similar to the associativity of the group Q_8 , which we discussed in Section 1.4.

This ring \mathcal{H} is a division ring, since one can check that

$$(a + bi + cj + dk)^{-1} = \frac{a - bi - cj - dk}{\|a + bi + cj + dk\|}$$

where

$$\|a + bi + cj + dk\| := a^2 + b^2 + c^2 + d^2.$$

In the equation above, $\|a + bi + cj + dk\|$ is a nonzero real number if $a + bi + cj + dk$ is not the zero element. The quantity $\|a + bi + cj + dk\|$ is called the **norm** of the quaternion $a + bi + cj + dk$.

- (7) If X is a set and R is a ring, let $\text{Fun}(X, R)$ be the collection of set-theoretic functions from X to R , and consider the pointwise addition and multiplication of functions:

$$(f + g)(x) := f(x) + g(x) \quad \text{and} \quad (f \cdot g)(x) := f(x) \cdot g(x)$$

The set $\text{Fun}(X, R)$ is a ring with these operations. In this ring, the zero is the function that is constantly equal to zero, and the identity is the constant function equal to 1. If X is a finite set and $|X| = n$, then $\text{Fun}(X, R)$ may be identified with $R^n = \underbrace{R \times \cdots \times R}_n$,

the direct product of n copies of R .

Just like with groups, there are constructions that allow us to take old rings and build new ones.

Definition 8.9 (Direct product of rings). Let R and S be two rings. The cartesian product $R \times S$ has a natural ring structure with addition and multiplication defined componentwise:

$$(a, b) + (c, d) = (a + c, b + d) \quad \text{and} \quad (a, b) \cdot (c, d) = (a \cdot c, b \cdot d).$$

The additive identity is $0_{R \times S} = (0_R, 0_S)$ and the multiplicative identity is $1_{R \times S} = (1_R, 1_S)$.

Exercise 48. Check that the direct product of two rings is a ring. Moreover, prove that $R \times S$ is a commutative ring if and only if R and S are both commutative.

Exercise 49. Show that the direct product of two fields is *never* a field.

Definition 8.10 (Polynomial ring). If R is any ring and x is a “variable”, then $R[x]$ denotes the collection of R -linear combination of powers of x — i.e., formal expressions of the form

$$r_0 + r_1x + r_2x^2 + \cdots + r_nx^n$$

with $n \geq 0$ and $r_i \in R$, and two such expressions are deemed equal if their coefficients are the same.

We make $R[x]$ into a ring by the usual rule for adding and multiplying polynomial expressions, treating x as commuting with all elements of R . So

$$(r_0 + r_1x + r_2x^2 + \cdots + r_nx^n) + (r'_0 + r'_1x + r'_2x^2 + \cdots + r'_mx^m) = (r_0 + r'_0) + (r_1 + r'_1)x + \cdots$$

or more precisely, setting $r_i = 0$ for $i > n$ and $r'_i = 0$ for $i > m$,

$$(r_0 + r_1x + r_2x^2 + \cdots + r_nx^n) + (r'_0 + r'_1x + r'_2x^2 + \cdots + r'_mx^m) = \sum_{i=0}^{\max m, n} (r_i + r'_i)x^i,$$

while

$$(r_0 + r_1x + r_2x^2 + \cdots + r_nx^n) \cdot (r'_0 + r'_1x + r'_2x^2 + \cdots + r'_mx^m) = \sum_k \left(\sum_{a+b=k} r_a r'_b \right) x^k.$$

This ring $R[x]$ is the **polynomial ring** in one variable over R . One can also talk about polynomial rings in many variables. For a finite set of variables x_1, \dots, x_n , the ring $R[x_1, \dots, x_n]$ can be constructed inductively by setting

$$R[x_1, \dots, x_n] = R[x_1, \dots, x_{n-1}][x_n].$$

More generally, given an infinite set of variables X , an element in the polynomial ring $R[X]$ can be obtained by formally adding finitely many **monomials** in X with coefficients in R , which are terms of the form $rx_1^{a_1} \cdots x_n^{a_n}$ with $x_i \in X$ and integers $a_i \geq 0$. Each polynomial in $R[X]$ uses only finitely many variables, and thus sums and products of two elements are obtained as in the polynomial ring in that finite set of variables.

Exercise 50. Check that if R is a ring then so is $R[x]$. Moreover, show that if R is commutative, then so is $R[x]$.

We will later discuss polynomial rings in more detail. For now, we note that in many circumstances when one says *a polynomial ring*, one often means a polynomial ring *over a field*.

8.2 Units and zerodivisors

Elements in a ring might have certain special properties:

Definition 8.11. An element a of a ring is called a **unit** if there exists $b \in R$ such that $ab = 1$ and $ba = 1$. The set of all units of a ring R is denoted R^\times .

Exercise 51. Show that if a is a unit in a ring R , then there is a unique $b \in R$ such that $ab = 1$ and $ba = 1$.

Definition 8.12. Let a be a unit in a ring R . The unique $b \in R$ such that $ab = 1 = ba$ is called the **inverse** of a , denoted by a^{-1} .

Exercise 52. Show that the set of units in a ring R forms a group (R^\times, \cdot) with respect to multiplication.

Example 8.13. (1) The units in \mathbb{Z} are $\mathbb{Z}^\times = \{\pm 1\}$.

(2) For all $n \geq 2$,

$$\mathbb{Z}/n^\times = \{[j]_n \mid \gcd(j, n) = 1\}.$$

(3) For all $n \geq 1$ and any field F ,

$$\mathrm{Mat}_n(F)^\times = \mathrm{GL}_n(F).$$

Exercise 53. Let R be a ring. Find all the units of $R[x]$.

Definition 8.14. A **zerodivisor** in a ring R is an element $x \in R$ such that $x \neq 0$ but either $xy = 0$ or $yx = 0$ for some $y \neq 0$.

Example 8.15. The ring $\text{Mat}_2(\mathbb{R})$ has lots of zerodivisors: for example,

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

is a zerodivisor since $A^2 = 0$.

Example 8.16. In the ring $\mathbb{Z}/6$, the element $[2]_6$ is a zerodivisor since $[2]_6[3]_6 = 0$.

Lemma 8.17. *Let R be any ring. There is no element $r \in R$ that is both a unit and a zerodivisor.*

Proof. Suppose that a is both a zerodivisor and a unit. Then there exists $b \neq 0$ such that $ab = 0$ or $ba = 0$. Multiplying either of these equations by a^{-1} gives $b = 0$, which is a contradiction. \square

Definition 8.18. A ring R is an **integral domain**, often shortened to **domain**, if R is commutative and has no zerodivisors.

Remark 8.19. If one allows the zero ring, then in the definition of a domain we should explicitly require $1 \neq 0$. Moreover, if one allows for nonunital rings, then we should also require all domains to be unital.

Remark 8.20. Any domain R satisfies what is known as the **cancellation rule**: given any nonzero element $a \in R$,

$$ab = ac \implies b = c.$$

Indeed, the equality

$$ab = ac \implies a(b - c) = 0,$$

but since a is not a zerodivisor we must have $b - c = 0$.

The cancellation rule does not hold if R is not a domain: if a and b are nonzero and $ab = 0$, then $ab = a \cdot 0$ even though $b \neq 0$.

Corollary 8.21. *Every field is a domain.*

Proof. If R is a field, then every nonzero $r \in R$ is a unit, and thus by Lemma 8.17 r is not a zerodivisor. Thus R has no zerodivisors, and must be a domain. \square

In contrast, not every domain must be a field.

Example 8.22. The ring \mathbb{Z} is a domain but not a field.

Example 8.23. Fix an integer $n \geq 2$ and consider the ring \mathbb{Z}/n . If n is composite, say $n = ab$ with $1 < a, b < n$, then $[a]_n[b]_n = 0$ in \mathbb{Z}/n . In particular, $[a]_n$ and $[b]_n$ are zerodivisors and \mathbb{Z}/n is not a domain.

In contrast, if n is prime then \mathbb{Z}/n is a field, and thus in particular it is a domain. Putting all this together, we see that

$$\mathbb{Z}/n \text{ is a domain} \iff n \text{ is prime} \iff \mathbb{Z}/n \text{ is a field.}$$

In fact, this is a special case of a more general fact:

Exercise 54. Show that every finite domain is a field.

Definition 8.24. An element a in a ring R is **nilpotent** if $a^n = 0$ for some integer $n \geq 1$.

Exercise 55. Show that if a is a nonzero nilpotent element, then a is a zerodivisor.

Thus there are no nontrivial nilpotent elements in a domain.

Exercise 56. Show that if a is a nilpotent element in a ring R , then $1 - a$ is a unit.

Exercise 57. Given an integer $n \geq 1$, describe all the nilpotent elements in \mathbb{Z}/n .

Definition 8.25. An element a in a ring R is **idempotent** if $a^2 = a$.

Exercise 58. Show that if e is an idempotent element in a ring R , then $1 - e$ is also an idempotent element.

Exercise 59. Show that if F is a field, then 0 and 1 are the only idempotent elements.

8.3 Subrings

Definition 8.26. A **subring** of a ring R is a subset $S \subseteq R$ such that S is a ring under the operations of R and $1_S = 1_R$. When R is a field, a subring of R that is also a field is called a **subfield** of R .

Some authors do not include the condition that $1_S = 1_R$ in their definition of subring. However, we think of the identity as part of the basic data of the ring, and thus it is desirable for it to be shared with any subring. As we will see later when we define ideals, this will make our definition of ideal *quite* different in practice from what we would get if we allowed a subring to not be unital, or not share the multiplicative identity with the original ring.

Exercise 60. Prove that for a ring R , a subset S of R is a subring if and only if $1_R \in S$ and for all $x, y \in S$ we have $x - y \in S$ and $xy \in S$.

Exercise 61. Any subring of a commutative ring is a commutative ring. Any subring of a domain is a domain.

Exercise 62. Prove that the set of \mathbb{R} -linear combinations of

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} \sqrt{-1} & 0 \\ 0 & -\sqrt{-1} \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & \sqrt{-1} \\ \sqrt{-1} & 0 \end{bmatrix}.$$

forms a subring of $\text{Mat}_2(\mathbb{C})$.

We will later defined what it means for two rings to be isomorphic. The ring in Exercise 62 is isomorphic to the quaternions ring \mathcal{H} .

Remark 8.27. Let F be a ring and $R = \text{Mat}_n(F)$ with $n \geq 2$. Let S be the subset of R consisting of matrices whose only nonzero element is in the upper left corner. Then S is a ring under same operations as R , and in fact $S \cong R$, but S is not a subring of R according to our definition, since $1_S \neq 1_R$.

Example 8.28. • The following is a chain of subrings:

$$\mathbb{Z} \subseteq \mathbb{Q} \subseteq \mathbb{R} \subseteq \mathbb{C} \subseteq \mathcal{H}.$$

In the last containment, we think of \mathbb{C} as those elements $a + bi + cj + dk$ of \mathcal{H} with $c = d = 0$.

- For any ring R and integer $n \geq 1$, the set of scalar matrices

$$\{rI_n \mid r \in R\}$$

is a subring of $\text{Mat}_n(R)$.

- For any ring R and integer $n \geq 1$, the set of all diagonal matrices is a subring of $\text{Mat}_n(R)$.
- The set

$$\mathbb{Z}[i] = \{a + bi \mid a, b \in \mathbb{Z}\}$$

is a subring of \mathbb{C} called the ring of **Gaussian integers**.

Definition 8.29. The **center** of a ring R is the set

$$Z(R) := \{z \in R \mid zr = rz \text{ for all } r \in R\}.$$

An element in R is called **central** if it is in the center of R .

Exercise 63. Show that the center $Z(R)$ is a subring of R .

Example 8.30. If R is commutative, then $Z(R) = R$.

The center measures how far R is from being commutative.

Exercise 64. Show that the center of \mathcal{H} is \mathbb{R} .

Exercise 65. Show that for any commutative ring R , the center of $\text{Mat}_n(R)$ is the collection of scalar matrices.

Lemma 8.31. Let d be a squarefree integer, meaning that the prime factorization of d has no repeated primes. Then

$$\mathbb{Q}(\sqrt{d}) := \{a + b\sqrt{d} \mid a, b \in \mathbb{Q}\}$$

is a subfield of the field \mathbb{C} . Moreover,

$$\mathbb{Z}[\sqrt{d}] := \{a + b\sqrt{d} \mid a, b \in \mathbb{Z}\}$$

is a subring of $\mathbb{Q}(\sqrt{d})$.

Proof. We leave this as an exercise to prove that $\mathbb{Q}(\sqrt{d})$ and $\mathbb{Z}[\sqrt{d}]$ are closed under subtractions and products and contain 1, and thus are subrings of \mathbb{C} by Exercise 60.

It remains to show that $\mathbb{Q}(\sqrt{d})$ is a field, which amounts to the claim that $\mathbb{Q}(\sqrt{d})$ is also closed inside \mathbb{C} under taking inverses of nonzero elements. Suppose $r + q\sqrt{d} \neq 0$. Then its inverse in \mathbb{C} is

$$(r + q\sqrt{d})^{-1} = \frac{r - q\sqrt{d}}{r^2 - dq^2} \in \mathbb{Q}(\sqrt{d}).$$

A slightly subtle point here is that the fraction above makes sense. To see that, note that if $r^2 - dq^2 = 0$, then either $r = q = 0$ or $d = (r/q)^2$. But $r = q = 0$ contradicts the assumption that $r + q\sqrt{d} \neq 0$, so that's impossible. If $d = (r/q)^2$, since d is an integer then q^2 must divide r^2 , and thus q divides r . Therefore, $d = (r/q)^2$ is a square, contradicting our assumption that d is squarefree. \square

Remark 8.32. In Lemma 8.31, note that we do allow d to be negative. For instance, Lemma 8.31 applies to $\mathbb{Q}(\sqrt{-5})$ and $\mathbb{Z}[\sqrt{-5}]$. Indeed, this is a somewhat interesting example, as $\mathbb{Z}[\sqrt{-5}]$ is a classic example of a ring that is not UFD, something we will discuss later.

It does make sense to speak of $\mathbb{Q}(\sqrt{d})$ and $\mathbb{Z}[\sqrt{d}]$ when d has repeated prime factors, but it just leads to redundant examples. For instance, if $d = 12$, then $\mathbb{Q}(\sqrt{12}) = \mathbb{Q}(\sqrt{3})$ and $\mathbb{Z}[\sqrt{12}] = \mathbb{Z}[\sqrt{3}]$.

Example 8.33. The ring $\mathbb{Z}[\sqrt{d}]$ is an integral domain: it is a subring of \mathbb{C} , and \mathbb{C} is a domain and thus a field by Corollary 8.21.

Remark 8.34. The difference in notation (more precisely, in the parenthesis) between $\mathbb{Z}[\sqrt{d}]$ and $\mathbb{Q}(\sqrt{d})$ will be explained next semester. In short, if R is a subring of S and $s \in S$, then $R[s]$ is the smallest subring of S that contains both R and s , which for a subfield F of a field L and an element $a \in L$, $F(a)$ denotes the smallest subfield of L containing F and a . In this case, it just happens that the sets $\mathbb{Z}[\sqrt{d}]$ and $\mathbb{Q}(\sqrt{d})$ look surprisingly similar.

8.4 Ideals

Notation 8.35. Given a ring R and a subset $S \subseteq R$, we write

$$RS := \{ra \mid a \in S, r \in R\} \quad \text{and} \quad SR := \{ar \mid a \in S, r \in R\}.$$

If $S = \{a\}$, then we Ra instead of $R\{a\}$ and aR instead of $\{a\}R$. Finally, given $a, b \in R$, we write

$$Ra + Rb := \{ra + sb \mid r, s \in R\}.$$

Definition 8.36. For a ring R , an **ideal** (or a **two sided ideal**) of R is a nonempty subset I such that

- Closure under addition: $(I, +)$ is a subgroup of $(R, +)$.
- Absorption:² for all $r \in R$ and $a \in I$, we have $ra \in I$ and $ar \in I$. More concisely: $RI \subseteq I$ and $IR \subseteq I$.

²One might even write $RIR \subseteq I$.

For noncommutative rings, one speaks also about left ideals and right ideals.

Definition 8.37. A **left ideal** of a ring R is a subgroup I of $(R, +)$ which satisfies $RI \subseteq I$, while a **right ideal** is a subgroup I of $(R, +)$ which satisfies $IR \subseteq I$.

Our definition of rings, or more precisely our insistence that all rings have 1, makes ideals and subrings very different beasts.

Remark 8.38. If an ideal I contains 1, then by the absorption property we must have $I = R$, since for all $a \in R$ we have

$$a = a \cdot 1 \in I.$$

Thus the only subset of R that is both an ideal and a subring is R itself.

Here are some examples of ideals:

Example 8.39. (1) Every ring R has at least two ideals: $\{0\}$ and R itself.

(2) The ideals of \mathbb{Z} are of the form $\mathbb{Z} \cdot n$ for various n , but we will prove this later.

One can show (exercise!) that

$$\mathbb{Z} \cdot 6 + \mathbb{Z} \cdot 10 = \{m \cdot 6 + n \cdot 10 \mid m, n \in \mathbb{Z}\}$$

is also an ideal, and so it must have the form $\mathbb{Z} \cdot n$ for some n . Indeed,

$$\mathbb{Z} \cdot 6 + \mathbb{Z} \cdot 10 = \mathbb{Z} \cdot 2.$$

(3) The sets $R_i = \left\{ \begin{bmatrix} 0 & 0 & \cdots & 0 \\ \dots & \dots & \cdots & \dots \\ a_{i1} & a_{i2} & \cdots & a_{in} \\ \dots & \dots & \cdots & \dots \\ 0 & 0 & \cdots & 0 \end{bmatrix} \right\}$ and $L_j = \left\{ \begin{bmatrix} 0 & \cdots & a_{j1} & \cdots & 0 \\ \dots & \dots & \cdots & \cdots & \dots \\ 0 & \cdots & a_{ji} & \cdots & 0 \\ \dots & \dots & \cdots & \cdots & \dots \\ 0 & \cdots & a_{jn} & \cdots & 0 \end{bmatrix} \right\}$ are a right ideal and a left ideal of $\text{Mat}_n(R)$ respectively. Neither of these are two-sided ideals if $n \geq 2$.

Definition 8.40. An ideal I in a ring R is a **proper ideal** if $I \neq R$, and **nontrivial** if $I \neq \{0\}$.

Some authors might say an ideal is nontrivial to mean it is proper and nontrivial.

Exercise 66. Prove that an ideal I is proper if and only if I contains no units.

Exercise 67. Let R be a commutative ring. Show that R is a field if and only if R has only two ideals, $\{0\}$ and R .

Definition 8.41. A ring R is a **simple ring** if it has no proper nontrivial ideals, meaning that the only ideals of R are R and $\{0\}$.

Exercise 68. If F is a field or, more generally, a division ring, and $n \geq 1$ is an integer, prove that $\text{Mat}_{n \times n}(F)$ is a simple ring.

Here are some operations that one can perform with ideals.

Lemma 8.42. *Let R be a ring and let I and J be ideals of R . Then*

(1) *The sum of ideals*

$$I + J := \{a + b \mid a \in I, b \in J\}$$

is an ideal.

(2) *The intersection of ideals is an ideal: $I \cap J$ is an ideal, and more generally the intersection*

$$\bigcap_{\alpha \in J} I_\alpha$$

of any collection of ideals I_α of R is an ideal.

(3) *The product of ideals is an ideal:*

$$IJ := \left\{ \sum_{i=1}^n a_i b_i \mid n \geq 0, a_i \in I, b_i \in J \right\}$$

is an ideal such that $IJ \subseteq I \cap J$.

The set of all ideals of a ring R is a lattice with respect to the partial order given by containment. In this lattice, the supremum of a pair of ideals I and J is $I + J$ and the infimum is $I \cap J$.

Exercise 69. Prove Lemma 8.42.

Remark 8.43. However, the union of ideals is typically *not* an ideal. For example, in \mathbb{Z} , the sets of even integers $I = 2\mathbb{Z}$ and multiples of 3 $J = 3\mathbb{Z}$ are both ideals, but $I \cup J$ is not ideal since it contains 2 and 3 but it does not contain

$$1 = 3 - 2.$$

However, the union of *nested* ideals is an ideal.

Exercise 70. Let $\{I_\lambda\}_{\lambda \in \Lambda}$ be a chain of ideals, meaning that for all $\alpha, \beta \in \Lambda$ we have $I_\alpha \subseteq I_\beta$ or $I_\beta \subseteq I_\alpha$. Show that

$$\bigcup_{\lambda \in \Lambda} I_\lambda$$

is an ideal.

Definition 8.44. Let R be a ring and consider a subset $S \subseteq R$. The **ideal generated by S** , denoted (S) , is the intersection of all the ideals of R that contain S . When $S = \{a_1, \dots, a_n\}$, we may write (a_1, \dots, a_n) instead of $(\{a_1, \dots, a_n\})$.

Remark 8.45. Let S be a subset of a ring R . By Lemma 8.42, the ideal generated by S is indeed an ideal.

The ideal generated by S is the smallest ideal of R that contains S .

Exercise 71. Let A be any subset A of a ring R . The ideal generated by A is given by

$$(A) = \left\{ \sum_{i=1}^n x_i a_i y_i \mid n \geq 0, a_i \in A, x_i, y_i \in R \right\}.$$

If R is commutative and A is any subset, then we can simplify this to

$$(A) = \left\{ \sum_{i=1}^n r_i a_i \mid n \geq 0, r_i \in R, a_i \in A \right\}.$$

Definition 8.46. Let R be a ring. Given an ideal I and a subset S of R , we say that S generates I if $(S) = I$, and we call the elements of S generators of I .

Remark 8.47. Suppose that R is a commutative ring. Given generators for I and J , say

$$I = (S) \quad \text{and} \quad J = (T),$$

the set $\{st \mid s \in S, t \in T\}$ generates IJ , while the set $S \cup T$ generates $I + J$.

Definition 8.48. We say an ideal I is **finitely generated** if $I = (S)$ for some finite subset S of R .

Remark 8.49. Note that if $A = \{a_1, \dots, a_n\}$ and R is commutative, then

$$(a_1, \dots, a_n) = Ra_1 + \cdots + Ra_n = \{r_1 a_1 + \cdots + r_n a_n \mid r_i \in R\}.$$

Definition 8.50. An ideal of R is **principal** if it can be generated by one element, meaning that $I = (a)$ for some $a \in R$.

Example 8.51. In $R = \mathbb{Z}[x]$, we have

$$I = (2, x) = \{2f(x) + xg(x) \mid f(x), g(x) \in \mathbb{Z}[x]\}.$$

Thus I is the set of polynomials with integer coefficients whose constant term is even. One can show that this ideal *cannot* be generated by a single element, so it is not a principal ideal.

We will primarily use this notion when R is commutative.

Remark 8.52. Note that if R is commutative and $I = (a)$, then

$$I = Ra = \{ra \mid r \in R\}$$

by Exercise 71, since an expression of the form

$$r_1 a + \cdots + r_m a$$

can be rewritten as ra with $r = r_1 + \cdots + r_m$. Note, however, that this does not work for noncommutative rings.

Example 8.53.

- (1) We will later show that every ideal of \mathbb{Z} is principal, so all ideals in \mathbb{Z} are of the form $I = (n) = \mathbb{Z} \cdot n$ for some $n \in \mathbb{Z}$.
- (2) We will later show that for any field F , every ideal of $F[x]$ is principal.
- (3) For any field F , every ideal in $F[x_1, \dots, x_n]$ is finitely generated, but not necessarily principal when $n \geq 2$. This fact is the Hilbert Basis Theorem, an elementary result in Commutative Algebra which we will *not* prove in the class.

8.5 Homomorphisms

A homomorphism of rings is a function between two rings that preserves the ring structure: the addition, multiplication, and 1.

Definition 8.54. For rings R and S , a **ring homomorphism** (aka, a **ring map**) from R to S is a function $f: R \rightarrow S$ that satisfies the following properties:

- (1) $f(x + y) = f(x) + f(y)$ for all $x, y \in R$,
- (2) $f(x \cdot y) = f(x) \cdot f(y)$ for all $x, y \in R$, and
- (3) $f(1_R) = 1_S$.

Remark 8.55. Equivalently, f is a ring homomorphisms if f is a homomorphism of abelian groups $(R, +) \longrightarrow (S, +)$ and a *homomorphism of monoids* from (R, \cdot) to (S, \cdot) .³

We really must require $f(1_R) = 1_S$, since this is not a consequence of the first two conditions.

Example 8.56. The map from \mathbb{R} to $\text{Mat}_2(\mathbb{R})$ sending

$$r \mapsto \begin{bmatrix} r & 0 \\ 0 & 0 \end{bmatrix}$$

preserves addition and multiplication, but it does not sent 1 to 1.

Example 8.57. The map $\mathbb{R} \rightarrow \text{Mat}_{n \times n}(\mathbb{R})$ sending r to rI_n is a ring homomorphism.

Exercise 72 (\mathbb{Z} is an initial object in the category of rings). Prove that for any ring S there is a unique ring homomorphism $f: \mathbb{Z} \rightarrow S$ given by sending n to $n \cdot 1_S$.

Example 8.58. Fix a commutative ring R , an element $a \in R$, and an indeterminant x . The evaluation at a map is the function $f: R[x] \rightarrow R$ given by e

$$f\left(\sum_i r_i x^i\right) = \sum_i r_i a^i$$

This is a ring homomorphism.

Exercise 73. Prove that for any commutative ring R and any element $a \in R$, there is a unique ring homomorphism $\mathbb{Z}[x] \rightarrow R$ that sends x to a .

Definition 8.59. Let $f: R \longrightarrow S$ be a ring homomorphism. The **kernel** of f is

$$\ker(f) := \{x \in R \mid f(x) = 0\}.$$

Lemma 8.60. If $f: R \rightarrow S$ is a ring homomorphism, then the following properties hold:

³By definition, a homomorphism of monoids preserves the binary operations and sends the identity to the identity.

- (1) $f(0_R) = 0_S$,
- (2) $f(-x) = -f(x)$,
- (3) If $u \in R^\times$ then $f(u) \in S^\times$ and $f(u^{-1}) = f(u)^{-1}$.
- (4) The image $\text{im}(f)$ is a subring of S .
- (5) The kernel $\ker(f)$ is an ideal of R .
- (6) The map f is injective if and only if $\ker(f) = \{0\}$.

Proof. By definition, f is a homomorphism of additive groups, and thus

$$f(0_R) = 0_S \text{ and } f(-x) = -f(x)$$

are an application of Lemma 1.73.

The fact that units must be sent to units is actually a general property of homomorphisms of monoids. Indeed, since f sends 1 to 1 by assumption, we have

$$1 = f(1) = f(uu^{-1}) = f(u)f(u^{-1})$$

and similarly

$$f(u^{-1})f(u) = f(u^{-1}u) = f(1) = 1.$$

Thus $f(u^{-1}) = f(u)^{-1}$ by the uniqueness of two-sided inverses of units.

To show that the image of f is a subring, first note that $1_S = f(1_R) \in \text{im}(f)$. Moreover, given $a, b \in \text{im}(f)$, say $a = f(x)$ and $b = f(y)$, we have

$$a - b = f(x) - f(y) = f(x - y) \in \text{im}(f) \quad \text{and} \quad ab = f(x)f(y) = f(xy) \in \text{im}(f).$$

By Exercise 60, $\text{im}(f)$ must be a subring of S .

The kernel $\ker(f)$ is already known to be a subgroup under $+$ by Lemma 3.8. Moreover, for $a \in \ker(f)$ and $r \in R$, we have

$$f(ra) = f(r)f(a) = f(r) \cdot 0 = 0,$$

so that $ra \in \ker(f)$ and similarly $ar \in \ker(f)$.

Finally, (7) follows immediately from Lemma 1.78, which is the corresponding fact about group homomorphisms, since f is in particular a homomorphism between the additive groups of R and S . \square

Remark 8.61. In fact, we will later show that a subset I of a ring R is an ideal if and only if it is the kernel of some ring homomorphism with source R .

Definition 8.62. Given rings R and S , a **ring isomorphism** from R to S is a ring homomorphism $f: R \rightarrow S$ such that there exists a ring homomorphism $g: S \rightarrow R$ with

$$f \circ g = \text{id}_S \quad \text{and} \quad g \circ f = \text{id}_R.$$

In that case, we write f^{-1} to denote the homomorphism g . Two rings R and S are isomorphic, written $R \cong S$, if there is an isomorphism from R to S .

Exercise 74. Show that if $f : R \rightarrow S$ is a bijective ring homomorphism, then f is an isomorphism. Moreover, show that the composition of two ring homomorphisms (respectively, isomorphisms) is again a ring homomorphism (respectively, isomorphism).

Exercise 75. Fix a ring R and integer $n \geq 1$. Recall that the collection S of all diagonal matrices in $\text{Mat}_n(R)$ is a subring of $\text{Mat}_n(R)$. Prove that

$$S \cong \underbrace{R \times \cdots \times R}_{n \text{ times}}$$

Exercise 76. Show that the following are ring isomorphism invariants:

- (1) All group isomorphism invariants of the additive group, including the isomorphism class, meaning that if $R \cong S$ then $(R, +) \cong (S, +)$.
- (2) The properties of being commutative, a division ring, a field, or an integral domain.
- (3) The cardinality of the set of zero divisors.
- (4) All group isomorphism invariants of the group of units, including the isomorphism class, that is, if $R \cong S$ then $(R^\times, \cdot) \cong (S^\times, \cdot)$.
- (5) The isomorphism type of the center: if $R \cong S$ then $Z(R) \cong Z(S)$.

Exercise 77. Let $f : R \rightarrow S$ be a ring homomorphism. Show the following:

- (1) Let I be an ideal in R . Then $f(I)$ is an ideal of $f(R)$.
- (2) Let I be an ideal of S . Then $f^{-1}(I)$ is an ideal of R .

Warning! The image of an ideal by a ring homomorphism is however not necessarily an ideal of the target ring.

Example 8.63. Let k be a field and x be an indeterminate. Consider the subring of $S = k[x]$ of polynomials where all the terms have even degree, given by

$$R = k[x^2] := \{r_0 + r_1x^2 + \cdots + r_nx^{2n} \mid r_i \in R\}.$$

The inclusion map $i : R \rightarrow S$ is a ring homomorphism. Moreover, consider the ideal $I = (x^2)$ of R . Its image $J = i(I)$ under i is *not* an ideal of S : for example, because $x^2 \in J$ but $x \cdot x^2 = x^3 \notin J$.

One might however consider the expansion of I into S :

Definition 8.64. Let R and S be commutative rings. Given a ring homomorphism $f : R \rightarrow S$ and an ideal I in R , the **expansion** of I into S is the ideal of S given by $Sf(I)$, sometimes denoted simply by SI .

8.6 Quotient rings

We should think of a two-sided ideal as analogous to a normal subgroup of a group, for two related reasons:

- They are the things that occur as kernels of homomorphisms.
- They are the things you are allowed to mod out by.

Suppose I is a proper ideal of a ring R . Recall this includes the fact that I is a subgroup of $(R, +)$, and hence it is a normal subgroup since $(R, +)$ is abelian. Thus, R/I is an abelian group under $+$. Since we use additive notation, a typical element of this group is of the form $r + I$ for $r \in R$, and

$$a + I = b + I \iff a - b \in I.$$

This quotient group also inherits a ring structure from R :

Theorem 8.65. *If R is a ring and I is a proper (two-sided) ideal, then the binary operation*

$$(r + I) \cdot (s + I) := rs + I$$

on R/I is well-defined and makes $(R/I, +, \cdot)$ into a ring, where $+$ is the operation induced by addition on R . The one in this ring is $1 + I$. Moreover, the map $\pi: R \rightarrow R/I$ with $\pi(r) = r + I$ is a ring homomorphism.

Proof. The main point is the well-definedness of the operation. To show that, suppose

$$r + I = r' + I \quad \text{and} \quad s + I = s' + I.$$

Then $r = r' + a$ and $s = s' + b$ for $a, b \in I$, and hence

$$rs = r's' + r'b + as' + ab.$$

Since I is a two-sided ideal, $r'b$, as' , and ab all belong to I and thus so does their sum. It follows that $rs + I = r's' + I$. This proves that the operation is well-defined.

To show that R/I is a ring, note that we already know it is an abelian group under addition. The fact that multiplication is associative follows from the formula and the fact that multiplication is associative in R . Moreover, from the formula that $1 + I$ is a multiplicative identity, since 1 is one for R . Likewise, the distributive laws are consequences of the distributive laws in R .

To show that π is a group homomorphism, note that

$$\pi(1) = 1 + I$$

by definition, and we already know that π is a group homomorphism, so we only need to prove it preserves products. But indeed, that follows from the definition of the product on R/I . \square

Definition 8.66. The ring R/I with the operations $+$ and \cdot induced from R is the **quotient ring** of R modulo I . The ring homomorphism $\pi: R \rightarrow R/I$ sending r to $r + I$ is called the **canonical surjection**, **canonical map**, or the **quotient map**.

Remark 8.67. In the quotient ring R/I , the zero element is $0 + I$ and the one is $1 + I$.

Example 8.68. Given an ideal $I = (n)$ in the ring \mathbb{Z} , the quotient ring $\mathbb{Z}/(n)$ is the familiar ring \mathbb{Z}/n .

Example 8.69. Let $R = \mathbb{R}[x]$ and $I = (x^2 + 1)$. Then we may form the quotient ring

$$R/I = \mathbb{R}[x]/(x^2 + 1).$$

Intuitively, we are starting with \mathbb{R} , adjoined an element x , and then dictated that $x^2 = -1$, and so we should be getting \mathbb{C} . We will prove this carefully in Example 8.74.

Example 8.70. More generally, let R be any commutative ring, let x be an indeterminant, and suppose $f(x)$ is a monic polynomial, say

$$f(x) = x^n + r_{n-1}x^{n-1} + \cdots + r_1x + r_0$$

for some $r_0, \dots, r_n \in R$. Set $S = R[x]/(f(x))$. One should think of this as adjoining a new ring element \bar{x} to S and imposing the relation given by f :

$$\bar{x}^n = -r_{n-1}\bar{x}^{n-1} + \cdots + r_1\bar{x} + r_0.$$

In fact, the elements of S are in bijective correspondence with the collection of polynomials of degree at most $n - 1$: the function

$$\{a_0 + \cdots + a_{n-1}x^{n-1} \mid a_0, \dots, a_{n-1} \in R\} \longrightarrow S$$

sending g to $g + I$ is a bijection of sets.

For instance, the ring

$$S = \mathbb{Q}[x]/(x^4 + x^3 + x^2 + x + 1)$$

can be thought of taking the ring \mathbb{Q} and adjoining an element ζ_5 such that

$$\zeta_5^4 + \zeta_5^4 + \zeta_5^3 + \zeta_5^2 + \zeta_5 + 1 = 0 \implies -\zeta_5(\zeta_5^4 + \zeta_5^3 + \zeta_5^2 + \zeta_5 + 1) = 1.$$

Thus this new element ζ_5 is invertible; in fact, one can show that S is a field and is isomorphic to $\mathbb{Q}(\zeta_5)$, the smallest subfield of \mathbb{C} containing both \mathbb{Q} and $\zeta_5 = e^{2\pi i/5} \in \mathbb{C}$.

Example 8.71. Many rings of interest in commutative algebra arise from the construction

$$F[x_1, \dots, x_n]/I$$

for some field F , some integer $n \geq 1$, and some ideal I in $F[x_1, \dots, x_n]$. By the Hilbert Basis Theorem, every such ideal is finitely generated, so that such a ring has the form

$$F[x_1, \dots, x_n]/(f_1, \dots, f_m)$$

where each f_j is a polynomial expression in x_1, \dots, x_n . You should think of this as starting with F , adjoining n new elements, and the imposing m relations on these elements. Though keep in mind that in the setting of commutative rings, relations involve both addition and multiplication.

8.7 The Isomorphism Theorems for rings

Theorem 8.72 (Universal Mapping Property for Quotient Rings). *Let R be a ring and I a (two-sided) ideal in R , and let $\pi : R \rightarrow R/I$ be the canonical surjection. If $f : R \rightarrow S$ is a ring homomorphism such that $I \subseteq \ker(f)$, there exists a unique ring homomorphism $\bar{f} : R/I \rightarrow S$ such that the following diagram commutes:*

$$\begin{array}{ccc} R & \xrightarrow{f} & S \\ \pi \downarrow & \nearrow \bar{f} & \\ R/I & & \end{array}$$

meaning that

$$\bar{f} \circ \pi = f.$$

Proof. Ignoring the multiplication operation, we already know from Theorem 4.39 that there is a unique group homomorphism \bar{f} of abelian groups from $(R/I, +)$ to $(S, +)$ such that

$$\bar{f} \circ \pi = f.$$

It remains only to check that \bar{f} preserves multiplication and sends 1 to 1. Given elements $r + I, s + I \in R/I$, we have

$$\bar{f}((r + I)(s + I)) = \bar{f}(rs + I) = f(rs) = f(r)f(s) = f(r + I)f(s + I),$$

since f preserves multiplication. Finally,

$$\bar{f}(1_{R/I}) = \bar{f}(1_R + I) = f(1_R) = 1_S$$

since f sends 1_R to 1_S . □

Theorem 8.73 (First Isomorphism Theorem for Rings). *If $f : R \rightarrow S$ is a ring homomorphism, there is an isomorphism*

$$\begin{aligned} \bar{f} : R/\ker(f) &\xrightarrow{\cong} \text{im}(f) \\ r + \ker(f) &\longmapsto f(r). \end{aligned}$$

In particular, if f is surjective, then

$$R/\ker(f) \cong S.$$

Proof. Taking $I = \ker(f)$ in the [UMP for quotient rings](#), we have a ring homomorphism $\bar{f} : R/\ker(f) \rightarrow S$. By the formula for \bar{f} we immediately get that $\text{im}(\bar{f}) = \text{im}(f)$. Its kernel is

$$\{r + I \mid f(r) = 0\} = \{0_{R/I}\}$$

and hence \bar{f} is injective. The result follows. □

Here is a nice application of the [First Isomorphism Theorem](#):

Example 8.74. Recall that $\mathbb{R}[x]/(x^2 + 1)$ ought to be \mathbb{C} . To prove this, we define a map

$$\phi: \mathbb{R}[x] \longrightarrow \mathbb{C}$$

sending $f(x)$ to $f(i)$, the evaluation of f at i . It is easy to check ϕ is a ring homomorphism, but we leave the details as an exercise. This map is surjective since elements of the form $a + bx$ in the source map to all possible complex numbers under ϕ .

We claim the kernel of ϕ is $(x^2 + 1)$. Note that

$$x^2 + 1 \in \ker(\phi)$$

and it follows that

$$(x^2 + 1) \subseteq \ker(\phi),$$

since $\ker(\phi)$ is a two-sided ideal.

Suppose $\phi(f(x)) = 0$. By the Division Algorithm in the polynomial ring $\mathbb{R}[x]$, which we will cover in more detail later, we can write

$$f(x) = (x^2 + 1)q(x) + r(x)$$

with the degree of $r(x)$ at most 1. So $r(x) = a + bx$ for real numbers a and b . If $r(x) \neq 0$, so that at least one of a or b is nonzero, then

$$r(i) = a + bi \neq 0$$

since a complex number is 0 only if both components are, which would contradict the fact that $f(i) = 0$. So we must have $r(x) = 0$ and hence $f(x) \in (x^2 + 1)$.

Applying the [First Isomorphism Theorem for rings](#), we get

$$\mathbb{R}[x]/(x^2 + 1) \cong \mathbb{C}$$

via the map sending $f(x) + (x^2 + 1)$ to $f(i)$.

Example 8.75. Similarly, we may define $\phi: \mathbb{Q}[x] \rightarrow \mathbb{C}$ by $\phi(p(x)) = p(\zeta_5)$. We will skip the details, but its image of $\mathbb{Q}(\zeta_5)$ and its kernel is $(x^4 + x^3 + x^2 + x + 1)$ and hence we declared in Example 8.70 $\mathbb{Q}[x]/(x^4 + x^3 + x^2 + x + 1) \cong \mathbb{Q}(\zeta_5)$.

Exercise 78. Let S be a subring of a ring R and let I be an ideal of R . Show that Show that

$$S + I = \{s + i \mid s \in S, i \in I\}$$

is a subring of R and $S \cap I$ is an ideal of S .

Theorem 8.76 (Second Isomorphism Theorem for rings). *Let S be a subring of a ring R and let I be an ideal of R . Then*

$$S + I = \{s + i \mid s \in S, i \in I\}$$

is a subring of R , $S \cap I$ is an ideal of S , and

$$\frac{S + I}{I} \cong \frac{S}{S \cap I}.$$

Proof. The first two facts are Exercise 78. The map $f : S + I \rightarrow \frac{S}{S \cap I}$ sending $s + i$ to $s + i + I = s + I$ is a homomorphism of rings since it is the composition of a subring inclusion with the canonical quotient map. It is surjective by definition, and the kernel is

$$\ker(f) = \{s + i \mid s \in S, i \in I, s + I = I\} = I.$$

The result now follows from the First Isomorphism Theorem for rings. \square

Theorem 8.77 (Third Isomorphism Theorem for rings). *If R is a ring and $I \subseteq J$ are two ideals of R , then J/I is an ideal of R/I and*

$$\frac{R/I}{J/I} \cong R/J \quad \text{via} \quad (r + I) + J/I \mapsto r + J.$$

Proof. If we ignore multiplication, we know that $(J/I, +)$ is a subgroup of $(R/I, +)$ and that there is an isomorphism of abelian groups

$$(R/I)/(J/I) \cong R/J$$

given by

$$(r + I) + J/I \mapsto r + J.$$

One just needs to check that J/I is a two-sided ideal of R/I and the indicated bijection preserves multiplication, which we leave as an elementary exercise. \square

The following will be helpful in discussing some interesting examples:

Exercise 79 (Reduction homomorphism). Given a ring map $\phi: R \rightarrow S$ between commutative rings, there is an induced ring map

$$\rho: R[x] \rightarrow S[x] \quad \text{given by} \quad \rho\left(\sum_i r_i x^i\right) = \sum_i \phi(r_i)x^i.$$

That is, ρ consists of applying ϕ to the coefficients of each polynomials.

The proof is just a tedious check of the axioms, and so we leave it as an exercise.

Example 8.78. In particular, for I an ideal of R , taking $S = R/I$ and ϕ to be the canonical homomorphism, Exercise 79 implies that there is a ring homomorphism

$$\rho: R[x] \rightarrow \frac{R}{I}[x]$$

given by

$$\rho\left(\sum_i r_i x^i\right) = \sum_i (r_i + I)x^i$$

Thus ρ is given by modding out the coefficients by I . In this case, the kernel of ρ is the collection of polynomials with coefficient in I , which we denote by $I[x]$. By the First Isomorphism Theorem, we conclude that

$$\frac{R[x]}{I[x]} \cong \frac{R}{I}[x].$$

Example 8.79. Consider the ideal $J = (2, x^2 + x + 1)$ of $\mathbb{Z}[x]$. Explicitly, by Exercise 71 we have

$$J = \{p(x) \cdot 2 + q(x)(x^2 + x + 1) \mid p(x), q(x) \in \mathbb{Z}[x]\}.$$

Suppose we want to understand $\mathbb{Z}[x]/J$. Then the [Third Isomorphism Theorem](#) is our friend. Set $I = (2) = \mathbb{Z}[x] \cdot 2$ and note that $I \subseteq J$, and so by the [Third Isomorphism Theorem](#) we have

$$\frac{\mathbb{Z}[x]}{J} \cong \frac{\mathbb{Z}[x]/I}{J/I}.$$

By the example above,

$$\frac{\mathbb{Z}[x]}{I} \cong \frac{\mathbb{Z}}{2}[x].$$

As we did for groups, we will write J/I to denote the image of J under the quotient map $\pi : \mathbb{Z}[x] \rightarrow \mathbb{Z}[x]/I$. Since J is generated by 2 and $x^2 + x + 1$ and I is generated by 2, one can show that $J/(2)$ is the principal ideal of $\mathbb{Z}[x]/(2)$ generated by the coset represented by $x^2 + x + 1$. Under the identification

$$\mathbb{Z}[x]/(2) \cong (\mathbb{Z}/2)[x],$$

this ideal $J/(2)$ corresponds to the principal ideal of $(\mathbb{Z}/2)[x]$ generated by $x^2 + x + 1 \in (\mathbb{Z}/2)[x]$. We obtain a ring isomorphism

$$\mathbb{Z}[x]/J \cong \frac{(\mathbb{Z}/2)[x]}{(x^2 + x + 1)}.$$

Looking ahead a bit, we note that the quadratic polynomial $x^2 + x + 1$ has no roots in the field $\mathbb{Z}/2$, as the only possibilities are 0 and 1, and neither is a root. As we will prove in soon, this implies $(\mathbb{Z}/2)[x]/(x^2 + x + 1)$ is a field, and thus $\mathbb{Z}[x]/J$ is a field.

As discussed before Lemma 8.42, the set of all ideals in a ring R is a partially ordered set with respect to the order given by containment.

Theorem 8.80 (Lattice Theorem for Quotient Rings). *Suppose R is a ring and I is a two-sided ideal of R , and write $\pi : R \rightarrow R/I$ for the quotient map. There is a bijection*

$$\begin{aligned} \{ \text{ideals of } R \text{ containing } I \} &\longleftrightarrow \{ \text{ideals of } R/I \}. \\ J &\longmapsto \pi(J) = J/I \\ \pi^{-1}(L) &\longleftarrow L \end{aligned}$$

Proof. By Theorem 4.51, we know that there is a bijection of subgroups (under \oplus) of R that contain I and subgroups of R/I , given by these formulas. It remains to prove that this correspondence preserves the property of being an ideal, which we leave as an exercise. \square

Example 8.81. We claimed in Example 8.79 that $\mathbb{Z}[x]/(2, x^2 + x + 1)$ is a field. Since a field has only two ideals, $\{0\}$ and the field itself, we deduce, using the [Lattice Isomorphism Theorem](#), that there are only two ideals in $\mathbb{Z}[x]$ that contain $(2, x^2 + x + 1)$, namely

$$(2, x^2 + x + 1) = \pi^{-1}(0) \quad \text{and} \quad \mathbb{Z}[x] = \pi^{-1}(F).$$

8.8 Prime and maximal ideals in commutative rings

Definition 8.82. A **maximal ideal** of a ring R is an ideal that is maximal with respect to containment among all *proper* ideals of R . More precisely, an ideal M is maximal if $M \neq R$ and for all ideals I in R ,

$$M \subseteq I \implies M = I \text{ or } I = R.$$

Thus the only ideals of R containing M are M and R .

Let R be a commutative ring. A **prime ideal** of R is a *proper* ideal P such that

$$xy \in P \implies x \in P \text{ or } y \in P.$$

Example 8.83. In \mathbb{Z} , the prime ideals are (0) and the ideals generated by prime integers $P = (p)$, where p is a prime integer. The maximal ideals are the ideals generated by prime integers. In particular, (0) is prime but not maximal.

Example 8.84. In $\mathbb{Z}[i]$, we claim that the ideal (13) is not prime. On the one hand,

$$13 = (3 + 2i)(3 - 2i) \in (13)$$

but we claim that

$$3 + 2i \notin (13) \quad \text{and} \quad 3 - 2i \notin (13).$$

To see this, let N be the square of the complex norm function, meaning that $N(a + bi) = a^2 + b^2$ for any $a, b \in \mathbb{R}$. Now note that if $3 \pm 2i = 13\alpha$ for some $\alpha \in \mathbb{Z}[i]$, then

$$N(3 \pm 2i) = N(13)N(\alpha),$$

so it would follow that

$$13 = N(3 \pm 2i) = 13^2 N(\alpha)$$

with $N(\alpha) \in \mathbb{Z}$, which is impossible.

Theorem 8.85. Let R be a commutative ring and let Q be an ideal of R .

- (1) The ideal Q is maximal if and only if R/Q is a field.
- (2) The ideal Q is prime if and only if R/Q is a domain.
- (3) Every maximal ideal of R is prime.

Proof. By the [Lattice Isomorphism Theorem](#), the ideals of R/Q are of the form I/Q , where I is an ideal in R containing Q .

By Exercise 67, R/Q is a field if and only if R/Q has only two ideals, $\{0\} = Q/Q$ and R/Q . Thus R/Q is a field if and only if the only ideals that contain Q are Q and R .

Now suppose Q is prime. If

$$(r + I)(r' + I) = 0 + I,$$

then $rr' \in I$ and hence either $r \in I$ or $r' \in I$, so that either

$$r + I = 0 \quad \text{or} \quad r' + I = 0.$$

Since R is commutative, then R/I is also commutative, and since Q is a proper, then R/I is not the zero ring. This proves that R/Q is a domain.

Conversely, suppose that R/Q is a domain. Since R/Q is not the zero ring, Q is proper. If $x, y \in R$ satisfy $xy \in I$, then

$$(x + I)(y + I) = 0$$

in R/Q , and hence either $x + Q = 0$ or $y + Q = 0$. It follows $x \in Q$ or $y \in Q$. This proves that Q is prime.

If Q is maximal, then R/Q is a field, which in particular implies that R/Q is a domain, and thus Q is prime. \square

Exercise 80. Show that the ideal $(2, x)$ in $\mathbb{Z}[x]$ is maximal (and thus prime). In contrast, the ideals (2) and (x) are prime but not maximal.

Example 8.86. For a field F , the ideal $I = (x_1 - a_1, \dots, x_n - a_n)$ of the polynomial ring $F[x_1, \dots, x_n]$ is maximal. This holds because I is the kernel of the surjective ring homomorphism $F[x_1, \dots, x_n] \rightarrow F$ given by evaluating polynomials at (a_1, \dots, a_n) .

Exercise 81. Show that $f : R \rightarrow S$ is a ring homomorphism and S is a domain, then $\ker(f)$ is a prime ideal.

Theorem 8.87. *Every commutative ring has a maximal ideal.*

Fun fact: this is actually *equivalent* to the Axiom of Choice. We will prove it (but not its equivalence to the Axiom of Choice!) using Zorn's Lemma, another equivalent version of the Axiom of Choice. Zorn's Lemma is a statement about partially ordered sets. Given a partially ordered set S , a chain in S is a totally ordered subset of S .

Theorem 8.88 (Zorn's Lemma). *Let S be a nonempty partially ordered set S such that every chain in S has an upper bound in S . Then S contains at least one maximal element.*

We can now prove every ring has a maximal ideal; in fact, we will prove something stronger:

Theorem 8.89. *Given a commutative ring R , every proper ideal $I \neq R$ is contained in some maximal ideal.*

Chapter 9

Nice domains

In this chapter, all rings in this chapter are commutative. We will introduce three special classes of domains: Euclidean domains, PIDs, and UFDs. We will also show that

$$\text{Fields} \subsetneq \text{Euclidean Domains} \subsetneq \text{PIDs} \subsetneq \text{UFDs} \subsetneq \text{Domains}.$$

9.1 Euclidean domains

An Euclidean domain is a domain with some additional structure, designed to mimic the parallel facts that there is a notion of division with remainder in both \mathbb{Z} and $F[x]$, with F a field.

Definition 9.1. An **Euclidean domain** is an integral domain R together with a function

$$N: R \setminus \{0\} \rightarrow \mathbb{Z}_{\geq 0}$$

satisfying the following property: for any two elements $a, b \in R$ with $b \neq 0$, there are elements q and r of R such that

$$a = bq + r \text{ and } r = 0 \text{ or } N(r) < N(b).$$

The function N is an **Euclidean function** for R . If N satisfies $N(ab) = N(a)N(b)$, then N is called a **norm function**.

One sometimes says that an Euclidean domain has a division algorithm, but that is misleading: there need not be an algorithm to find q and r given a and b , and neither q nor r need to be unique. Finally, the Euclidean function N is *not* required to satisfy any sort of multiplicative property, but in some examples it does, and in those examples it is called a norm function.

Example 9.2. A degenerate example of an Euclidean domain is a field F equipped with the trivial norm $N(x) = 0$ for all $x \neq 0$, or really any function $N: F \setminus \{0\} \rightarrow \mathbb{Z}_{\geq 0}$. Indeed, given $a, b \in F$ with $b \neq 0$, we have

$$a = b(ab^{-1}) + 0,$$

thus $q = ab^{-1}$ and $r = 0$ satisfy the definition.

This calculation shows, more generally, that if b is a unit, then for all a there exists an equation $a = bq + r$ with $r = 0$, not matter what N we use.

The canonical example of an Euclidean domain is \mathbb{Z} .

Theorem 9.3 (Division Algorithm for \mathbb{Z}). *For any two integers a, b with $b \neq 0$, there are (unique) integers q and r such that*

$$a = qb + r \quad \text{and} \quad 0 \leq r < |b|.$$

Example 9.4. Let $R = \mathbb{Z}$ with $N(m) = |m|$ for all $m \neq 0$. This ring is an Euclidean Domain because of the familiar [Division Algorithm for integers](#). Notice however that the [Division Algorithm](#) gives us something stronger: if we add in the additional requirement that when dividing a by b the remainder r must satisfy $0 \leq r < |b|$, then that remainder is unique.

However, this uniqueness is not part of the abstract theory since it does not generalize to all cases well. And in fact, even in this case there is no uniqueness: following only the definition, we have nonunique remainders, as for example when $a = 12$ and $b = 5$, then both

$$12 = 2 \cdot 5 + 2 \quad \text{and} \quad 12 = 3 \cdot 5 + (-3)$$

are equally acceptable, since $|-3| < 5$.

Definition 9.5. Let R be a commutative ring with $1 \neq 0$. Consider a nonzero polynomial

$$f = \sum_{i=0}^n a_i x^i \in R[x]$$

with $a_n \neq 0$. The **degree** of f is $\deg(f) = n$, and the **leading coefficient** of f is $\text{lc}(f) = a_n$. The 0 polynomial does not have a degree nor a leading coefficient.

Lemma 9.6. *Let R be an integral domain and $f, g \in R[x]$ be nonzero polynomials. Then:*

- (1) *The product fg is nonzero and $\text{lc}(f \cdot g) = \text{lc}(f) \cdot \text{lc}(g)$. In particular, $R[x]$ is a domain.*
- (2) *We have $\deg(fg) = \deg(f) + \deg(g)$.*
- (3) *The units of $R[x]$ are the constant polynomials given by units of R : $R[x]^\times = R^\times$.*

Proof. If $f = a_n x^n + \text{lower order terms}$, with $a_n \neq 0$ and $g = b_m x^m + \text{lower order terms}$ with $b_m \neq 0$, then $fg = a_n b_m x^{m+n} + \text{lower order terms}$. Since R is a domain, $a_n b_m \neq 0$, so

$$fg \neq 0, \quad \text{lc}(f \cdot g) = \text{lc}(f) \cdot \text{lc}(g), \quad \text{and} \quad \deg(fg) = \deg(f) + \deg(g).$$

If $r \in R$ is a unit, then the constant polynomial r is also a unit in $R[x]$. Conversely, suppose that $f \in R[x]^\times$ has inverse g . Then

$$0 = \deg(1) = \deg(fg) = \deg(f) + \deg(g) \implies \deg(f) = \deg(g) = 0. \quad \square$$

Remark 9.7. If R is a commutative ring and $\text{lc}(f)$ is a nonzerodivisor, then equality in part (1) and the conclusion of part (2) above hold by the same proof.

Corollary 9.8. *If F is a field, then $f \in F[x]$ is a unit if and only if $f \neq 0$ and $\deg(f) = 0$.*

There is also a well-known Division Algorithm for polynomials in one variable.

Theorem 9.9 (Division Algorithm for polynomials). *Let A be a commutative ring, and consider $R = A[x]$. Given polynomials f and g in $A[x]$ such that $\text{lc}(g) \in A^\times$, there exist unique polynomials q and r such that*

$$f = gq + r \quad \text{and} \quad r = 0 \text{ or } \deg(r) < \deg(g).$$

Proof. We first show existence. Fix f and $g \neq 0$. If $\deg(g) = 0$, then g is a unit, so consider $q = g^{-1}f$ and $r = 0$, and note that

$$f = g(g^{-1}f) = qf + r.$$

Now when $\deg(g) > 0$, let $g = a_nx^n + \text{lower order terms}$, with $a_n \in A^\times$ and $n > 0$. If $f = 0$, then $q = r = 0$ works, so we might as well assume $f = b_mx^m + \text{lower order terms}$, with $b_m \neq 0$ and $m \geq 0$. We proceed by complete induction on $m = \deg(f)$. If $m < n$, we may take $q = 0$ and $r = f$. Assume $m \geq n$, and consider

$$h := f - g \cdot (b_m/a_m)x^{m-n} = (b_m - a_m(b_m/a_m))x^m + \text{lower order terms}.$$

We have $\deg(h) < m$, and thus by induction, $h = g \cdot q' + r$ with $r = 0$ or $\deg(r) < \deg(g)$. Thus

$$f = h + g \cdot (b_m/a_m)x^{m-n} = g \cdot q' + r' + g \cdot (b_m/a_m)x^{m-n} = gq + r$$

where $q = q' + (b_m/a_m)x^{m-n}$.

For uniqueness, if $qg + r = f = q'g + r'$ with q, r and q', r' as in the statement, we have $(q - q')g = (r - r')$. If $r - r' \neq 0$, then this is a polynomial of degree less than n , but by the remark above, if $(q - q')g \neq 0$, its degree must be at least n . This is a contradiction, so $r - r' = 0$, so $r = r'$, and $(q - q')g = 0$. Again by the remark, this implies $q - q' = 0$ so $q = q'$. \square

Corollary 9.10. *Given a field F , $F[x]$ is an Euclidean domain. In particular, the function $N: F[x] \setminus \{0\} \rightarrow \mathbb{Z}_{\geq 0}$ given by $N(f(x)) := \deg(f(x))$ is an Euclidean function.*

Proof. Since F is a field, for any nonzero polynomial in $F[x]$, the leading coefficient is a unit. Apply the Division Algorithm for polynomials. \square

Corollary 9.11. *Let F be a field and let $f = a_nx^n + \dots + a_0 \in F[x]$ be a polynomial with $a_n \neq 0$ and $n \geq 1$. Then every nonzero element of $F[x]/(f)$ is of the form $g + (f)$ for some polynomial g of degree $\deg(g) < n$. Moreover, if g and h are distinct polynomials in $F[x]$ of degree strictly less than n , then $g + (f) \neq h + (f)$.*

Thus, there is a bijection

$$\begin{array}{ccc} \frac{F[x]}{(f)} & \longleftrightarrow & \{\text{polynomials of degree } < n\} \\ g + (f) & \longleftarrow & g \end{array}$$

Proof. First, consider any $g \in F[x]$. By the Division Algorithm, we can find polynomials q and r such that $g = qf + r$ and $r = 0$ or $\deg(r) < \deg(f) = n$. Then

$$g + (f) = qf + r + (f) = r + (f).$$

On the other hand, consider any nonzero element $g + (f) = h + (f)$ in $F[f]/(f)$, so that in particular $f, g \neq 0$, and assume that $\deg(g), \deg(h) < n$. Then $g - h \in (f)$. By Lemma 9.6, the degree of any nonzero multiple of f is at least n , but $\deg(g - h) < n$. We conclude that $g - h$ must be zero, so that $g = h$. \square

Theorem 9.12. *The ring $R = \mathbb{Z}[i]$ of Gaussian integers is a Euclidean domain with N the usual complex (Euclidean) square norm $N(a + bi) = a^2 + b^2$.*

Proof. Let $\alpha, \beta \in \mathbb{Z}[i]$. Note that

$$\mathbb{Z}[i] \subseteq \mathbb{Q}(i) = \{a + bi \mid a, b \in \mathbb{Q}\},$$

and consider

$$\frac{\alpha}{\beta} = p + qi \in \mathbb{Q}(i).$$

Now pick $s, t \in \mathbb{Z}$ so that $|p - s| \leq \frac{1}{2}$ and $|q - t| \leq \frac{1}{2}$. We have

$$\alpha = \beta(s + ti) + \beta(p + qi) - \beta(s + ti).$$

Set $q = s + ti \in \mathbb{Z}[x]$, and

$$r = \beta(p + qi) - \beta(s + ti) = \beta(s + ti - (p + qi)) \in \mathbb{Z}[i].$$

Moreover, note that

$$\alpha = \beta(s + ti) + r.$$

If $r = 0$, then we are done. If $r \neq 0$, we need to check that $N(r) < N(\beta)$. Using that N is multiplicative, the Pythagorean Theorem, and the choice for s, t , we have

$$N(r) = N(\beta(s + ti - (p + qi))) = N(\beta)N(s + ti - (p + qi)) \leq N(\beta) \cdot \left(\frac{1}{4} + \frac{1}{4}\right) < N(\beta).$$

Thus the norm function N makes $\mathbb{Z}[i]$ into a Euclidean domain. \square

9.2 Principal ideal domains (PIDs)

One of the key features of Euclidean domains is that they are examples of PIDs:

Definition 9.13. A **principal ideal domain**, often shortened to **PID**, is a domain R where all ideals are principal, meaning that for every ideal I there exists $a \in R$ such that $I = (a)$.

Theorem 9.14. *Every Euclidean domain is a PID.*

Proof. Let N be a norm function making R into a Euclidean domain, and fix an ideal I in R . If I is the zero ideal, then $I = (0)$ is principal. Otherwise, pick a nonzero element $b \in I$ with $N(b)$ as small as possible. Note that such b exists by the Well-Ordering Principle. We claim that $I = (b)$. On the one hand, since $b \in I$ then $(b) \subseteq I$. On the other hand, given $a \in I$,

$$a = bq + r$$

and either $r = 0$ or $N(r) < N(b)$. But note that $r = a - bq \in I$, and we cannot have both $r \neq 0$ and $N(r) < N(b)$ since b was chosen to have smallest possible norm among elements of I . So it must be that $r = 0$, and hence $a \in (b)$. \square

Corollary 9.15. *Let F be a field. The rings \mathbb{Z} , $\mathbb{Z}[i]$, and $F[x]$ are all PIDs.*

Proof. As we saw in the previous section, all of these rings are Euclidean domains: the fact that \mathbb{Z} is an Euclidean domain is Example 9.4; Theorem 9.12 says that $\mathbb{Z}[i]$ is an Euclidean domain; and Corollary 9.10 says that $F[x]$ is an Euclidean domain. \square

Exercise 82. Show that $\mathbb{Z}[\sqrt{-2}]$ is a PID.

Example 9.16. The ring $\mathbb{Z}[x]$ is not a Euclidean domain. This follows from Theorem 9.14, since $\mathbb{Z}[x]$ is not a PID — for example, the ideal $(2, x)$ is not principal. Similarly, the ring $F[x, y]$ is not a Euclidean domain since it is not a PID (e.g., (x, y) is not principal).

The converse of Theorem 9.14 is false:

Example 9.17 (A PID that is not an Euclidean domain). The ring

$$R = \mathbb{Z}\left[\frac{1 + \sqrt{-19}}{2}\right]$$

is a PID that is not a Euclidean domain.

Definition 9.18. Let R be a commutative ring and let $a, b \in R$.

- The element b is a **divisor** of a , and a is a **multiple** of b , written $b \mid a$, if there is an element $x \in R$ with $a = bx$. Equivalently, $b \mid a$ iff $a \in (b)$.
- We say a and b are **associates** if $a = ub$ for some unit $u \in R$. Note that this condition is symmetric, since if $a = ub$ then $b = u^{-1}a$ and u^{-1} is also a unit.
- A **greatest common divisor**, or **gcd**, of a and b is an element $d \in R$ satisfying $d \mid a$, $d \mid b$, and

$$e \mid a \quad \text{and} \quad e \mid b \quad \implies \quad e \mid d.$$

- A **least common multiple**, or **lcm**, of a and b is an element $m \in R$ satisfying $a \mid m$, $b \mid m$, and whenever $a \mid m'$ and $b \mid m'$ then $m \mid m'$.

Lemma 9.19. *Assume R is a domain and $x, y \in R$. The following are equivalent:*

(1) x and y are associates,

(2) $(x) = (y)$, and

(3) x and y divide each other, meaning that $x \mid y$ and $y \mid x$.

Proof. The equivalence of the latter two is clear (and does not require that R be a domain), since $x \mid y$ if and only if $y \in (x)$ if and only if $(y) \subseteq (x)$.

Assume (3) holds. Then $x \in (y)$ and so $x = yu$ for some $u \in R$. Similarly $y = xs$ and hence $y = yus$, which implies $y(1 - us) = 0$. Since R is a domain, either $y = 0$ or $su = 1$. If $y = 0$, then $x = yu = 0 = y$. If $y \neq 0$ then u is a unit (with inverse s).

Conversely, suppose (1) holds, so that $x = uy$ for some unit u . Then $y \mid x$, and since we also have $y = u^{-1}x$, it follows that $x \mid y$. \square

Remark 9.20. Greatest common divisors and least common multiples are not uniquely defined. For example, in \mathbb{Z} , both 2 and -2 are greatest common divisors of 4 and 6. But, at least in a domain, they are unique up to associates. That is, if g and g' are both gcds of the same pair of elements in a domain R , then g and g' are associates, and similarly for lcms. This follows from Lemma 9.19 since, by definition, g and g' would have to divide each other.

Gcds (and lcms) need not exist, in general, but here is a situation in which they do:

Lemma 9.21. *If R is a PID and $a, b \in R$, then $(a, b) = (g)$ for some $g \in R$, and any such g is a gcd of a and b .*

Proof. The existence of g is granted by the definition in a PID: the ideal (a, b) must be principal. Now since $a, b \in (g)$, we have $g \mid a$ and $g \mid b$, so g is a common divisor of a and b . Given any other h such that $h \mid a$ and $h \mid b$, we have $a, b \in (h)$, so $(g) = (a, b) \subseteq (h)$ since (h) is an ideal. As a consequence, $g \in (h)$, and hence $h \mid g$. We conclude that g is a greatest common divisor of a and b . \square

Remark 9.22. Let R be a PID. Using Lemma 9.19 we may describe all the ideals that contain a given ideal $(a) \subseteq R$: they are given by the collection of divisors of a up to associates. For instance, in $\mathbb{Q}[x]$ there are 8 ideals that contain $(x^4 - 1)$, since

$$x^4 - 1 = (x^2 + 1)(x - 1)(x + 1)$$

has 8 divisors (including 1 and $x^4 - 1$ itself).

Remark 9.23. If R is not only a PID but also an Euclidean domain, then the Euclidean algorithm can be used to compute a gcd of any two nonzero $a, b \in R$. This is slightly misleading, since the “division algorithm” in the definition of an Euclidean domain is not really an algorithm. But for \mathbb{Z} and $F[x]$ it is truly an algorithm, and you probably used it to find gcds before in your life.

Definition 9.24. Let R be a domain.

(1) An element $p \in R$ is a **prime** element if $p \neq 0$, p is not a unit, and

$$p \mid ab \implies p \mid a \text{ or } p \mid b.$$

(2) An element $r \in R$ is **irreducible** if $r \neq 0$, r is not a unit, and for all $x, y \in R$

$$r = xy \implies x \text{ is a unit or } y \text{ is a unit.}$$

Remark 9.25. The condition that a nonzero nonunit element $p \in R$ is a prime element can be rephrased as follows:

$$ab \in (p) \implies a \in (p) \text{ or } b \in (p).$$

That is, p is a prime element if and only if (p) is a nonzero prime ideal.

Example 9.26.

- (1) The prime elements of \mathbb{Z} are the prime integers (where we allow both positive and negative primes); these are also the irreducible elements.
- (2) Any element $a \in \mathbb{Z}[i]$ with $N(a)$ a prime integer is irreducible (exercise!). For example, $1 + 2i$ is irreducible.
- (3) The element $13 = (2 + 3i)(2 - 3i)$ is not irreducible in $\mathbb{Z}[i]$.
- (4) We claim that the polynomial $x^2 + x + 1 \in (\mathbb{Z}/2)[x]$ is irreducible. Indeed, if it factors nontrivially, it must factor as a product of two linear polynomials, say

$$x^2 + x + [1] = (x + [a])(x + [b]).$$

Then $-[b]$ is a root for $x^2 + x + [1]$. But neither $[0]$ nor $[1]$ are roots for this polynomial, which is a contradiction.

Theorem 9.27. Let R be a domain and let $r \in R$.

- (1) If r is a prime element, then r is irreducible.
- (2) Assume R is a PID. The following are equivalent:
 - (a) r is prime,
 - (b) r is irreducible, and
 - (c) the ideal (r) generated by r is a maximal ideal.

Proof. Suppose R is a domain and that r is prime. Then by definition $r \neq 0$ and r is not a unit. Suppose $r = yz$. Then $yz \in (r)$ and hence by definition either $y \in (r)$ or $z \in (r)$. If $y \in (r)$, we have $y = rt$ for some t and so $y = yzt$. Since $r \neq 0$, $y \neq 0$, and R is a domain, we must have $zt = 1$, showing that z is a unit.

Assume R is a PID. We just showed that (a) implies (b). To show that (b) implies (c), assume r is irreducible. Then by definition r is not a unit, and hence (r) is a proper ideal. This ideal is therefore contained in some maximal ideal M by Theorem 8.89. We will show that $(r) = M$, and hence (r) is a maximal ideal. Since R is a PID, all ideals are principal, and thus $M = (y)$ for some y , so that $r = yt$ for some t . But r is irreducible and y is not a unit, which forces t to be a unit and hence $(r) = (y) = M$.

Finally, (c) implies (a) since, by Theorem 8.85, all maximal ideals are prime. In particular, (r) is a prime ideal and hence r is a prime element. \square

Corollary 9.28. In any PID, every nonzero prime ideal is maximal.

Proof. Let Q be a nonzero prime ideal in the PID R . Since R is a PID, $Q = (r)$ for some nonzero element $r \in R$, and in particular r is a prime element. By Theorem 9.27, $Q = (r)$ must be a maximal ideal. \square

Example 9.29. Let F be a field and let $p \in F[x]$ be a nonzero polynomial. Since $F[x]$ is a PID, by Corollary 9.28 the quotient $F[x]/(p)$ is a field if and only if p is irreducible.

If p is quadratic, then it is irreducible if and only if it has no roots. For example, we deduce from these observations that the ring $(\mathbb{Z}/2)[x]/(x^2+x+1)$ is a field, which we claimed in Example 8.79.

9.3 Unique factorization domains (UFDs)

Definition 9.30. A ring R is called a **unique factorization domain**, or **UFD** for short, if R is an integral domain and the following hold:

- (1) For every nonzero element $r \in R$ we have

$$r = up_1 \cdots p_n$$

for some unit u , some integer $n \geq 0$, and some (not necessarily distinct) irreducible elements $p_1, \dots, p_n \in R$.

- (2) Such factorizations are unique up to ordering and associates: if

$$r = vq_1 \cdots q_m$$

is another such factorization with v a unit and each q_i irreducible, then $m = n$ and there is a permutation σ such that, for all i , the elements p_i and $q_{\sigma(i)}$ are associates.

Remark 9.31. Note that units admit irreducible factorizations according to this definition by taking $n = 0$.

Example 9.32. (1) The ring \mathbb{Z} is a UFD by the Fundamental Theorem of Arithmetic.

- (2) Given a field F , $F[x]$ is a UFD: $F[x]$ is an Euclidean domain and we will soon show that all Euclidean domains are UFDs.
- (3) It follows that $F[x_1, \dots, x_n]$ is a UFD for all n . Note that if $n > 1$, this ring is not a PID and hence not a Euclidean domain.

Theorem 9.33. If R is a UFD, then $R[x]$ is also a UFD.

We will give a proof of this theorem later, time permitting.

Example 9.34 (A UFD that is not a PID). Let F be a field and fix an integer $n \geq 1$. Since $F[x]$ is a UFD, by applying Theorem 9.33 repeatedly we conclude that $F[x_1, \dots, x_n]$ is also a UFD. However, $F[x_1, \dots, x_n]$ is not a PID when $n > 1$, as one can show that (x_1, \dots, x_n) is not a principal ideal.

Example 9.35 (Another UFD that is not a PID). The ideal $(2, x)$ in $\mathbb{Z}[x]$ is not principal. Thus $\mathbb{Z}[x]$ is not a PID, and therefore it is also not an Euclidean domain. On the other hand, \mathbb{Z} is a UFD and thus by Theorem 9.33 $\mathbb{Z}[x]$ must also be a UFD.

Example 9.36 (A domain that is not a UFD). We claim that the ring $\mathbb{Z}[\sqrt{-5}]$ is a domain that is *not* a UFD. Note that

$$6 = (1 + \sqrt{-5})(1 - \sqrt{-5}) = 2 \cdot 3,$$

and one can show that each of $1 + \sqrt{-5}$, $1 - \sqrt{-5}$, 2, and 3 are irreducible by checking their norms (exercise!). Moreover, recall the only units in this ring are ± 1 , so these elements are not associates of each other.

Notice also that $\mathbb{Z}[\sqrt{-5}]$ contains elements that are irreducible but not prime: for example, 2 is irreducible but not prime. Compare with Theorem 9.37 below.

Exercise 83. Let R be a UFD. Given $a, b \in R$, let

$$a = up_1^{e_1} \cdots p_m^{e_m} \quad \text{and} \quad b = vp_1^{f_1} \cdots p_m^{f_m}$$

for irreducible elements p_1, \dots, p_m such that p_i and p_j are not associates for all $i \neq j$, integers $e_i \geq 0, f_j \geq 0$ and units u and v . Show that:

(1) We have $a \mid b$ if and only if $e_i \leq f_i$ for all i .

(2) The gcd of a and b exists and is given by

$$\gcd(a, b) = p_1^{h_1} \cdots p_m^{h_m}$$

with

$$h_i = \min\{e_i, f_i\}$$

for all i (or any associate of this).

(3) The lcm of a and b exists and is given by

$$\operatorname{lcm}(a, b) = p_1^{g_1} \cdots p_m^{g_m}$$

with

$$g_i = \max\{e_i, f_i\}$$

for all i (or any associate of this).

Theorem 9.37. *If R is a UFD, then an element of R is irreducible if and only if it is prime.*

Proof. By Theorem 9.27, every prime element in R is irreducible. Suppose $r \in R$ is irreducible and that $r \mid ab$ for some $a, b \in R$. We must show that $r \mid a$ or $r \mid b$. Let

$$a = up_1 \cdots p_s \quad \text{and} \quad b = vq_1 \cdots q_t$$

with u and v units, each p_i and q_j irreducible, and $s, t \geq 0$. Since r is irreducible,

$$r = uva_1 \cdots a_s b_1 \cdots b_t$$

gives two irreducible factorization of the same element. So we must have either

$$s = 0 \quad \text{and} \quad r \cdot (uv)^{-1} = b$$

or

$$t = 0 \quad \text{and} \quad r \cdot (uv)^{-1} = a.$$

Thus $r \mid b$ or $r \mid a$. This proves that r is prime. \square

Our next goal is to show that every PID is a UFD. First, we show the following partial converse to Theorem 9.37.

Theorem 9.38 (Uniqueness of factorizations under certain conditions). *Assume R is a domain such that every irreducible element is a prime element. Given a nonzero $r \in R$, if*

$$r = up_1 \cdots p_n = vq_1 \cdots q_m$$

are two different irreducible factorization of r , then $n = m$ and there is a permutation σ such that, for all i , the elements p_i and $q_{\sigma(i)}$ are associates.

Proof. Without loss of generality, assume $n \leq m$. We will use induction on m .

If $m = 0$, since we assume $n \leq m$, we must have $n = 0$ too, and we are done. So assume $m > 0$ and that all irreducible factorizations with at most $m - 1$ irreducible factors are unique up to reordering and taking associates.

Since we are assuming that all irreducible elements are prime elements, in particular q_m is prime. Since q_m divides $r = vp_1 \cdots p_n$, we must have that q_m divides p_j for some j . Note that q_m cannot divide a unit or else it would be a unit. In particular, $n \geq 1$. After reordering, we may assume $j = n$. Thus $p_n = q_m w$ for some $w \in R$. Since p_n is irreducible and q_m is not a unit, w must be a unit and hence p_n and q_m are associates. We get

$$vq_1 \cdots q_m = (uw)p_1 \cdots p_{n-1}q_m$$

with $uw \in R^\times$. Since R is a domain, we may divide by q_m to obtain

$$vq_1 \cdots q_{m-1} = (uw)p_1 \cdots p_{n-1}$$

By the induction hypothesis, $n - 1 = m - 1$, and hence $n = m$, and p_1, \dots, p_{n-1} are associates of q_1, \dots, q_{m-1} in some order. Since p_n and q_m are associates, this completes our proof. \square

Theorem 9.39. *Every PID is a UFD.*

Proof. Let R be a PID. By Theorem 9.27, every irreducible element is a prime element. By Theorem 9.38, irreducible factorizations are unique when they exist. It remains to show that every nonzero element $r \in R$ has at least one irreducible factorization. Suppose this is not the case. Then r must not be a unit and it must not be irreducible, and so r must factor nontrivially as $r = x_1y_1$ with neither x_1 nor y_1 a unit. Likewise, both x_1 and y_1 cannot be irreducible. Without loss of generality, say it is y_1 , so that y_1 admits a nontrivial factorization $y_1 = x_2y_2$. At least one of these is not irreducible, say it is y_2 so that $y_2 = x_3y_3$ and $r = x_1x_2x_3y_3$. Continuing in this way, we construct an infinite sequence of elements

y_1, y_2, \dots . Since $y_i = y_{i+1}x_{i+1}$ we have $(y_i) \subseteq (y_{i+1})$, and since x_{i+1} is not a unit $(y_i) \subsetneq (y_{i+1})$ for all i . That is, we have constructed an infinite, strictly ascending chain of ideals

$$(y_1) \subsetneq (y_2) \subsetneq (y_3) \subsetneq \dots$$

I claim this is not possible. To show that, let

$$I = \bigcup_i (y_i).$$

While the union of ideals is not usually an ideal, the union of any *nested* chain of ideals is in fact an ideal, by Exercise 70. Since R is a PID, we must have $I = (z)$ for some z . But then $z \in (y_i)$ for some i , and it follows that

$$(y_i) = (y_{i+1}) = \dots$$

This is a contradiction, and thus we conclude that R is in fact a UFD. \square

Remark 9.40. The proof of Theorem 9.39 works just as well if R is a *noetherian* domain. In a noetherian ring, every ideal is finitely generated. In fact, as long as the ideal I constructed in the proof is finitely generated, say by z_1, \dots, z_m , there is an i such that $z_1, \dots, z_m \in (y_i)$ and hence $I \subseteq (y_i)$, which leads to a contradiction.

Thus, every noetherian integral domain having the property that all irreducible elements are prime elements must be a UFD.

Remark 9.41. There exist UFDs that are not noetherian. For instant, any polynomial ring

$$R = F[x_1, x_2, \dots]$$

in a countably infinite list of variables with coefficients in a field F is a UFD but it is not noetherian, because the ideal

$$(x_1, x_2, \dots)$$

generated by all the variables is not finitely generated.

Chapter 10

Polynomials and irreducibility

We will also be interested in understanding when a polynomial is irreducible. In this chapter, we will discuss a few useful irreducibility criteria that we will use often later on.

10.1 Gauss' Lemma

Definition 10.1 (Field of fractions). Let R be a domain. The **field of fractions** of R is the field

$$\text{Frac}(R) := \left\{ \frac{r}{u} \mid r, u \in R, u \neq 0 \right\} / \sim$$

where \sim is the equivalence relation given by

$$\frac{r}{u} \sim \frac{r'}{u'} \text{ if } ru' = r'u.$$

The operations are given by

$$\frac{r}{v} + \frac{s}{w} = \frac{rw + sv}{vw} \quad \text{and} \quad \frac{r}{v} \frac{s}{w} = \frac{rs}{vw}.$$

The zero in $\text{Frac}(R)$ is the element $\frac{0}{1}$ and the identity is the element $\frac{1}{1}$. There is an injective ring homomorphism

$$\begin{aligned} R &\rightarrow \text{Frac}(R) \\ r &\mapsto \frac{r}{1} \end{aligned}$$

We write elements in $\text{Frac}(R)$ in the form $\frac{r}{u}$ even though they are equivalence classes of such expressions.

Exercise 84. Check that $\text{Frac}(R)$ is indeed a field and the map given above is a ring homomorphism.

Example 10.2. (1) For $R = \mathbb{Z}$, the construction of the field of fractions of \mathbb{Z} recovers \mathbb{Q} .

(2) The field of fractions of $R = \mathbb{R}[x]$ is the field of rational functions $\mathbb{R}(x)$.

(3) We may identify the field of fractions of $R = \mathbb{Z}[i]$ with $\mathbb{Q}(i)$.

Exercise 85. Establish the following universal mapping property for the field of fractions construction:

Let R be an integral domain and F its field of fractions. Given an injective ring homomorphism $f: R \rightarrow E$ where E is a field, there is a unique ring homomorphism $\tilde{f}: F \rightarrow E$ such that $\tilde{f} \circ \iota = f$. Moreover, \tilde{f} is also injective. In fact,

$$\tilde{f}\left(\frac{a}{b}\right) = \frac{f(a)}{f(b)}.$$

Lemma 10.3. Suppose R is an integral domain, $f, g \in R[x]$, and that p is a prime element of R . If p divides all of the coefficients of fg , then p divides all of the coefficients of f or all the coefficients of g .

Proof. Let $R[x] \rightarrow (R/(p))[x]$ be the map $h(x) \mapsto \bar{h}(x)$ that mods out the coefficients by p . Since this is a ring homomorphism, we have

$$\overline{fg}(x) = \bar{f}(x)\bar{g}(x).$$

Since we assume p divides every coefficient of f , we have

$$\bar{f}(x)\bar{g}(x) = \overline{\bar{f} \cdot \bar{g}}(x) = 0$$

in $(R/(p))[x]$. Since p is prime, $R/(p)$ is an integral domain and thus, as we proved before, $R/(p)[x]$ is also an integral domain. We must therefore have $\bar{f}(x) = 0$ or $\bar{g}(x) = 0$; that is, either p divides every coefficient of f or it divides every coefficient of g . \square

Theorem 10.4 (Gauss' Lemma). *Let R be a UFD with field of fractions F . Regard R as a subring of F (via the canonical map) and view elements in $R[x]$ as also being elements of $F[x]$ via the induced map $R[x] \hookrightarrow F[x]$. If f is irreducible in $R[x]$, then f remains irreducible when regarded as an element of $F[x]$.*

Remark 10.5. This result is at least a tiny bit surprising. Note that there are many irreducible polynomials in $\mathbb{R}[x]$ that do *not* remain irreducible in the larger ring $\mathbb{C}[x]$, such as $x^2 + 1$. So, in general, one might think that passing to a larger ring of coefficients would cause some irreducible polynomial to become reducible. Gauss' Lemma says that this is *not* the case if the larger ring is the field of fractions of the smaller one (provided the smaller one is a UFD).

Proof. We will prove the contrapositive, so we will show that if $f \in R[x]$ is reducible in $F[x]$, then it is also reducible in $R[x]$. Suppose f factors nontrivially as $f = AB$ in $F[x]$. Since F is a field, the units of $F[x]$ are the nonzero constant polynomials, and so having a nontrivial factorization means $\deg(A), \deg(B) > 0$. All the coefficients of A and B are fractions, and so we may clear denominators — that is, we can find nonzero elements $r, s \in R$ (e.g., by taking the product of all the denominators) such that $a := rA$ and $b := sB$ both belong to $R[x]$. Set $d = rs$ and observe that we have

$$df = ab$$

with $d \in R$ and $f, a, b \in R[x]$.

If d is a unit in R , then we are done since then

$$f = (d^{-1}a)b$$

is a nontrivial factorization in $R[x]$, given that $R[x]^\times = R^\times$ and that $\deg(a), \deg(b) > 0$.

Since R is a UFD, we have $d = p_1 \cdots p_m$, for some $m \geq 1$, with each p_i irreducible and hence prime. Since p_m divides every coefficient of df , by Lemma 10.3 p_m must also either divide every coefficient of a or divide every coefficient of b . So, upon dividing through by p_1 we obtain

$$d_1 f = a_1 b_1$$

with $a_1, b_1 \in R[x]$ and $d_1 = p_1 \cdots p_{m-1} \in R$. More precisely, if p divides a then $a_1 = a/p$ and $b_1 = b$ and if p divides b then $a_1 = a$ and $b_1 = b/p$.

By the same reasoning, we may divide by p_{m-1} to obtain

$$d_2 f = a_2 b_2$$

with $a_2, b_2 \in R[x]$ and $d_2 = p_1 \cdots p_{m-3} \in R$. Continuing in this way, we arrive at an equation of the form

$$f = a_m b_m$$

in $R[x]$ with $\deg(a_m) = \deg(A) > 0$ and $\deg(b_m) = \deg(B) > 0$. This proves f is reducible in $R[x]$. \square

Theorem 10.6. *Let R be a UFD with field of fractions F . Regard R as a subring of F (via the canonical map) and view elements in $R[x]$ as also being elements of $F[x]$ via the induced map $R[x] \hookrightarrow F[x]$. Let $f \in R[x]$. If f is irreducible when regarded as an element in $F[x]$ and the gcd of the coefficients of f is a unit in R , then f is irreducible as an element of $R[x]$.*

Remark 10.7. This is false if the gcd of the coefficients of f is not a unit. To see this, note that $2x + 6$ is irreducible in $\mathbb{Q}[x]$ but not in $\mathbb{Z}[x]$, since it factors as $2(x + 3)$. In $\mathbb{Q}[x]$, however, this factorization is trivial because 2 is a unit.

Proof. We again prove the contrapositive: we will show that if f is reducible in $R[x]$ then either the gcd of the coefficients of f is not a unit or f remains reducible in $F[x]$.

Suppose f factors nontrivially in $R[x]$ as $f = gh$ with g and h nonunits. If both g and h have positive degree, then they remain nonunits in $F[x]$, and so f is reducible in that ring too. Otherwise, suppose g is the constant polynomial c . Then, since c is a nonunit in R and $f = ch$, the gcd of the coefficients of f is not a unit. \square

Example 10.8. Let us use Gauss's Lemma to show that the polynomial

$$f = x^4 + 7x^3 + 18x^2 + 31$$

is irreducible in $\mathbb{Q}[x]$. First, one can check that f has no roots in \mathbb{Q} by the Rational Root Test, but that does not mean it does not factor as a product of two irreducible quadratics.

By Gauss's Lemma, if f is irreducible in $\mathbb{Z}[x]$ then it is irreducible in $\mathbb{Q}[x]$. Working in $\mathbb{Z}[x]$ has the advantage that we can mod out by a prime:

Suppose f did factor nontrivially in $\mathbb{Z}[x]$. Then, since f is monic, it would factor as $f = gh$ with g and h monic polynomials in $\mathbb{Z}[x]$ each of degree at least one. For any prime integer p , we would have

$$\bar{f} = \bar{g}\bar{h}$$

in $(\mathbb{Z}/p)[x]$ with $\deg(\bar{g}) = \deg(g)$ and $\deg(\bar{h}) = \deg(h)$, since g and h are monic.

Let $p = 2$. We have

$$\bar{f} = x^4 + x^3 + 1 \in (\mathbb{Z}/2)[x].$$

This polynomial does not have a root, as the only possibilities are 0 and 1, and hence it has no linear factors. Therefore, \bar{g} and \bar{h} must be irreducible of degree 2. But the only irreducible polynomial of degree 2 in $(\mathbb{Z}/2)[x]$ is $q = x^2 + x + 1$, since we can check one by one and see that all the other three quadratic polynomials have roots. Since

$$q^2 = x^4 + x^2 + 1 \neq \bar{f},$$

we have reached a contradiction. We conclude that f is irreducible in $\mathbb{Q}[x]$.

10.2 Eisenstein's Criterion

Theorem 10.9 (Eisenstein's Criterion). *Let R be a domain and consider a monic polynomial $f = x^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0 \in R[x]$ of degree $n \geq 1$. If there exists a prime ideal P of R such that $a_0, \dots, a_{n-1} \in P$ and $a_0 \notin P^2$, then f is irreducible in $R[x]$.*

Proof. Suppose, by contradiction, that f is reducible. Since it is monic, we would be able to factor it as $f = gh$, where g and h are polynomials in $R[x] \setminus R[x]^\times$. Since the leading coefficients of g and h multiply to 1, those coefficients must be units in R . We may thus assume that g and h are monic by multiplying each of these by the inverse of their leading coefficient.

Consider the canonical quotient $R \rightarrow R/P$ and the induced reduction homomorphism $R[x] \rightarrow (R/P)[x]$, and write \bar{p} for the image of p in $(R/P)[x]$. Then in the ring $(R/P)[x]$ we have the identity

$$x^n = \bar{f}(x) = \bar{g}(x)\bar{h}(x).$$

Set $T = R/P$ and notice that T is a domain, by Theorem 8.85. We now need an auxiliary claim.

Claim: If T is a domain and $\bar{g}, \bar{h} \in T[x]$ are monic polynomials such that $\bar{g}\bar{h} = x^n$, then $\bar{g} = x^m$ and $\bar{h} = x^{n-m}$ for some $1 \leq m \leq n$.

Proof of claim: Let

$$g = x^m + a_{m-1}x^{m-1} + \cdots + a_0 \quad \text{and} \quad h(x) = x^{n-m} + b_{n-m-1}x^{n-m-1} + \cdots + b_0.$$

Let j be the least integer such that $a_j \neq 0$ and i the least integer such that $b_i \neq 0$. Set $a_m = 1 = b_{n-m}$. The coefficient of x^{i+j} in $g(x)h(x)$ is $\sum_{s+t=i+j} a_s b_t$. The only nonzero term here is the term $a_j b_i$, which is indeed nonzero since R is a domain, and hence the degree $i+j$ term of gh is non-zero. This forces $i = m$ and $j = n$.

The Claim thus gives that \bar{g} and \bar{h} have zero constant terms or, in other words, the constant terms of g and h are both in P . The constant term of $f = g \cdot h$ is thus in P^2 , which is a contradiction. \square

If R is UFD, such as \mathbb{Z} , we may consider the special case where P is a principal ideal.

Corollary 10.10. *Let R be a UFD and consider*

$$f = x^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0 \in R[x]$$

with $n \geq 1$. If there is a prime element p such that $p \mid a_i$ for $i = 0, \dots, n-1$, and $p^2 \nmid a_0$, then f is irreducible.

Example 10.11. For example, $x^n - p \in \mathbb{Z}[x]$ is irreducible for all $n \geq 1$ and all primes p . By Gauss's Lemma, it is irreducible in $\mathbb{Q}[x]$ too. This implies that

$$\mathbb{Q}[x]/(x^n - p)$$

is a field. In fact, this field is isomorphic to $\mathbb{Q}(\sqrt[n]{p})$, the smallest subfield of \mathbb{C} that contains \mathbb{Q} and $\sqrt[n]{p}$.

Example 10.12. Let F be any field. We claim that the polynomial

$$f(x, y) = x^3 + y^5x + y$$

is irreducible in $F[x, y]$. To prove this, let us think of $F[x, y]$ as $F[x, y] = R[x]$ where $R = F[y]$, so that

$$f = x^3 + r_1x + r_0$$

with $r_1 = y^5$ and $r_0 = y$. Note that y is a prime element of R that divides r_1 and r_0 , but y^2 does not divide r_0 . So, by Eisenstein's Criterion, f is irreducible.

10.3 Applications of Gauss' Lemma

In this section, we will give a couple more applications of Gauss' Lemma. First we make a convenient definition.

Definition 10.13. Suppose R is a UFD and $f(x) \in R[x]$. A *content* of $f(x)$, denoted $\text{cont}(f) \in R$, is a gcd of the coefficients of $f(x)$. We say $f(x)$ is *primitive* if $\text{cont}(f(x))$ is a unit.

Note that a content of f is a unit if and only if every other content of f is also a unit. By Gauss' Lemma, in the setting of the definition a primitive polynomial in $R[x]$ that is irreducible over $\text{Frac}(R)[x]$ is irreducible over $R[x]$.

Exercise 86. Assume R is a UFD and $f(x), g(x) \in R[x]$. Then $f(x)$ and $g(x)$ are primitive if and only if $f(x)g(x)$ is primitive.

Lemma 10.14. *Suppose R is a UFD. Any primitive polynomial $f(x)$ of positive degree in $R[x]$ factors as a product of primitive irreducible polynomials.*

Proof. If $f(x)$ is irreducible, there is nothing to prove. Otherwise there is a nontrivial factorization $f(x) = a(x)b(x)$, and since $f(x)$ is primitive, we must have $\deg(a(x)) < \deg(f(x))$ and $\deg(b(x)) < \deg(f(x))$. Moreover, by the exercise above, each of $a(x)$ and $b(x)$ must also be primitive. The existence of irreducible factorizations for primitive polynomials thus follows by induction on degree. \square

Theorem 10.15. *Let R be a UFD. Then the polynomial ring $R[x]$ is a UFD.*

Proof. We first show factorizations exist. Let $f(x) \in R[x]$. Since R is UFD, we may consider its content $c = \text{cont}(f)$, so that $f(x) = cf'(x)$ with $f'(x)$ primitive. Now c factors in R into irreducibles and this remains an irreducible factorization of c in $R[x]$. So it suffices to prove $f'(x)$ has an irreducible factorization too, which follows from the previous lemma.

For uniqueness, suppose we have two products of irreducible elements of $R[x]$ that are equal. Among the factors involved, we first write the constant factors and the nonconstant ones, getting an equation of the form

$$d_1 \cdots d_m p_1(x) \cdots p_n(x) = e_1 \cdots e_s q_1(x) \cdots q_t(x),$$

where d_i, e_j are irreducible elements of R and p_i, q_j are irreducible polynomials of degree at least one. Note that each of p_i, q_j must be primitive (since a nonprimitive polynomial $p(x)$ factors as $\text{cont}(p)p'(x)$). By the exercise, $p_1(x) \cdots p_n(x)$ and $q_1(x) \cdots q_t(x)$ are also primitive. It follows that $d_1 \cdots d_m$ is the content of $d_1 \cdots d_m p_1(x) \cdots p_n(x)$ and $e_1 \cdots e_s$ is the content of $e_1 \cdots e_s q_1(x) \cdots q_t(x)$. Since these are equal, $d_1 \cdots d_m$ and $e_1 \cdots e_s$ agree up to a unit factor and hence, since R is a UFD, we have $s = m$ and, after reordering, d_i and e_i are associates, for all i .

We may now cancel d_1, \dots, d_m to get that

$$p_1(x) \cdots p_n(x) = u q_1(x) \cdots q_t(x)$$

for some unit u of R , and it remains to prove $n = t$ and, after reordering, that p_i and q_i are associates, for all i . Let F be the field of fractions of R . We know $F[x]$ is a Euclidean domain and hence it is a PID and hence a UFD. Moreover, by Gauss' Lemma, each p_i, q_j remains irreducible in $F[x]$. Thus $n = t$ and, after reordering, $p_i(x)$ and $q_i(x)$ are associate in $F[x]$, for all i . This means that for each i we have $p_i(x) = \frac{r_i}{s_i} q_i(x)$, for some non-zero elements r_i, s_i of R , and hence $s_i p_i(x) = r_i q_i(x)$. But since p_i, q_j are primitive, we have $s_i = \text{cont}(s_i p_i(x))$ and $r_i = \text{cont}(r_i q_i(x))$. It follows that s_i and r_i are associates in R and hence $p_i = u q_i$ for some unit u . \square

Corollary 10.16. *Let F be a field. Then $F[x_1, \dots, x_n]$ and $\mathbb{Z}[x_1, \dots, x_n]$ are UFDs.*

The next application may be familiar from elementary algebra.

Theorem 10.17 (Rational root test). *Let R be a UFD, and $F = \text{Frac}(R)$. Let $f(x) = a_n x^n + \cdots + a_1 x + a_0 \in R[x]$, and $a_n, a_0 \neq 0$ in R . Suppose that $p/q \in F$ is a root of f with p/q in lowest terms, meaning p and q have no common nonunit divisor in R . Then $p | a_0$ and $q | a_n$.*

Proof. If p/q is a root of f , then $x - p/q$ divides f in $F[x]$ and likewise $qx - p$ divides f in $F[x]$, so we can write $f = (qx - p)g$ for some $g \in F[x]$; we can then write $df = (qx - p)g'$ for some $g' \in R[x]$ and $d \in R$. By Gauss' Lemma, since $qx - p$ is irreducible over $F[x]$ and is primitive, it is irreducible over $R[x]$. Since $R[x]$ is a UFD, $qx - p$ is a prime element, and for degree reasons, $qx - p \nmid d$, so $qx - p \mid f$. By distributing coefficients, one finds that the leading coefficient of f must be a multiple of q , and the constant coefficient must be a multiple of p . \square

Example 10.18. The polynomial $x^5 - x + 6 \in \mathbb{Q}[x]$ has no rational roots. Indeed, by the rational root test, any rational root would be of the form p/q with $p \mid 6$ and $q \mid 1$. Thus, it suffices to check $\pm 1, \pm 2, \pm 3, \pm 6$, and none of these is a root.

Part III

Modules

Chapter 11

Modules

Modules are a generalization of the concept of a vector space to any ring of scalars. But while vector spaces make for a great first example of modules, many of the basic facts we are used to from linear algebra are often a little more subtle over a general ring. These differences are features, not bugs. We will introduce modules, study some general linear algebra, and discuss the differences that make the general theory of modules richer and even more fun.

11.1 Modules: definition and examples

Definition 11.1. Let R be a ring with $1 \neq 0$. A **left R -module** is an abelian group $(M, +)$ together with an action $R \times M \rightarrow M$ of R on M , written as $(r, m) \mapsto rm$, such that for all $r, s \in R$ and $m, n \in M$ we have the following:

- $(r + s)m = rm + sm$,
- $(rs)m = r(sm)$,
- $r(m + n) = rm + rn$, and
- $1m = m$.

A **right R -module** is an abelian group $(M, +)$ together with an action of R on M , written as $M \times R \rightarrow M$, $(m, r) \mapsto mr$, such that for all $r, s \in R$ and $m, n \in M$ we have

- $m(r + s) = mr + ms$,
- $m(rs) = (mr)s$,
- $(m + n)r = mr + nr$, and
- $m1 = m$.

By default, we will be studying left R -modules. To make the writing less heavy, we will sometimes say **R -module** rather than left R -module whenever there is no ambiguity.

Remark 11.2. If R is a commutative ring, then any left R -module M may be regarded as a right R -module by setting $mr := rm$. Likewise, any right R -module may be regarded as a left R -module. Thus for commutative rings, we just refer to modules, and not left or right modules.

Lemma 11.3 (Arithmetic in modules). *Let R be a ring with $1_R \neq 0_R$ and M be an R -module. Then $0_R m = 0_M$ and $(-1_R)m = -m$ for all $m \in M$.*

Proof. Let $m \in M$. Then

$$0_R m = (0_R + 0_R)m = 0_R m + 0_R m.$$

Since M is an abelian group, the element $0_R m$ has an additive inverse, $-0_R m$, so adding it on both sides we see that

$$0_M = 0_R m.$$

Moreover,

$$m + (-1_R)m = 1_R m + (-1_R)m = (1_R - 1_R)m = 0_R m = 0_M,$$

so $(-1_R)m = -m$. \square

Typically, one first encounters modules in an undergraduate linear algebra course: the vector spaces from linear algebra are modules over fields. Later we will see that vector spaces are much simpler modules than modules over other rings. So while one might take linear algebra and vector spaces as an inspiration for what to expect from a module, be warned that this perspective can often be deceiving.

Definition 11.4. Let F be a field. A **vector space** over F is an F -module.

We will see more about vector spaces soon. Note that many of the concepts we will introduce have special names in the case of vector spaces. Here are some other important examples:

Lemma 11.5. *Let M be a set with a binary operation $+$. Then*

- (1) *M is an abelian group if and only if M is a \mathbb{Z} -module.*
- (2) *M is an abelian group such that $nm := \underbrace{m + \cdots + m}_{n \text{ times}} = 0_M$ for all $m \in M$ if and only if M has a \mathbb{Z}/n -module structure.*

Proof. First, we show 1). If M is a \mathbb{Z} -module, then $(M, +)$ is an abelian group by definition of module. Conversely, if $(M, +)$ is an abelian group then there is a unique \mathbb{Z} -module structure on M given by the formulas below. The uniqueness of the \mathbb{Z} action follows from the identities below in which the right hand side is determined only by the abelian group structure of M . The various identities follow from the axioms of a module:

$$\begin{cases} i \cdot m = (\underbrace{1 + \cdots + 1}_i) \cdot m = \underbrace{1 \cdot m + \cdots + 1 \cdot m}_i = \underbrace{m + \cdots + m}_i & \text{if } i > 0 \\ 0 \cdot m = 0_M \\ i \cdot m = -(-i) \cdot m = -(\underbrace{m + \cdots + m}_{-i}) & \text{if } i < 0. \end{cases}$$

We leave it as an exercise to check that this \mathbb{Z} -action really satisfies the module axioms.

Now we show 2). If M is a \mathbb{Z}/n module, then $(M, +)$ is an abelian group by definition, and $nm = \underbrace{m + \cdots + m}_n = \underbrace{[1]_n \cdot m + \cdots + [1]_n \cdot m}_n = [0]_n m = 0_M$.

Conversely, there is a unique \mathbb{Z}/n -module structure on M given by the formulas below, which are analogous to the ones above:

$$\begin{cases} [i]_n \cdot m = (\underbrace{[1]_n + \cdots + [1]_n}_i) \cdot m = \underbrace{[1]_n \cdot m + \cdots + [1]_n \cdot m}_i = \underbrace{m + \cdots + m}_i & \text{if } i > 0 \\ 0 \cdot m = 0_M \\ [i]_n \cdot m = -(-[i]_n) \cdot m = -(m + \cdots + m) & \text{if } i < 0. \end{cases}$$

These formulas are well-defined, meaning they are independent of the choice of representative for $[i]_n$, because of the assumption that $nm = 0_M$. Again checking that this \mathbb{Z}/n -action really satisfies the module axioms is left as an exercise. \square

The proposition above says in particular that any group of the form

$$G = \mathbb{Z}^\ell \times \mathbb{Z}/d_1 \times \cdots \times \mathbb{Z}/d_m$$

is a \mathbb{Z} -module, and if $\ell = 0, m \geq 1$ and $d_i \mid n$ for $1 \leq i \leq m$ then G is also a \mathbb{Z}/n -module. In particular, the Klein group is a $\mathbb{Z}/2$ -module.

In contrast to vector spaces, for M a module over a ring R , it can happen that $rm = 0$ for some $r \in R$ and $m \in M$ such that $r \neq 0_R$ and $m \neq 0_M$. For example, in the Klein group K_4 viewed as a \mathbb{Z} -module we have $2m = 0$ for all $m \in K_4$.

- Example 11.6.** (1) The trivial R -module is $0 = \{0\}$ with $r0 = 0$ for any $r \in R$.
- (2) If R is any ring, then R is a left and right R -module via the action of R on itself given by its internal multiplication.
- (3) If I is a left (respectively, right) ideal of a ring R then I is a left (respectively, right) R -module with respect to the action of R on I by internal multiplication.
- (4) If R is a subring of a ring S , then S is an R -module with respect to the action of R on S by internal multiplication in S .
- (5) If R is a commutative ring with $1 \neq 0$, then $R[x_1, \dots, x_n]$ is an R -module for any $n \geq 1$. This is a special case of (4).
- (6) If R is a commutative ring, let $\text{Mat}_n(R)$ denote the set of $n \times n$ matrices with entries in R . Then $\text{Mat}_n(R)$ is an R -module for $n \geq 1$, with the R -action given by multiplying all the entries of the given matrix by the given element of R .

(7) The **free module** over R of rank n is the set

$$R^n = \left\{ \begin{bmatrix} r_1 \\ \vdots \\ r_n \end{bmatrix} \mid r_i \in R, 1 \leq i \leq n \right\}$$

with componentwise addition and multiplication by elements of R , as follows:

$$\begin{bmatrix} r_1 \\ \vdots \\ r_n \end{bmatrix} + \begin{bmatrix} r'_1 \\ \vdots \\ r'_n \end{bmatrix} = \begin{bmatrix} r_1 + r'_1 \\ \vdots \\ r_n + r'_n \end{bmatrix} \text{ and } r \begin{bmatrix} r_1 \\ \vdots \\ r_n \end{bmatrix} = \begin{bmatrix} rr_1 \\ \vdots \\ rr_n \end{bmatrix}.$$

We will often write the elements of R^n as n -tuples (r_1, \dots, r_n) instead. Notice that R is the free R -module of rank 1.

(8) More generally, given a collection of R -modules $\{A_i\}$, the abelian group

$$\bigoplus_i A_i = \{(a_i)_i \mid a_i \in A_i, a_i = 0 \text{ for all } i \text{ but finitely many}\}$$

is an R -module with the R -action $r(a_i) := (ra_i)$.

11.2 Submodules and restriction of scalars

Definition 11.7. Let R be a ring and let M be a left R -module. An **R -submodule** of M is a subgroup N (under addition) of M satisfying $rn \in N$ for all $r \in R$ and $n \in N$.

The submodules of an R -module M are precisely the subsets of M which are modules in their own right, via the same R -action as we are considering for M .

Exercise 87. Show that if N is a submodule of M , then N is an R -module via the restriction of the action of R on M to the subset N .

Example 11.8. Every R -module M has two **trivial submodules**: M itself and the **zero module** $0 = \{0_M\}$. A submodule N of M is **nontrivial** if $N \neq M$ and $N \neq 0$.

Lemma 11.9 (Submodule tests). *Let R be a ring with $1 \neq 0$ and let M be a left R -module. Let N be a nonempty of M .*

- (1) (*Two-step test*) N is an R -submodule of M if and only if $n + n' \in N$ and $rn \in N$ for all $n, n' \in N$ and $r \in R$.
- (2) (*One-step test*) N is an R -submodule of M if and only if $rn + n' \in N$ for all $n, n' \in N$ and $r \in R$.

Proof. Exercise. □

Example 11.10. Let R be a ring and let M be a subset of R . Then M is a left (respectively, right) R -submodule of R if and only if M is a left (respectively, right) ideal of R .

Exercise 88. Let R be a ring and let A and B be submodules of an R -module M . Then the **sum** of A and B ,

$$A + B := \{a + b \mid a \in A, b \in B\},$$

and $A \cap B$ are both R -submodules of M .

Exercise 89. Let R be a commutative ring with $1 \neq 0$, let I be an ideal of R and let M be an R -module. Show that

$$IM := \left\{ \sum_{k=1}^n j_k m_k \mid n \geq 0, j_k \in I, m_k \in M \text{ for } 1 \leq k \leq n \right\}$$

is a submodule of M .

Example 11.11. When R is a field, the submodules of a vector space V are precisely the subspaces of V . When $R = \mathbb{Z}$, then the class of R -modules is simply the class of all abelian groups, by Lemma 11.5. The submodules of a \mathbb{Z} -module M coincide with the subgroups of the abelian group M .

Definition 11.12. Let R be a ring with $1 \neq 0$ and let M be an R -module. Given elements $m_1, \dots, m_n \in M$, the **submodule generated** by m_1, \dots, m_n is the subset of M given by

$$Rm_1 + \cdots + Rm_n := \{r_1 m_1 + \cdots + r_n m_n \mid r_1, \dots, r_n \in R\}.$$

Exercise 90. Let R be a ring with $1 \neq 0$ and M be an R -module. Given $m_1, \dots, m_n \in M$, the submodule generated by m_1, \dots, m_n is indeed a submodule of M . Moreover, this is the smallest submodule of M that contains m_1, \dots, m_n , meaning that every submodule of M containing m_1, \dots, m_n must also contain $Rm_1 + \cdots + Rm_n$.

Definition 11.13. Let R be a ring with $1 \neq 0$. An R -module M is **cyclic** if there exists an element $m \in M$ such that

$$M = Rm := \{rm \mid r \in R\}.$$

Given an R -module M , the ring R is sometimes referred to as the **ring of scalars**, by analogy to the vector space case. Given an action of a ring of scalars on a module, we can sometimes produce an action of a different ring of scalars on the same set, producing a new module structure.

Lemma 11.14 (Restriction of scalars). *Let $\phi: R \rightarrow S$ be a ring homomorphism. Any left S -module M may be regarded via **restriction of scalars** as a left R -module with R -action defined by $rm := \phi(r)m$ for any $m \in M$. In particular, if R is a subring of a ring S , then any left S -module M may be regarded via restriction of scalars as a left R -module with R -action defined by the action of the elements of R viewed as elements of S .*

Proof. Let $r, s \in R$ and $m, n \in M$. One checks that the axioms in the definition of a module hold for the given action using properties of ring homomorphisms. For example:

$$(r + s)m = \phi(r + s)m = (\phi(r) + \phi(s))m = \phi(r)m + \phi(s)m = rm + sm.$$

The remaining properties are left as an exercise. □

Note that the second module structure on M obtained via restriction of scalars is induced by the original module structure, so the two are related. In general, one can give different module structures on the same abelian group over different, possibly unrelated, rings.

Example 11.15. If I is an ideal of a ring R , applying restriction of scalars along the quotient homomorphism $q: R \rightarrow R/I$ tells us that any left R/I -module is also a left R -module. In particular, applying this to the R/I -module R/I makes R/I a left and right R -module by restriction of scalars along the quotient homomorphism. Thus the R -action on R/I is given by

$$r \cdot (a + I) := ra + I.$$

Example 11.16. Given any ring R there exists a unique ring homomorphism $\mathbb{Z} \rightarrow R$. Thus any R -module can be given the structure of a \mathbb{Z} -module by restriction of scalars along this unique map. Note also that a module over any ring is in particular an abelian group, so we can always regard any R -module as a \mathbb{Z} -module by forgetting the R -action and focusing only on the abelian group structure. These two constructions — the restriction of scalars to \mathbb{Z} and the *forgetful functor* — actually coincide.

The next example explains why restriction of scalars is called a *restriction*.

Example 11.17. Let R be a subring of S , and let $i: R \rightarrow S$ be the inclusion map, which must by definition be a ring homomorphism. Applying restriction of scalars to an S -module M via i is the same as simply *restricting* our scalars to the elements of R .

11.3 Module homomorphisms and isomorphisms

Definition 11.18. Given R -modules M and N , an **R -module homomorphism** from M to N is a function $f: M \rightarrow N$ such that for all $r \in R$ and $m, n \in M$ we have

- $f(m + n) = f(m) + f(n)$
- $f(rm) = rf(m).$

Remark 11.19. The condition $f(m + n) = f(m) + f(n)$ says that f is a homomorphism of abelian groups, and the condition $f(rm) = rf(m)$ says that f is R -linear, meaning that it preserves the R -action. Since f is a homomorphism of abelian groups, it follows that $f(0) = 0$ must hold.

Definition 11.20. Let M and N be vector spaces over a field F . A **linear transformation** from M to N is an F -module homomorphism $M \rightarrow N$.

Example 11.21. Let R be a commutative ring and M be an R -module. For each $r \in R$, the multiplication map $\mu_r: M \rightarrow M$ given by $\mu_r(m) = rm$ is a homomorphism of R -modules: indeed, by the definition of R -module we have

$$\mu_r(m + n) = r(m + n) = rm + rn = \mu_r(m) + \mu_r(n),$$

and

$$\mu_r(sm) = r(sm) = (rs)m = (sr)m = s(rm) = s\mu_r(m).$$

Note that R is not commutative, the left multiplication map $\mu_r : M \rightarrow M$ is not a homomorphism of (left) R -modules.

Definition 11.22. An R -module homomorphism $h : M \rightarrow N$ is an **R -module isomorphism** if there is an R -module homomorphism $g : N \rightarrow M$ such that $h \circ g = \text{id}_N$ and $g \circ h = \text{id}_M$. We say M and N are **isomorphic**, denoted $M \cong N$, if there exists an isomorphism $M \rightarrow N$.

To check that an R -module homomorphism $f : M \rightarrow N$ is an isomorphism, it is sufficient to check that it is bijective.

Exercise 91. Let $f : M \rightarrow N$ be a homomorphism of R -modules. Show that if f is bijective, then its set-theoretic inverse $f^{-1} : N \rightarrow M$ is an R -module homomorphism. Therefore, every bijective homomorphism of R -modules is an isomorphism.

One should think of a module isomorphism as a relabelling of the names of the elements of the module. If two modules are isomorphic, that means that they are *essentially the same*, up to renaming the elements.

Definition 11.23. Let $f : M \rightarrow N$ be a homomorphism of R -modules. The **kernel** of f is

$$\ker(f) := \{m \in M \mid f(m) = 0\}.$$

The **image** of f , denoted $\text{im}(f)$ or $f(M)$, is

$$\text{im}(f) := \{f(m) \mid m \in M\}.$$

Exercise 92. Let R be a ring with $1 \neq 0$, let M be an R -module, and let N be an R -submodule of M . Then the inclusion map $i : N \rightarrow M$ is an R -module homomorphism.

Exercise 93. If $f : M \rightarrow N$ is an R -module homomorphism, then $\ker(f)$ is an R -submodule of M and $\text{im}(f)$ is an R -submodule of N .

Definition 11.24. Let R be a ring and let M and N be R -modules. Then $\text{Hom}_R(M, N)$ denotes the set of all R -module homomorphisms from M to N , and $\text{End}_R(M)$ denotes the set $\text{Hom}_R(M, M)$. We call $\text{End}(M)$ the **endomorphism ring** of M , and elements of $\text{End}(M)$ are called **endomorphisms** of M .

The endomorphism ring of an R -module M is called that because it *is* a ring, with multiplication given by composition of endomorphisms, 0 given by the zero map (the constant equal to 0), and 1 given by the identity map. However, two homomorphisms from M to N are not composable unless $M = N$, so $\text{Hom}_R(M, N)$ is not a ring.

When R is commutative, $\text{Hom}_R(M, N)$ is, however, an R -module; let us describe its R -module structure. Given $f, g \in \text{Hom}_R(M, N)$, $f + g$ is the map defined by

$$(f + g)(m) := f(m) + g(m),$$

and given $r \in R$ and $f \in \text{Hom}_R(M, N)$, $r \cdot f$ is the R -module homomorphism defined by

$$(r \cdot f)(m) := r \cdot f(m) = f(rm).$$

The zero element of $\text{Hom}_R(M, N)$ is the **zero map**, the constant equal to 0_N .

Lemma 11.25. *Let M and N be R -modules over a commutative ring R . Then the addition and multiplication by scalars defined above make $\text{Hom}_R(M, N)$ an R -module.*

Proof. There are many things to check, including:

- The addition and the R -action are both well-defined: given $f, g \in \text{Hom}_R(M, N)$ and $r \in R$, we always have $f + g, rf \in \text{Hom}_R(M, N)$.
- The axioms of an R -module are satisfied for $\text{Hom}_R(M, N)$.

We leave the details as exercises. \square

We will see later that for an n -dimensional vector space V over a field F , there is an isomorphism of vector spaces $\text{End}_F(V) \cong M_n(F)$. This says that every linear transformation $T : V \rightarrow V$ corresponds to some $n \times n$ matrix. However, the story for general R -modules is a lot more complicated.

Lemma 11.26. *For any commutative ring R with $1 \neq 0$ and any R -module M there is an isomorphism of R -modules $\text{Hom}_R(R, M) \cong M$.*

Proof. Let $f : M \rightarrow \text{Hom}_R(R, M)$ be given for each $m \in M$ by $f(m) = \phi_m$ where ϕ_m is the map defined by $\phi_m(r) = rm$ for all $r \in R$. Now we have many things to check:

- f is well-defined, meaning that for any $m \in M$, its image $f(m) = \phi_m$ is an element of $\text{Hom}_R(R, M)$, since

$$\phi_m(r_1 + r_2) = (r_1 + r_2)m = r_1m + r_2m = \phi_m(r_1) + \phi_m(r_2)$$

$$\phi_m(r_1r_2) = (r_1r_2)m = r_1(r_2m) = r_1\phi_m(r_2)$$

for all $r_1, r_2 \in R$.

- f is an R -module homomorphism, since

$$\phi_{m_1+m_2}(r) = r(m_1 + m_2) = rm_1 + rm_2 = \phi_{m_1}(r) + \phi_{m_2}(r)$$

$$\phi_{r'm}(r) = r(r'm) = (rr')m = r'(rm) = r'\phi_m(r)$$

- f is injective, since $\phi_m = \phi_{m'}$ implies in particular that $\phi_m(1_R) = \phi_{m'}(1_R)$, which by definition of ϕ_- means that $m = m'$.
- f is surjective, since for $\psi \in \text{Hom}_R(R, M)$ we have $\psi(r) = \psi(r1_R) = r\psi(1_R)$ for all $r \in R$, so $\psi = \phi_{\psi(1_R)}$.

This shows that f is an R -module isomorphism. \square

Definition 11.27. Let R be a commutative ring with $1_R \neq 0_R$. An **R -algebra** is a ring A with $1_A \neq 0_A$ together with a ring homomorphism $f : R \rightarrow A$ such that $f(R)$ is contained in the center of A .

Given an R -algebra A , the R -algebra structure on A induces a natural R -module structure: given elements $r \in R$ and $a \in A$, the R -action is defined by

$$r \cdot a := f(r)a,$$

where the product on the right is the multiplication in A . Similarly, we get a natural right R -module structure on A , and since by definition $f(R)$ is contained in the center of A , we obtain what is called a *balanced bimodule* structure on A .

Example 11.28. Let R be a commutative ring with $1_R \neq 0_R$. The ring $R[x_1, \dots, x_n]$ together with the inclusion map $R \hookrightarrow R[x_1, \dots, x_n]$ is an R -algebra. More generally, any quotient of $R[x_1, \dots, x_n]$ is an R -algebra.

The ring of matrices $M_n(R)$ with the homomorphism $r \mapsto rI_n$ is also an R -algebra.

Lemma 11.29. Let R be a commutative ring with $1 \neq 0$ and let M be an R -module. Then $\text{End}_R(M)$ is an R -algebra, with addition and R -action defined as above, and multiplication defined by composition $(fg)(m) = f(g(m))$ for all $f, g \in \text{End}_R(M)$ and all $m \in M$.

Proof. There are many things to check here, including that:

- The axioms of a (unital) ring are satisfied for $\text{End}_R(M)$.
- There is a ring homomorphism $f: R \rightarrow \text{End}_R(M)$ such that $f(1_R) = 1_{\text{End}_R(M)} = \text{id}_M$ and $f(R) \subseteq Z(\text{End}_R(M))$.

We will just check the last item and leave the others as exercises. Define $f: R \rightarrow \text{End}_R(M)$ by $f(r) = r\text{id}_M$. Notice that this element $f(r)$ is the map μ_r from Example 11.21. Then

$$f(r+s) = (r+s)\text{id}_M = r\text{id}_M + s\text{id}_M = f(r) + f(s)$$

and

$$f(rs) = (rs)\text{id}_M = (r\text{id}_M) \circ (s\text{id}_M) = f(r)f(s)$$

show that f is a ring homomorphism. Moreover, $\text{id}_M \in Z(\text{End}_R(M))$, and one can check easily that $\mu_r \in Z(\text{End}_R(M))$: given any other $g \in \text{End}_R(M)$, and any $m \in M$, since g is R -linear we have

$$(g \circ \mu_r)(m) = g(\mu_r(m)) = g(rm) = rg(m) = (\mu_r \circ g)(m).$$

This shows that $f(R)$ is contained in the center of $\text{End}_R(M)$. □

Remark 11.30. Let R be a commutative ring with $1 \neq 0$ and let M be an R -module. Then M is also an $\text{End}_R(M)$ -module with the action $\phi m = \phi(m)$ for any $\phi \in \text{End}_R(M)$, $m \in M$.

Definition 11.31. Let R be a ring, let M be an R -module, and let N be a submodule of M . The *quotient module* M/N is the quotient group M/N with R action defined by

$$r(m+N) := rm + N$$

for all $r \in R$ and $m + N \in M/N$.

Lemma 11.32. Let R be a ring, let M be an R -module, and let N be a submodule of M . The quotient module M/N is an R -module, and the quotient map $q: M \rightarrow M/N$ is an R -module homomorphism with kernel $\ker(q) = N$.

Proof. Among the many things to check here, we will only check the well-definedness of the R -action on M , and leave the others as exercises. To check well-definedness, consider $m + N = m' + N$. Then $m - m' \in N$, so $r(m - m') \in N$ by the definition of submodule. This gives that $rm - rm' \in N$, hence $rm + N = rm' + N$. \square

Definition 11.33. Given an R -module M and a submodule N of M , the map $q: M \rightarrow M/N$ is the **canonical quotient map**, or simply the canonical map from M to N .

Example 11.34. If R is a field, quotient modules are the same thing as quotient vector spaces. When $R = \mathbb{Z}$, recall that \mathbb{Z} -modules are the same as abelian groups, by Lemma 11.5. Quotients of \mathbb{Z} -modules coincide with quotients of abelian groups.

Theorem 11.35 (Universal mapping property for quotient modules). *Let N be a submodule of M , let T be an R -module, and let $f: M \rightarrow T$ be an R -module homomorphism. If $N \subseteq \ker f$, then the function*

$$\begin{aligned} M/N &\xrightarrow{\bar{f}} T \\ m + N &\longmapsto f(m) \end{aligned}$$

is a well-defined R -module homomorphism. In fact, $\bar{f}: M/N \rightarrow T$ is the unique R -module homomorphism such that $\bar{f} \circ q = f$, where $q: M \rightarrow M/N$ denotes the canonical map.

We can represent this in a more visual way by saying that \bar{f} is the unique R -module homomorphism that makes the diagram

$$\begin{array}{ccc} M & \xrightarrow{f} & T \\ q \searrow & \nearrow \exists! \bar{f} & \\ M/N & & \end{array}$$

commute.

Proof. By 817, we already know that \bar{f} is a well-defined homomorphism of groups under $+$ and that it is the unique one such that $\bar{f} \circ q = f$. It remains only to show \bar{f} is an R -linear map:

$$\bar{f}(r(m + N)) = \bar{f}(rm + N) = f(rm) = rf(m) = r\bar{f}(m + N).$$

where the third equation uses that f preserves scaling. \square

Theorem 11.36 (First Isomorphism Theorem). *Let N be an R -module and let $h: M \rightarrow N$ be an R -module homomorphism. Then $\ker(h)$ is a submodule of M and there is an R -module isomorphism $M/\ker(h) \cong \text{im}(h)$.*

Proof. If we forget the multiplication by scalars in R , by the First Isomorphism Theorem for Groups, we know that there is an isomorphism of abelian groups under $+$, given by

$$\begin{aligned}\bar{h} : M/\ker(h) &\xrightarrow{\cong} \text{im}(h) \\ m + \ker(h) &\longmapsto h(m).\end{aligned}$$

It remains only to show this map preserves multiplication by scalars. And indeed:

$$\begin{aligned}\bar{h}(r(m + \ker(h))) &= \bar{h}(rm + \ker(h)) \quad \text{by definition of the } R\text{-action on } M/\ker(h) \\ &= h(rm) \quad \qquad \qquad \qquad \text{by definition of } \bar{h} \\ &= rh(m) \quad \qquad \qquad \qquad \text{since } h \text{ is an } R\text{-module homomorphism} \\ &= r\bar{h}(m + \ker h) \quad \qquad \qquad \qquad \text{by definition of } h.\end{aligned}$$

Theorem 11.37 (Diamond Isomorphism Theorem). *Let A and B be submodules of M , and let $A+B = \{a+b \mid a \in A, b \in B\}$. Then $A+B$ is a submodule of M , $A \cap B$ is a submodule of A , and there is an R -module isomorphism $(A+B)/B \cong A/(A \cap B)$.*

Proof. By Exercise 88, $A+B$ and $A \cap B$ are submodules of M . By the Diamond Isomorphism Theorem for Groups, there is an isomorphism of abelian groups

$$\begin{aligned}h : A/(A \cap B) &\xrightarrow{\cong} (A+B)/B \\ a + (A \cap B) &\longmapsto a + B\end{aligned}$$

It remains only to show h preserves multiplication by scalars:

$$h(r(a + (A \cap B))) = h(ra + A \cap B) = ra + B = r(a + B) = rh(a + (A \cap B)). \quad \square$$

Theorem 11.38 (Cancelling Isomorphism Theorem). *Let A and B be submodules of M with $A \subseteq B$. Then there is an R -module isomorphism $(M/A)/(B/A) \cong M/B$.*

Proof. From 817, we know that B/A is a subgroup of M/A under $+$. Given $r \in R$ and $b + A \in B/A$ we have $r(b + A) = rb + A$ which belongs to B/A since $rb \in B$. This proves B/A is a submodule of M/A . By the Cancelling Isomorphism Theorem for Groups, there is an isomorphism of abelian groups

$$\begin{aligned}(M/A)/(B/A) &\longrightarrow M/B \\ (m + A) + B/A &\longmapsto m + B\end{aligned}$$

and it remains only to show this map is R -linear:

$$\begin{aligned}h(r((m + A) + B/A)) &= h(r(m + A) + B/A) = h((rm + A) + B/A) \\ &= rm + B = r(m + B) \\ &= rh((m + A) + B/A).\end{aligned} \quad \square$$

Theorem 11.39 (Lattice Isomorphism Theorem). *Let R be a ring, let N be a R -submodule of an R -module M , and let $q: M \rightarrow M/N$ be the quotient map. Then the function*

$$\begin{array}{ccc} \{R\text{-submodules of } M \text{ containing } N\} & \xrightarrow{\Psi} & \{R\text{-submodules of } M/N\} \\ K & \longmapsto & K/N \end{array}$$

is a bijection, with inverse defined by

$$\Psi^{-1}(T) := q^{-1}(T) = \{a \in M \mid a + N \in T\}$$

for each R -submodule T of M/N . Moreover, Ψ and Ψ^{-1} preserve sums and intersections of submodules.

Proof. From 817, we know there is a bijection between the set of subgroups of M and those that contain N and subgroups of the quotient group M/N , given by the same map Ψ . We just need to prove that these maps send submodules to submodules. If K is a submodule of M containing N , then by the Cancelling Isomorphism Theorem we know that K/N is a submodule of M/N . If T is a submodule of M/N , then $\pi^{-1}(T)$ is an abelian group, by 817. For $r \in R$ and $m \in \pi^{-1}(T)$, we have $\pi(m) \in T$, and hence $\pi(rm) = r\pi(m) \in T$ too, since T is a submodule. This proves $\pi^{-1}(T)$ is a submodule. \square

11.4 Module generators, bases and free modules

Definition 11.40. Let M be an R -module. A **linear combination** of finitely many elements a_1, \dots, a_n of M is an element of M of the form $r_1m_1 + \dots + r_nm_n$ for some $r_1, \dots, r_n \in R$.

Definition 11.41. Let R be a ring with $1 \neq 0$ and let M be an R -module. For a subset A of M , the submodule of M **generated by** A is

$$RA := \{r_1a_1 + \dots + r_na_n \mid n \geq 0, r_i \in R, a_i \in A\}.$$

We say M is **generated by** A if $M = RA$. If M is an F -vector space, we may say that M is **spanned by** a set A instead of generated by A .

A module M is **finitely generated** if there is a finite subset A of M that generates M . If $A = a$ has a single element, the module $RA = Ra$ is called **cyclic**.

Exercise 94. Let M be an R -module and let $A \subseteq M$. Then RA is the smallest submodule of M containing A , that is

$$RA = \bigcap_{A \subseteq N, N \text{ submodule of } M} N.$$

Exercise 95. Being finitely generated and being cyclic are R -module isomorphism invariants.

Example 11.42. Let R be a ring with $1 \neq 0$.

- (1) $R = R1$ is cyclic.
- (2) $R \oplus R$ is generated by $\{(1, 0), (0, 1)\}$.
- (3) $R[x]$ is generated as an R -module by the set $\{1, x, x^2, \dots, x^n, \dots\}$ of monic monomials in the variable x .
- (4) Let $M = \mathbb{Z}[x, y]$. M is generated by
 - $\{1, x, y\}$ as a ring,
 - $\{1, y, y^2, \dots, y^n, \dots\}$ as an $\mathbb{Z}[x]$ -module, and
 - $\{x^i y^j \mid i, j \in \mathbb{Z}_{\geq 0}\}$ as a group (\mathbb{Z} -module).

Lemma 11.43. *Let R be a ring with $1 \neq 0$, let M be an R -module, and let N be an R -submodule of M .*

- (1) *If M is finitely generated as an R -module, then so is M/N .*
- (2) *If N and M/N are finitely generated as R -modules, then so is M .*

Proof. The proof of (2) will be a problem set question. To show (1), note that if $M = RA$ then $M/N = R\bar{A}$, where $\bar{A} = \{a + N \mid a \in A\}$. \square

Definition 11.44. Let M be an R -module and let A be a subset of M . The set A is **linearly independent** if whenever $r_1, \dots, r_n \in R$ and a_1, \dots, a_n are distinct elements of A satisfying $r_1 a_1 + \dots + r_n a_n = 0$, then $r_1 = \dots = r_n = 0$. Otherwise A is **linearly dependent**.

Definition 11.45. A subset A of an R -module M is a **basis** of M if A is linearly independent and generates M . An R -module M is a **free** R -module if M has a basis.

We will later see that over a field, every module is free. However, when R is not a field, there are R -modules that are not free; in fact, *most* modules are not free.

Example 11.46. Here are some examples of free modules:

- (1) If we think of R as a module over itself, it is free with basis $\{1\}$.
- (2) The module $R \oplus R$ is free with basis $\{(1, 0), (0, 1)\}$.
- (3) The R -module $R[x]$ is free, and $\{1, x, x^2, \dots, x^n, \dots\}$ is a basis.
- (4) Let $M = \mathbb{Z}[x, y]$. Then $\{1, y, y^2, \dots, y^n, \dots\}$ is a basis for the $\mathbb{Z}[x]$ -module M , and $\{x^i y^j \mid i, j \in \mathbb{Z}_{\geq 0}\}$ is a basis for the \mathbb{Z} -module M .

Example 11.47. $\mathbb{Z}/2$ is not a free \mathbb{Z} -module. Indeed suppose that A is a basis for $\mathbb{Z}/2$ and $a \in A$. Then $2a = 0$ so A cannot be linearly independent, a contradiction.

Lemma 11.48. *If A is a basis of M then every nonzero element $0 \neq m \in M$ can be written uniquely as $m = r_1 a_1 + \dots + r_n a_n$ with a_i distinct elements of A and $r_i \neq 0$.*

Proof. Suppose that if $m \neq 0$ and A_1, A_2 are finite subsets of A such that

$$m = \sum_{a \in A_1} r_a a = \sum_{b \in A_2} s_b b$$

for some $r_a, s_b \in R$. Then

$$\sum_{a \in A_1 \cap A_2} (r_a - s_a) a + \sum_{a \in A_1 \setminus A_2} r_a a - \sum_{a \in A_2 \setminus A_1} s_a a = 0.$$

Since A is a linearly independent set, we conclude that $r_a = s_a$ for $a \in A_1 \cap A_2$, $r_a = 0_R$ for $a \in A_1 \setminus A_2$, and $s_a = 0_R$ for $a \in A_2 \setminus A_1$. Set

$$B := \{a \in A_1 \cap A_2 \mid r_a \neq 0_R\}.$$

Then

$$m = \sum_{a \in B} r_a a$$

is the unique way of writing m as a linear combination of elements of A with nonzero coefficients. \square

Theorem 11.49. *Let R be a ring, M be a free R -module with basis B , N be any R -module, and let $j : B \rightarrow N$ be any function. Then there is a unique R -module homomorphism $h : M \rightarrow N$ such that $h(b) = j(b)$ for all $b \in B$.*

Proof. We have two things to prove: existence and uniqueness.

Existence: By Lemma 11.48, any $0 \neq m \in M$ can be written uniquely as

$$m = r_1 b_1 + \cdots + r_n b_n$$

with $b_i \in B$ distinct and $0 \neq r_i \in R$. Define $h : M \rightarrow N$ by

$$\begin{cases} h(r_1 b_1 + \cdots + r_n b_n) = r_1 j(b_1) + \cdots + r_n j(b_n) & \text{if } r_1 b_1 + \cdots + r_n b_n \neq 0 \\ h(0_M) = 0_N & \end{cases}$$

One can check that this satisfies the conditions to be an R -module homomorphism (exercise!).

Uniqueness: Let $h : M \rightarrow N$ be an R -module homomorphism such that $h(b_i) = j(b_i)$. Then in particular $h : (M, +) \rightarrow (N, +)$ is a group homomorphism and therefore $h(0_M) = 0_N$ by properties of group homomorphisms. Furthermore, if $m = r_1 b_1 + \cdots + r_n b_n$ then

$$h(m) = h(r_1 b_1 + \cdots + r_n b_n) = r_1 h(b_1) + \cdots + r_n h(b_n) = r_1 j(b_1) + \cdots + r_n j(b_n)$$

by the definition of homomorphism, and because $h(b_i) = j(b_i)$. \square

Corollary 11.50. *If A and B are sets of the same cardinality, and fix a bijection $j : A \rightarrow B$. If M and N are free R -modules with bases A and B respectively, then there is an isomorphism of R -modules $M \cong N$.*

Proof. Let $g : M \rightarrow N$ and $h : N \rightarrow M$ be the module homomorphisms induced by the bijection $j : A \rightarrow B$ and its inverse $j^{-1} : B \rightarrow A$, which exist by Theorem 11.49. We will show that h and g are inverse homomorphisms. First, note that $g \circ h : N \rightarrow N$ is an R -module homomorphism and $(g \circ h)(b) = g(j^{-1}(b)) = j(j^{-1}(b)) = b$ for every $b \in B$. Since the identity map id_N is an R -module homomorphism and $\text{id}_N(b) = b$ for every $b \in B$, by the uniqueness in Theorem 11.49 we have $g \circ h = \text{id}_N$. Similarly, one shows that $h \circ g = \text{id}_M$. \square

The corollary gives that, up to isomorphism, there is only one free module with basis A , provided such a module exists. But does a free module generated by a given set A exist? It turns out it does.

Definition 11.51. Let R be a ring and let A be a set. The free R -module generated by A , denoted $F_R(A)$ is the set of formal sums

$$\begin{aligned} F_R(A) &= \{r_1a_1 + \cdots + r_na_n \mid n \geq 0, r_i \in R, a_i \in A\} \\ &= \left\{ \sum_{a \in A} r_a a \mid r_a \in R, r_a = 0 \text{ for all but finitely many } a \right\}, \end{aligned}$$

with addition defined by

$$\left(\sum_{a \in A} r_a a \right) + \left(\sum_{a \in A} s_a a \right) = \sum_{a \in A} (r_a + s_a) a$$

and R -action defined by

$$r \left(\sum_{a \in A} r_a a \right) = \sum_{a \in A} (rr_a) a.$$

Exercise 96. This construction $F_R(A)$ results in an R -module, which is free with basis A , and $F_R(A) \cong \bigoplus_{a \in A} R$.

Theorem 11.52 (Uniqueness of rank over commutative rings). *Let R be a commutative ring with $1 \neq 0$ and let M be a free R -module. If A and B are both bases for M , then A and B have the same cardinality, meaning that there exists a bijection $A \rightarrow B$.*

Proof. You will show this in the next problem set (at least in the case where M has a finite basis). \square

Definition 11.53. Let R be a commutative ring with $1 \neq 0$ and let M be a free R -module. The **rank** of M is the cardinality of any basis of M .

Example 11.54. Let R be a commutative ring with $1 \neq 0$. The rank of R^n is n . Note that by Corollary 11.50, any free R -module of rank n must be isomorphic to R^n .

Earlier, we described the R -module structure on the direct sum of R -modules; this is how we construct R^n , by taking the direct sum of n copies of the R -module R . This construction can also be described as the direct product of n copies of R . However, the direct sum and direct product are two different constructions.

Definition 11.55. Let R be a ring. Let $\{M_a\}_{a \in J}$ be a collection of R -modules. The **direct product** of the R -modules M_a is the Cartesian product

$$\prod_{a \in J} M_a := \{(m_a)_{a \in J} \mid m_a \in M_a\}$$

with addition defined by

$$(m_a)_{a \in J} + (n_a)_{a \in J} := (m_a + n_a)_{a \in J}$$

and R -action defined by

$$r(m_a)_{a \in J} = (rm_a)_{a \in J}.$$

The **direct sum** of the R -modules M_a is the R -submodule $\bigoplus_{a \in J} M_a$ of the direct product $\prod_{a \in J} M_a$ given by

$$\bigoplus_{a \in J} M_a = \{(m_a)_{a \in J} \mid m_a = 0 \text{ for all but finitely many } a\}.$$

Exercise 97. The direct sum and the direct product of an arbitrary family of R -modules are R -modules.

Example 11.56. Suppose that $|A| = n < \infty$. Let M_1, \dots, M_n be R -modules. The direct product module $M_1 \times \dots \times M_n$ is the abelian group $M_1 \times \dots \times M_n$ with ring action given by $r(m_1, \dots, m_n) = (rm_1, \dots, rm_n)$ for all $r \in R$ and $m_i \in M_i$. Comparing the definitions we see that

$$M_1 \times \dots \times M_n = M_1 \oplus \dots \oplus M_n.$$

If $M_i = R$ for $1 \leq i \leq n$, then we denote $R^n = \underbrace{R \times \dots \times R}_n = \underbrace{R \oplus \dots \oplus R}_n$.

It is useful to talk about maps from the factors/summands to the direct product/ direct sum and conversely.

Definition 11.57. For $i \in J$ the *inclusion of the i -th factor* into a direct product or direct sum is the map

$$\iota_i: M_i \rightarrow \prod_{a \in J} M_a \text{ or } \iota_i: M_i \rightarrow \bigoplus_{a \in J} M_a, \iota_i(m) = (m_a)_{a \in J}, \text{ where } m_a = \begin{cases} m & \text{if } a = i \\ 0 & \text{if } a \neq i \end{cases}.$$

For $i \in J$ the i -th *projection map* from a direct product or a direct sum module is

$$\pi_i: \prod_{a \in J} M_a \rightarrow M_i \text{ or } \pi_i: \bigoplus_{a \in J} M_a \rightarrow M_i, \pi_i((m_a)_{a \in J}) = m_i.$$

Lemma 11.58. *Projections from direct products or sums of R -module, inclusions into direct products or sums of R -modules, and products of R -module homomorphisms are R -module homomorphisms. Furthermore, inclusions are injective, projections are surjective, and*

$$\pi_i \circ \iota_i = \text{id}_{M_i}.$$

Also, $\iota_i(M_i)$ is an R -submodule of the direct product/sum which is isomorphic to M_i .

Note, however, that $\iota_i \circ \pi_i \neq \text{id}$.

Chapter 12

Vector spaces and linear transformations

12.1 Classification of vector spaces and dimension

Recall that for a subset A of an F -vector space V , the **span** of A , denoted $\text{span}(A)$, is the subspace generated by A :

$$\text{span}(A) := \left\{ \sum_{i=1}^n c_i a_i \mid n \geq 0, c_i \in F, a_i \in A \right\}.$$

Lemma 12.1. *Suppose I is a linearly independent subset of an F -vector space V and $v \in V \setminus \text{span}(I)$, then $I \cup \{v\}$ is also linearly independent.*

Proof. Let w_1, \dots, w_n be any list of distinct elements of $I \cup \{v\}$ and suppose that $\sum_i c_i w_i = 0$ for some $c_i \in F$. If none of the w_i 's is equal to v , then $c_i = 0$ for all i , since I is linearly independent. Without loss of generality, say $w_1 = v$. If $c_1 = 0$ then $c_i = 0$ for all i by the same reasoning as in the previous case. If $c_1 \neq 0$, then

$$v = \sum_{i \geq 2} \frac{c_i}{c_1} w_i \in \text{span}(I),$$

contrary to assumption. This proves that $I \cup \{v\}$ is a linearly independent set. \square

Theorem 12.2 (Every vector space has a basis). *Let V be an F -vector space and assume $I \subseteq S \subseteq V$ are subsets such that I is linearly independent and S spans V . Then there is a subset B with $I \subseteq B \subseteq S$ such that B is a basis.*

Before we prove this theorem, we note the following corollary:

Corollary 12.3. *Every vector space V has a basis, and hence is a free module. Moreover, every linearly independent subset of V is contained in some basis, and every set of vectors that spans V contains some basis.*

Proof. For this first part, apply the theorem with $I = \emptyset$ and $S = V$. For the second and third, use I arbitrary and $S = V$ and $I = \emptyset$ and S arbitrary, respectively. \square

Example 12.4. \mathbb{R} has a basis as a \mathbb{Q} -vector space; just don't ask me what it looks like.

Proof of Theorem 12.2. Let \mathcal{P} denote the collection of all subsets X of V such that $I \subseteq X \subseteq S$ and X is linearly independent. We make \mathcal{P} into a poset by the order relation given by set containment \subseteq . We note that \mathcal{P} is not empty since, for example $I \in \mathcal{P}$.

Let \mathcal{T} be any nonempty chain in \mathcal{P} . Let $Z = \bigcup_{Y \in \mathcal{T}} Y$. We claim $Z \in \mathcal{P}$. Given $z_1, \dots, z_m \in Z$, for each i we have $z_i \in Y_i$ for some $Y_i \in \mathcal{T}$. Since \mathcal{T} is totally ordered, one of Y_1, \dots, Y_m contains all the others and hence contains all the z_i 's. Since Y_i is linearly independent, this shows z_1, \dots, z_m are linearly independent. Thus Z is linearly independent. Since \mathcal{T} is non-empty, $Z \supseteq I$ and hence $Z \in \mathcal{P}$. It is an upper bound for \mathcal{T} by construction.

By Zorn's Lemma, \mathcal{P} has a maximal element B , which we claim is a basis for V . Note that B is linearly independent and $I \subseteq B \subseteq S$ by construction. We need to show that it spans V . Suppose not. Since S spans V , if $S \subseteq \text{span}(B)$, then $\text{span}(B)$ would have to be all of V . So, there is at least one $v \in S$ such that $v \notin \text{span}(B)$, and set $X := B \cup \{v\}$. Clearly, $I \subset X \subseteq S$ and, by Lemma 12.1, X is linearly independent. This shows that X is an element of \mathcal{P} that is strictly bigger than B , contrary to the maximality of B . \square

Corollary 12.5. *Let F be a field and W be a subspace of the F -vector space V . Then every basis of W extends to a basis of V , that is, if B is a basis of W then there exists a basis \tilde{B} of V such that B is a subset of \tilde{B} .*

Proof. Apply Corollary 12.3 with $B = I$ and $S = V$. Since B is a basis of W , B is linearly independent, and B remains linearly independent when regarded as a subset of V . \square

Remark 12.6. It is *not* true that, with the notation of the previous Corollary, if \tilde{B} is a basis of V then there exists a basis B of W such that B is a subset of \tilde{B} . For instance, take $F = \mathbb{R}$, $V = \mathbb{R}^2$, $\tilde{B} = \{(1, 0), (0, 1)\}$ and W the subspace spanned by $(1, 1)$.

Definition 12.7. A vector space is **finite dimensional** if there is spanned by a finite subset.

Thanks to Theorem 12.2, this is equivalent to the property that it has a finite basis. In the language of modules, a finite dimensional vector space is just a finitely generated F -module.

The following is an essential property of vector spaces that eventually will allow us to compare bases in terms of size. We first prepare with some quick remarks on size for possibly infinite sets. We say that two sets X, Y have the same size and write $|X| = |Y|$ if there is a bijective function from X to Y . If there is an injective function $X \hookrightarrow Y$, then we write $|X| \leq |Y|$. It is a consequence of the axiom of choice called the Cantor-Bernstein Theorem that $|X| \leq |Y|$ and $|Y| \leq |X|$ implies $|X| = |Y|$.

Lemma 12.8 (Exchange Property). *Let B be a basis for a vector space V and consider any set of linearly independent vectors $I \subseteq V$. Then there is a subset $A \subseteq B$ such that $|I| = |A|$ and $(B \setminus A) \cup I$ is also a basis for V .*

Proof. First we show we can swap out one element of B for one nonzero element $\{a\}$. In this case, we will show the stronger statement that for any subset $B_0 \subseteq B$, and any element $a \notin \text{span}(B_0)$, there is some $b \in B \setminus B_0$ such that $B \setminus \{b\} \cup \{a\}$ is a basis for V .

Since B is a basis, we can write

$$a = \sum_i \lambda_i b_i$$

for some elements $b_i \in B$. Since $a \notin \text{span}(B_0)$, we have $\lambda_i \neq 0$ for some $b_i \notin B_0$; say $\lambda_1 \neq 0$. We claim that $B' := B \setminus \{b_1\} \cup \{a\}$ is a basis.

First, to see that B' spans, it suffices to show that $b_1 \in \text{span}(B \setminus \{b_1\} \cup \{c\})$ since we then have $B \subseteq \text{span}(B \setminus \{b_1\} \cup \{c\})$, so $V = \text{span}(B) \subseteq \text{span}(B \setminus \{b_1\} \cup \{c\})$. To see that, we can solve the equation above for b_1 in terms of c and the other b_i 's.

Second, to see that B' is linearly independent, note that any nonzero relation on B' have a nonzero coefficient on c , since the other elements are part of a basis and hence linearly independent. Then, given any linear dependence relation on c and b_2, \dots, b_n , we could solve for c and express c in terms of b_2, \dots, b_n . But this would contradict that c has a unique expression in terms of the basis B , in light of the given expression we started with. Thus, B' is a basis.

Consider the collection of pairs (I', A') with $I' \subseteq I$, $A' \subseteq A$, and $|I'| = |A'|$ with the property that $B \setminus A' \cup I'$ is a basis for V . By a Zorn's Lemma argument (left as an exercise), there is a maximal such pair under the partial order $(I', A') \leq (I'', A'')$ if $I' \subseteq I''$ and $A' \subseteq A''$. Let (I_0, A_0) be a maximal element. We will argue that $I_0 = I$.

To obtain a contradiction, suppose otherwise, and let $a \in I \setminus I_0$. Apply the special case above to the basis $(B \setminus A_0) \cup I_0$ and special subset I_0 : since I is linearly independent, $a \notin \text{span}(I_0)$. Then by the special case, there is some $b \in B \setminus A_0$ such that $(B \setminus (A_0 \cup \{b\})) \cup (I_0 \cup \{a\})$ is a basis. This contradicts the maximality of (I_0, A_0) , so we deduce that $I_0 = I$ as required. \square

It follows that all bases for the same vector space have the same cardinality.

Theorem 12.9 (Dimension Theorem). *Any two bases of the same vector space have the same cardinality.*

Proof. Let B, B' be two bases for V . Applying the Exchange Lemma with $C = B'$, there is a subset $C \subseteq B$ with $|C| = |B'|$, so $|B'| \leq |B|$. Swapping roles, one has $|B| \leq |B'|$ as well. \square

Definition 12.10. The **dimension** of a vector space V , denoted $\dim_F(V)$ or $\dim(V)$, is the cardinality of any of its bases.

Example 12.11. $\dim_F(F^n) = |\{e_1, e_2, \dots, e_n\}| = n$.

While one can talk about infinite cardinals, we'll generally say that dimension is a natural number or ∞ .

Theorem 12.12 (Classification of finitely generated vector spaces). *Let F be a field.*

- (1) *Every finitely generated vector space over F is isomorphic to F^n for $n = \dim_F(V)$.*
- (2) *For any $m, n \in \mathbb{Z}_{\geq 0}$, $F^m \cong F^n$ if and only if $m = n$.*
- (3) *Two finite dimensional vector spaces V, V' are isomorphic if and only if $\dim_F(V) = \dim_F(V')$.*

Proof. To show (1), let V be a finite dimensional F -vector space. Then F has a finite spanning set S and by Theorem 12.2 there is a basis $B \subseteq S$ for V . Notice that B is necessarily finite and $V = FB$. Set $|B| = n$ and $B = \{b_1, \dots, b_n\}$. By the UMP for free modules, there is a linear transformation $f : F^n \rightarrow V$ such that $f(e_i) = b_i$ as well as a linear transformation $g : V \rightarrow F^n$ such that $g(b_i) = e_i$. Then both $f \circ g : V \rightarrow V$ and $g \circ f : F^n \rightarrow F^n$ are linear transformations which agree with the identity map on a basis. Hence by the uniqueness part of Theorem 11.49 we have $f \circ g = \text{id}_V$ and $g \circ f = \text{id}_{F^n}$. Therefore, these maps are the desired isomorphisms.

To show (2), let $\varphi : F^m \cong F^n$ be a vector space isomorphism and let B be a basis of F^m . We claim that $\varphi(B)$ is a basis for F^n . Indeed, if

$$\sum_{i=1}^m c_i \varphi(b_i) = 0 \quad \text{then} \quad \varphi \left(\sum_{i=1}^m c_i b_i \right) = 0, \quad \text{so} \quad \sum_{i=1}^m c_i b_i = 0$$

since φ is injective. But B is linearly independent, so we must have $c_i = 0$ for all $1 \leq i \leq m$. If $v \in F^n$, then since B spans F^m we have

$$\varphi^{-1}(v) = \sum_{i=1}^m c_i b_i$$

for some c_i . Thus

$$v = \sum_{i=1}^m c_i \varphi(b_i),$$

which shows $\varphi(B)$ spans F^n . By the Dimension Theorem, we have

$$\dim_F(F^n) = n = |\varphi(B)| = |B| = m. \quad \square$$

Part (3) is immediate from (1) and (2).

Remark 12.13. (1) The same proof as in part (1) of the classification of finitely generated vector spaces above shows that every finitely generated free R -module is isomorphic to R^n for some $n \geq 0$.

- (2) Part (2) of the classification of finitely generated vector spaces can be extended to modules over commutative rings as stated in the previous section; this is a problem in the homework.
- (3) The classification of finitely generated vector spaces yields that dimension is an isomorphism invariant. Moreover, it is a complete isomorphism invariant in the sense that if two module have the same dimension, then they are isomorphic.

Remark 12.14. Part (3) of the previous Theorem holds for infinite-dimensional vector spaces in the following form: two vector spaces V and V' are isomorphic if and only if their dimensions are equal in the sense that there exists a bijection between some basis of V and some basis of V' .

A word on infinite-dimensional vector spaces.

Example 12.15. Consider the vector space $F[x]$. This cannot be a finite dimensional vector space. For instance, if $\{f_1, \dots, f_n\}$ were a basis, then setting

$$M = \max_{1 \leq j \leq n} \{\deg(f_j)\}$$

we see that the element x^{M+1} is not in the span of $\{f_1, \dots, f_n\}$. We can find a basis for this space though. Consider the collection $B = \{1, x, x^2, \dots\}$. This set is linearly independent and spans $F[x]$, thus it forms a basis for $F[x]$. This basis is *countable*, so $\dim_F(F[x]) = |\mathbb{N}|$.

Example 12.16. Consider the real vector space

$$V := \mathbb{R}^{\mathbb{N}} = \mathbb{R} \times \mathbb{R} \times \mathbb{R} \times \dots .$$

This space can be identified with sequences $\{a_n\}$ of real numbers. One might be interested in a basis for this vector space. At first glance, the most obvious choice for a basis would be $E = \{e_1, e_2, \dots\}$. It turns out that E is the basis for the direct sum $\bigoplus_{i \in \mathbb{N}} \mathbb{R}$. However, it is immediate that this set does not span V , as $v = (1, 1, \dots)$ can not be represented as a finite linear combination of these elements. Since v is not in $\text{span}(E)$, then we know that $E \cup \{v\}$ is a linearly independent set. However, this new set $E \cup \{v\}$ does not span V either, as $(1, 2, 3, 4, \dots)$ is not in the span of $E \cup \{v\}$. We know that V has a basis, but it can be shown that no countable collection of vectors forms a basis for this space, and in fact $\dim_{\mathbb{R}}(\mathbb{R}^{\mathbb{N}}) = |\mathbb{R}|$.

Example 12.17. One can show that $\dim_{\mathbb{Q}}(\mathbb{R}) = |\mathbb{R}|$, and $\dim_{\mathbb{Q}}(\mathbb{C}) = |\mathbb{C}| = |\mathbb{R}|$, so $\mathbb{R} \cong \mathbb{C}$ as \mathbb{Q} -vector spaces. In particular, $(\mathbb{R}, +) \cong (\mathbb{C}, +)$ as groups.

We now deduce some formulas that relate the dimensions of various vector spaces.

Theorem 12.18. *Let W be a subspace of a vector space V . Then*

$$\dim(V) = \dim(W) + \dim(V/W).$$

Here the dimension of a vector space is understood to be either a nonnegative integer or ∞ , and the arithmetic of the formula is understood to follow the rules $n + \infty = \infty = \infty + \infty$ for any $n \in \mathbb{Z}_{\geq 0}$. The proof follows from the Problem #1 in Problem Set #2.

Example 12.19. Consider the vector space $V = \mathbb{R}^2$ and its subspace $W = \text{span}\{e_1\}$. Then the quotient vector space V/W is, by definition,

$$V/W = \{(x, y) + W \mid (x, y) \in \mathbb{R}^2\}.$$

Looking at each coset we see that

$$(x, y) + W = (x, y) + \text{span}\{e_1\} = \{(x, y) + (a, 0) \mid a \in \mathbb{R}\} = \{(t, y) \mid t \in \mathbb{R}\},$$

so $(x, y) + W$ is geometrically a line parallel to the x -axis and having the y -intercept y . It is intuitively natural to identify such a line with its intercept, which gives a map

$$V/W \rightarrow \text{span}\{e_2\} \quad (x, y) + W \mapsto (0, y).$$

It turns out that this map is a vector space isomorphism, hence

$$\dim(V/W) = \dim(\text{span}\{e_2\}) = 1$$

and we can check that

$$\dim(W) + \dim(V/W) = 1 + 1 = 2 = \dim(V).$$

If V and W are both infinite dimensional vector spaces, it can happen that V/W is finite dimensional but also that it is infinite dimensional.

Example 12.20. Let $V = F[x]$, which we saw in Example 12.15 is an infinite dimensional vector space over F . Fix a polynomial f with $\deg(f) = d$, and note that the ideal (f) of $F[x]$ generated by f is also an F -vector subspace of $F[x]$ via restriction of scalars. We will show later that $\dim(F[x]/(f)) = d$. In contrast, the subspace E of all even degree polynomials in $F[x]$ together with the zero polynomial satisfies $\dim(F[x]/E) = \infty$.

Definition 12.21. Let $T : V \rightarrow W$ be a linear transformation. The **nullspace** of T is $\ker(T)$. The **rank** of T is $\dim(\text{im}(T))$.

Corollary 12.22 (Rank-Nullity Theorem). *Let $f : V \rightarrow W$ be a linear transformation. Then*

$$\dim(\ker(f)) + \dim(\text{im}(f)) = \dim(V).$$

Proof. By the First Isomorphism Theorem for modules we have $V/\ker(f) \cong \text{im}(f)$, thus

$$\dim(V/\ker(f)) = \dim(\text{im}(V)).$$

By Theorem 12.18, we have

$$\dim(V) = \dim(\ker(V)) + \dim(V/\ker(f)).$$

Thus

$$\dim(V) = \dim(\ker(V)) + \dim(V/\ker(f)) = \dim(\ker(V)) + \dim(\text{im}(V)).$$

□

12.2 Linear transformations and homomorphisms between free modules

Definition 12.23 (The matrix of a homomorphism between free modules). Let R be a commutative ring with $1 \neq 0$. Let V be a finitely generated free R -module of rank n , and let W be a finitely generated free R -module of rank m . Let $B = \{b_1, \dots, b_n\}$ and $C = \{c_1, \dots, c_m\}$ be *ordered* bases of V, W . Given an R -module homomorphism $f : V \rightarrow W$, we define elements $a_{ij} \in R$ for $1 \leq i \leq m$ and $1 \leq j \leq n$ by the formulas

$$f(b_i) = \sum_{j=1}^m a_{j,i} c_j. \tag{12.2.1}$$

The matrix

$$[f]_B^C = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,n} \end{bmatrix}$$

is said to **represent** the homomorphism f with respect to the bases B and C .

Remark 12.24. By ??, the coefficients $a_{j,i}$ in equation 12.2.1 are uniquely determined by the $f(b_i)$ and the elements of C . The coefficients $a_{j,i}$ corresponding to $f(b_i)$ form the i th column of $[f]_B^C$. Note that $[f]_B^C$ is an $m \times n$ matrix with entries in R .

Definition 12.25. Let V and W be finite F -vector spaces of dimension n and m with ordered bases B and C , respectively, and let $f: V \rightarrow W$ be a linear transformation. The matrix $[f]_B^C$ is called the **matrix of the linear transformation** f with respect to the bases B and C .

Example 12.26. If $\text{id}_V: V \rightarrow V$ is the identity automorphism of an n -dimensional free R -module V , then for any basis B of V we have $\text{id}_V(b_i) = b_i$ for all i and hence

$$[\text{id}_V]_B^B = I_n.$$

Example 12.27. Let P_3 denote the the F -vector space of polynomials of degree at most 3 (including the zero polynomial) and consider the linear transformation $d: P_3 \rightarrow P_3$ given by taking the derivative $d(f) = f'$. Let $B = \{1, x, x^2, x^3\}$. Then

$$[d]_B^B = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 3 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Example 12.28. Let F be a field and consider a linear transformation $f: V \rightarrow W$, where $V = F^n$ and $W = F^m$. Consider also the standard ordered bases B and C , i.e. $b_i = e_i \in V$ and $c_i = e_i \in W$. Then for any

$$v = \begin{bmatrix} l_1 \\ \vdots \\ l_n \end{bmatrix} = \sum_i l_i b_i$$

in V we have

$$f \left(\sum_i l_i b_i \right) = \sum_i l_i f(b_i).$$

Each $f(b_i)$ can be written uniquely as a linear combination of the c_j 's as in (12.2.1):

$$f(b_i) = \sum_j a_{j,i} c_j.$$

Then we get

$$f(v) = \sum_i l_i \left(\sum_j a_{j,i} c_j \right) = \sum_j \left(\sum_i a_{j,i} l_i \right) c_j.$$

In other words, we have

$$f(v) = \begin{bmatrix} \sum_i a_{1,i} l_i \\ \vdots \\ \sum_i a_{m,i} l_i \end{bmatrix} = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,n} \end{bmatrix} \cdot \begin{bmatrix} l_1 \\ \vdots \\ l_n \end{bmatrix} = [f]_B^C \cdot v.$$

Then for any

$$v = \sum_i l_i b_i$$

in V we have

$$f\left(\sum_i l_i b_i\right) = \sum_i l_i f(b_i).$$

Each $f(b_i)$ is uniquely expressible as a linear combination of the c_j 's, say

$$f(b_i) = \sum_j a_{j,i} c_j.$$

Then we get

$$f(v) = \sum_i l_i \left(\sum_j a_{j,i} c_j \right) = \sum_j \left(\sum_i a_{j,i} l_i \right) c_j.$$

In other words, we have

$$f(v) = [f]_B^C \cdot v$$

where

$$[f]_B^C = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,n} \end{bmatrix}$$

and $[f]_B^C \cdot v$ denote the usual rule for matrix multiplication.

This says that any linear transformation $f : F^n \rightarrow F^m$ is given by multiplication by a matrix, since we noticed above that $f(v) = [f]_B^C \cdot v$. The same type of statement holds for free modules over commutative rings, and we will show it below in Theorem 12.29.

Theorem 12.29. *Let R be a commutative ring. Let V and W be finitely generated free R -modules of ranks n and m respectively. Fixing ordered bases B for V and C for W gives an isomorphism of R -modules*

$$\text{Hom}_R(V, W) \cong \text{Mat}_{m,n}(R) \quad f \mapsto [f]_B^C.$$

If $V = W$, so that in particular $m = n$, and $B = C$, then the above map is an R -algebra isomorphism $\text{End}_R(V) \cong \text{Mat}_n(R)$.

Proof. Let $\varphi : \text{Hom}_R(V, W) \rightarrow \text{Mat}_{m,n}(R)$ be defined by $\varphi(f) = [f]_B^C$. We need to check that φ is a homomorphism of R -modules, which translates into $[f+g]_B^C = [f]_B^C + [g]_B^C$ and $[\lambda f]_B^C = \lambda[f]_B^C$ for any $f, g \in \text{Hom}_R(V, W)$ and $\lambda \in R$. Let $A = [f]_B^C$ and $A' = [g]_B^C$. Then

$$(f+g)(b_i) = f(b_i) + g(b_i) = \sum_j a_{j,i} c_j + \sum_j a'_{j,i} c_j = \sum_j (a_{j,i} + a'_{j,i}) c_j$$

gives $[f+g]_B^C = A + A'$ and

$$(\lambda f)(b_i) = \lambda \left(\sum_j a_{j,i} c_j \right) = \sum_j (\lambda a_{j,i}) c_j$$

gives $[\lambda f]_B^C = \lambda A$. We leave the proof that for $f, g \in \text{End}_R(V)$ we have $[f \circ g]_B^C = [f]_B^C [g]_B^C$ as an exercise. (This also follows from a more general statement in the next section.)

Finally, the argument described in Example 12.28 also works for any ring R , and it can be adapted for any two chosen basis B and C , showing that φ is a bijection. \square

Corollary 12.30. *For any field F and finite F -vector spaces V and W of dimension n and m respectively, $\dim(\text{Hom}_F(V, W)) = mn$.*

Proof. The isomorphism $\text{Hom}_F(V, W) \cong \text{Mat}_{m,n}(F)$ gives

$$\dim(\text{Hom}_F(V, W)) = \dim(\text{Mat}_{m,n}(F)) = mn.$$

\square

To explain in what sense the matrix $[f]_B^C$ represents the linear transformation f , it is convenient to use the notion of coordinates with respect to a basis. We prepare for this with an exercise that we will also reuse later.

Exercise 98. Let R be a commutative ring and V be a free module with a basis B . Let M be an arbitrary R -module and let $\phi : V \rightarrow M$ be an R -module homomorphism. Then

- (1) ϕ is injective if and only if $\phi(B)$ is linearly independent.
- (2) ϕ is surjective if and only if $\phi(B)$ generates M .
- (3) ϕ is an isomorphism if and only if $\phi(B)$ is a basis for M .

Definition 12.31. Let R be a commutative ring and V be a free module with basis $B = \{b_1, \dots, b_n\}$. Consider the R -module homomorphism $\phi : V \rightarrow R^n$ with $\phi(b_i) = e_i$; there is a unique such map by the UMP for free modules, and it is an isomorphism by the previous exercise. We call $\phi(v)$ the **vector of B -coordinates** of v , denoted $[v]_B$.

Remark 12.32. Note that

$$[v]_B = (r_1, \dots, r_n) \Leftrightarrow v = r_1 b_1 + \dots + r_n b_n$$

since $\phi(r_1 b_1 + \dots + r_n b_n) = r_1 e_1 + \dots + r_n e_n = (r_1, \dots, r_n)$ and ϕ is injective.

Proposition 12.33. Let R be a commutative ring. Let V be a free module with ordered basis B and W be a free module with ordered basis C . Let $f: V \rightarrow W$ be a linear transformation.

$$[f(v)]_C = [f]_B^C \cdot [v]_B$$

for all $v \in V$.

Proof. Let $v \in V$ and write $[v]_B = (r_1, \dots, r_n)$, so $v = \sum_j r_j b_j$. Write $[f]_B^C = [a_{i,j}]$. Then

$$f(v) = f\left(\sum_j r_j b_j\right) = \sum_j r_j f(b_j) = \sum_j r_j \left(\sum_i a_{i,j} c_i\right) = \sum_i \left(\sum_j a_{i,j} r_j\right) c_i.$$

Thus the i -th entry of $[f(v)]_C$ is $\sum_j a_{i,j} r_j$. On the other hand, multiplying out $[f]_B^C \cdot [v]_B = [a_{i,j}](r_1, \dots, r_n)$, the i -th entry is also $\sum_j a_{i,j} r_j$. \square

We will note now an important fact that we will prove later in more generality.

Lemma 12.34. Let F be a field. Any invertible matrix over F is equal to a product of elementary matrices.

12.3 Change of basis

Definition 12.35. Let V be a finitely generated free module over a commutative ring R , and let B and C be bases of V . Let id_V be the identity map on V . Then $[\text{id}_V]_B^C$ is a matrix called the **change of basis matrix** from B to C .

We will show soon that $[\text{id}_V]_B^C$ is invertible with inverse $([\text{id}_V]_B^C)^{-1} = [\text{id}_V]_C^B$.

Example 12.36. Consider the subspace $V = P_2$ of $F[x]$ of all polynomials of degree up to 2, and the bases $B = \{1, x, x^2\}$ and $C = \{1, x - 2, (x - 2)^2\}$ of V . We calculate the change of basis matrix. We have

$$\begin{aligned} \text{id}_V(1) &= 1, \\ \text{id}_V(x) &= 2 \cdot 1 + 1 \cdot (x - 2), \\ \text{id}_V(x^2) &= 4 \cdot 1 + 4 \cdot (x - 2) + 1 \cdot (x - 2)^2. \end{aligned}$$

Thus, the change of basis matrix is given by $[\text{id}_V]_B^C = \begin{bmatrix} 1 & 2 & 4 \\ 0 & 1 & 4 \\ 0 & 0 & 1 \end{bmatrix}$.

Lemma 12.37. If V, W, U are finitely generated free R -modules with ordered bases B, C , and D , respectively, and $f: V \rightarrow W$ and $g: W \rightarrow U$ are R -module homomorphisms, then

$$[g \circ f]_B^D = [g]_C^D \cdot [f]_B^C.$$

Proof. It suffices to check that $[g \circ f]_B^D \cdot p = [g]_C^D \cdot [f]_B^C \cdot p$ for any $p \in R^n$ where $n = \text{rank}(V)$. (In fact, we can just take $p = e_j$ for each j , since Ae_j is the j th column of A .) We can write $p = [v]_B$ for some $v \in V$. Then

$$[g \circ f]_B^D [v]_B = [(g \circ f)(v)]_B = [g(f(p))]_B = [g]_C^D [f(v)]_C = [g]_C^D ([f]_B^C [v]_B) = ([g]_C^D [f]_B^C) [v]_B.$$

\square

Definition 12.38. Let V be a finitely generated free module over a commutative ring R . Two R -module homomorphisms $f, g : V \rightarrow V$ are **similar** if there is a bijective linear transformation $h : V \rightarrow V$ such that $g = h \circ f \circ h^{-1}$. Two $n \times n$ matrices A and B with entries in R are **similar** if there is an invertible $n \times n$ matrix P such that $B = PAP^{-1}$.

Remark 12.39. For elements $A, B \in \mathrm{GL}_n(R)$, the notions of similar and conjugate are the same.

Theorem 12.40. Let V, W be finitely generated free modules over a commutative ring R , let B and B' be bases of V , let C and C' be bases of W , and let $f : V \rightarrow W$ be a homomorphism. Then

$$[f]_{B'}^{C'} = [\mathrm{id}_W]_C^{C'} [f]_B^C [\mathrm{id}_V]_{B'}^B \quad (12.3.1)$$

In particular, if $g : V \rightarrow V$ is an R -module homomorphism, then $[g]_B^B$ and $[g]_{B'}^{B'}$ are similar.

Proof. Since $f = \mathrm{id}_W \circ f \circ \mathrm{id}_V$, by Lemma 12.37 we have

$$[f]_{B'}^{C'} = [\mathrm{id}_W]_C^{C'} [f]_B^C [\mathrm{id}_V]_{B'}^B.$$

Setting $V = W$, $B = C$, $B' = C'$, and $f = \mathrm{id}_V$ in (12.3.1) we have $[\mathrm{id}_V]_{B'}^{B'} = [\mathrm{id}_V]_B^{B'} [\mathrm{id}_V]_B^B [\mathrm{id}_V]_{B'}^B$. Notice that $[\mathrm{id}_V]_B^B = [\mathrm{id}_V]_{B'}^{B'} = I$ is the identity matrix, so the previous formula says that

$$I = [\mathrm{id}_V]_B^{B'} I [\mathrm{id}_V]_{B'}^B.$$

Setting $P = [\mathrm{id}_V]_B^{B'}$, we notice that the previous identity gives $P^{-1} = [\mathrm{id}_V]_{B'}^B$.

Now set $V = W$, $B = C$, $B' = C'$ and $f = g$ in (12.3.1) to obtain

$$[g]_{B'}^{B'} = [\mathrm{id}_V]_B^{B'} [g]_B^B [\mathrm{id}_V]_{B'}^B = P[g]_B^B P^{-1}. \quad \square$$

We now come to certain special changes of basis and their matrices:

Definition 12.41. Let R be a commutative ring with $1 \neq 0$, let M be a free R -module of finite rank n , and let $B = \{b_1, \dots, b_n\}$ be an ordered basis for M . An **elementary basis change operation** on the basis B is one of the following three types of operations to produce a new basis $B' = \{b'_1, \dots, b'_n\}$:

1. Replacing b_j by $rb_i + b_j$ for some $i \neq j$ and some $r \in R$; that is, $b'_j = rb_i + b_j$ and $b'_k = b_k$ for $k \neq i, j$.
2. Replacing b_i by ub_i for some i and some unit u of R ; that is, $b'_i = ub_i$ and $b'_k = b_k$ for $k \neq i$.
3. Swapping the indices of b_i and b_j for some $i \neq j$; that is, $b'_i = b_j$, $b'_j = b_i$, and $b'_k = b_k$ for $k \neq i, j$.

Definition 12.42. Let R be a commutative ring with $1 \neq 0$. An **elementary column operation** on a matrix $A \in \mathrm{Mat}_{m,n}(R)$ is one of the following three types of operations:

1. Adding an element of R times a column of A to a different column of A .
2. Multiplying a column of A by a unit of R .

3. Interchanging two columns of A .

We define a **elementary row operation** analogously.

Definition 12.43. Let R be a commutative ring with $1 \neq 0$. An **elementary matrix** over R is an $n \times n$ matrix of one of the following three forms:

- (1) For $r \in R$ and $1 \leq i, j \leq n$ with $i \neq j$, let $E_{i,j}(r)$ be the matrix with 1s on the diagonal, r in the (i, j) position, and 0 everywhere else.
- (2) For $u \in R^\times$ and $1 \leq i \leq n$ let $E_i(u)$ denote the matrix with (i, i) entry u , (j, j) entry 1 for all $j \neq i$, and 0 everywhere else.
- (3) For $1 \leq i, j \leq n$ with $i \neq j$, let $E_{(i,j)}$ denote the matrix with 1 in the (i, j) and (j, i) positions and in the (l, l) positions for all $l \notin \{i, j\}$, and 0 in all other entries.

Remark 12.44. The elementary matrices $E_i(u)$ and $E_{(i,j)}$ are symmetric and the transpose of $E_{i,j}(r)$ is $E_{j,i}(r)$. In particular, the transpose of an elementary matrix is an elementary matrix.

Lemma 12.45. Let E be an $n \times n$ elementary matrix.

- (1) E is the change of basis matrix $[\text{id}]_{B'}^B$ for the corresponding elementary basis change operation from B to B' .
- (2) If $B \in \text{Mat}_{m,n}(R)$, then the result of performing the corresponding elementary column operation on B is the product matrix BE . Explicitly,

- $AE_{i,j}(r)$ is the matrix obtained from A by replacing

$$\text{col}_j(A) \rightsquigarrow \text{col}_j(A) + r \cdot \text{col}_i(A).$$

- $AE_i(u)$ is the matrix obtained from A by replacing

$$\text{col}_i(A) \rightsquigarrow u \cdot \text{col}_i(A).$$

- $AE_{(i,j)}$ is the matrix obtained from A by replacing

$$\begin{aligned} \text{col}_i(A) &\rightsquigarrow \text{col}_j(A) \\ \text{col}_j(A) &\rightsquigarrow \text{col}_i(A) \end{aligned}$$

- (3) If $B \in \text{Mat}_{n,q}(R)$, then the result of performing the corresponding elementary row operation on A is the product matrix $E^T B$. Explicitly,

- $E_{i,j}(r)B$ is the matrix obtained from B by replacing

$$\text{row}_i(A) \rightsquigarrow \text{row}_i(A) + r \cdot \text{row}_j(A).$$

- $E_i(u)B$ is the matrix obtained from B by replacing

$$\text{row}_i(A) \rightsquigarrow u \cdot \text{row}_i(A).$$

- $E_{(i,j)}B$ is the matrix obtained from B by replacing

$$\begin{aligned}\text{row}_i(A) &\rightsquigarrow \text{row}_j(A) \\ \text{row}_j(A) &\rightsquigarrow \text{row}_i(A)\end{aligned}$$

Proof. (1) By definition, the j -th column of $[\text{id}]_B^B$ gives the coefficients for b'_j as a linear combination of the elements of B . In each case, we check that the matrix E agrees with the specified combinations in the definition of the basis operation.

- (2) It suffices to check this for a row vector, since the i -th row of BE can be computed as the i -th row of B multiplies by E . Then one can verify this by case-by-case multiplication.
- (3) Similar to (2).

□

Remark 12.46. To remember the relationship between elementary matrices and elementary operations, it suffices to remember that

- (1) Row operations correspond to multiplication on the left and column operations correspond to multiplication on the right, and
- (2) The elementary matrix corresponding to an elementary row or column operation is the matrix that results from applying that operation to the identity matrix.

Indeed, (2) follows from taking $A = I$ or $B = I$ in the Lemma.

12.4 Determinants

We briefly cover some of the key facts about determinants that we will need later.

Definition 12.47. Let R be a commutative ring. We define the function

$$\det : \text{Mat}_{n \times n}(R) \rightarrow R$$

by the rule

$$\det(A) = \sum_{\sigma \in S_n} \text{sgn}(\sigma) \prod_{i=1}^n a_{i,\sigma(i)}$$

for a matrix $A = [a_{i,j}]$. We call $\det(A)$ the **determinant** of A .

Example 12.48. If $A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$, then $\det(A) = a_{11}a_{22} - a_{12}a_{21}$.

If A is an upper triangular matrix, so that $a_{ij} = 0$ for $j > i$, then there is only one nonzero term in the sum, and $\det(A)$ is the product of the diagonal entries.

Definition 12.49. Let R be a commutative ring. Let $\phi : \underbrace{R^n \times \cdots \times R^n}_{n-\text{times}} \rightarrow R$ be a function.

We say that

(1) ϕ is **multilinear** if for each $i = 1, \dots, n$ we have

$$\phi(v_1, \dots, v_{i-1}, v_i + v'_i, v_{i+1}, \dots, v_n) = \phi(v_1, \dots, v_{i-1}, v_i, v_{i+1}, \dots, v_n) + \phi(v_1, \dots, v_{i-1}, v'_i, v_{i+1}, \dots, v_n)$$

and

$$\phi(v_1, \dots, v_{i-1}, rv_i, v_{i+1}, \dots, v_n) = \phi(v_1, \dots, v_{i-1}, v_i, v_{i+1}, \dots, v_n) + r\phi(v_1, \dots, v_{i-1}, v_i, v_{i+1}, \dots, v_n)$$

for all $v_1, \dots, v_n, v'_i \in R^n$ and $r \in R$; i.e., when all but one entry is fixed, the function $R^n \rightarrow R$ in the remaining output is an R -module homomorphism.

(2) ϕ is **alternating** if $\phi(v_1, \dots, v_n) = 0$ whenever $v_i = v_j$ for some $i \neq j$.

Lemma 12.50. *Let $\phi : \underbrace{R^n \times \cdots \times R^n}_{n\text{-times}} \rightarrow R$ be a multilinear alternating function. Then for any $\sigma \in S_n$ and any vectors $v_1, \dots, v_n \in R^n$, we have*

$$\phi(v_{\sigma(1)}, v_{\sigma(2)}, \dots, v_{\sigma(n)}) = \text{sgn}(\sigma)\phi(v_1, v_2, \dots, v_n).$$

Proof. First, we consider the case of the transposition (1 2). Note that

$$\begin{aligned} 0 &= \phi(v_1 + v_2, v_1 + v_2, \dots, v_n) = \phi(v_1, v_1 + v_2, \dots, v_n) + \phi(v_2, v_1 + v_2, \dots, v_n) \\ &= \phi(v_1, v_1, \dots, v_n) + \phi(v_2, v_1, \dots, v_n) + \phi(v_1, v_2, \dots, v_n) + \phi(v_2, v_2, \dots, v_n) \\ &= \phi(v_2, v_1, \dots, v_n) + \phi(v_1, v_2, \dots, v_n), \end{aligned}$$

so $\phi(v_2, v_1, \dots, v_n) = -\phi(v_1, v_2, \dots, v_n)$. The case of an arbitrary transposition follows in the same way. For an arbitrary permutation σ , we can write σ as a product of transpositions, and the claim follows by a straightforward induction. \square

Theorem 12.51. *Let R be a commutative ring. Identify $\text{Mat}_{n \times n}(R)$ with $\underbrace{R^n \times \cdots \times R^n}_{n\text{-times}}$ mapping A to the n -tuple of columns of A . Then \det is the unique function $\text{Mat}_{n \times n} \rightarrow R$ that is multilinear, alternating, and satisfies $\det(I) = 1$.*

Sketch of proof. The verification that \det has these properties is straightforward but messy. To show uniqueness, we can use multilinearity to show that the value of a function with these properties is determined by the values when each column is a standard vector e_i . We can then use the alternating property and the Lemma to show that the value is determined by the value at the identity matrix. \square

Our next goal is to prove the familiar multiplicative property for determinants.

Proposition 12.52. *Let R be a commutative ring. Let A be a square matrix and let B be a matrix obtained from A by a single elementary column operation:*

- If the operation is of type I, $\det(B) = \det(A)$.
- If the operation is of type II, given by multiplying a column of A by a unit u , then $\det(B) = u \det(A)$.

- If the operation is of type III, $\det(B) = -\det(A)$.

In particular, if A is an arbitrary matrix and E is an elementary matrix, then $\det(EA) = \det(E)\det(A)$.

Proof. The first case follows from multilinearity and alternating properties: For notational simplicity say $A = (v_1, v_2, \dots)$ and $B = (v_1 + rv_2, v_2, \dots)$. Then

$$\det(B) = \det(v_1, v_2, \dots) + r \det(v_2, v_2, \dots) = \det(A) + r \cdot 0 = \det(A)$$

The second case is immediate from (the second part of) R -multilinearity. The last case is a special case of Lemma.

The final claim comes from noting that $\det(E) = 1, u, -1$ in the three cases, respectively. \square

Corollary 12.53. For $R = F$ a field, we have $\det(A) \neq 0$ if and only if A is invertible.

Proof. If A is not invertible, then the span of the columns of A is a proper subspace of F^n and hence the columns of A must be linearly dependent. Say the i -th column is a linear combination of the rest: $v_i = \sum_{j \neq i} c_j v_j$. Then

$$\det(v_1, \dots, v_n) = \sum_{j \neq i} c_j \det(\text{a matrix with the } i\text{-th and } j\text{-th columns equal}) = 0.$$

If A is invertible, then we can write A as a product of elementary matrices (this is a result that we stated before, but will prove soon). The result thus follows from the Proposition and the fact that $\det(I_n) = 1$. \square

Theorem 12.54. Let R be an integral domain. Then for any matrices $A, B \in \text{Mat}_{n \times n}(R)$ we have

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

Proof. First we will consider the case where $R = F$ is a field.

If A is not invertible, neither is AB , since $\text{im}(AB) \subseteq \text{im}(A)$, and if B is not invertible, neither is AB , since $\ker(AB) \supseteq \ker(A)$. So, by the Proposition, if either A or B is not invertible, both sides of the equation are 0.

Assume now that A and B are both invertible. Then by the Proposition we have

$$A = E_1 \cdots E_n$$

and

$$B = F_1 \cdots F_m$$

and hence

$$AB = E_1 \cdots E_n F_1 \cdots F_m$$

where the E_i 's and F_j 's are elementary matrices.

Applying Corollary ?? repeatedly gives

$$\det(AB) = \det(E_1 \cdots E_n F_1 \cdots F_{m-1}) \det(F_m) = \cdots = \det(E_1) \cdots \det(E_n) \det(F_1) \cdots \det(F_m)$$

and similarly

$$\det(A) \det(B) = (\det(E_1) \cdots \det(E_n)) (\det(F_1) \cdots \det(F_m)).$$

Now, for an integral domain R , consider its fraction field F , and identify $R \subseteq F$ as a subring. To compute $\det(A)$, $\det(B)$, and $\det(AB)$ we can replace R by F , and are done by the field case. \square

Even when A and B aren't square, we can still say the following.

Proposition 12.55. *Let R be a commutative ring. Let $A \in \text{Mat}_{n \times m}(R)$ and $B \in \text{Mat}_{m \times n}(R)$ with $m \geq n$. For a subset $I = \{i_1, \dots, i_n\} \subseteq [m]$ with $|I| = n$, let A_I denote the submatrix of A with columns indexed by I (in increasing order). Then*

$$\det(AB) \in (\{\det(A_I) \mid I \subseteq [m], |I| = n\}).$$

Proof. Let a_1, \dots, a_m be the columns of A . We can write the j -th column of AB as $\sum_{i=1}^m b_{i,j} a_i$. Then, by multilinearity,

$$\begin{aligned} \det(AB) &= \det \left[\sum_{i_1=1}^m b_{i_1,1} a_{i_1} \quad \cdots \quad \sum_{i_n=1}^m b_{i_n,n} a_{i_n} \right] \\ &= \sum_{1 \leq i_1, \dots, i_n \leq m} b_{i_1,1} \cdots b_{i_n,n} \det \begin{bmatrix} a_{i_1} & \cdots & a_{i_n} \end{bmatrix}. \end{aligned}$$

By the alternating property, we can rewrite each $\det \begin{bmatrix} a_{i_1} & \cdots & a_{i_n} \end{bmatrix}$ as either zero, or the determinant of a submatrix with columns $i_1 < i_2 < \cdots < i_n$, up to sign. This gives $\det(AB)$ as an R -linear combination of the determinants $\det(A_I)$. \square

Chapter 13

Finitely generated modules over PIDs

We have seen that every module over a field is free. In contrast, whenever R is a commutative ring that is not a field, we can always construct modules that are not free. We will see that, however, every module is still a quotient of a free module. Describing that quotient explicitly is to give a presentation for the module, similarly to how we gave presentations for groups. We will study the particular case of finitely generated modules over PIDs in more detail.

13.1 The module presented by a matrix

Writing a given R -module M as a quotient of a free module is giving a **presentation** for M . In 817, we studied presentations for groups; these consisted of a set of generators and a set (normal subgroup) of relations among these generators. Presentations are important for modules as well. In this case, the relations are encoded by a matrix, or equivalently by a homomorphism between a pair of free modules. We study below how the change of basis techniques can be applied to unravel the structure of a module starting with its presentation.

Definition 13.1. Let R be a commutative ring with $1 \neq 0$, let $A \in \text{Mat}_{m,n}(R)$, and let $t_A : R^n \rightarrow R^m$ be the R -module homomorphism represented by A with respect to the standard bases; i.e., the homomorphism given by the rule $t_A(v) = Av$. The **R -module presented by A** is the R -module $R^m/\text{im}(t_A)$.

The R -module M presented by $A \in \text{Mat}_{m,n}(R)$ has m generators and n relations. Each row of A corresponds to a generator for M , while each column encodes a relation among those generators. More precisely, the relations among the m generators are themselves *generated* by the n generators of $\text{im}(t_A)$, which are the images of the standard basis of R^n by t_A .

Example 13.2. The \mathbb{Z} -module $M = \mathbb{Z}/6$ is presented by

$$\mathbb{Z} \xrightarrow{6} \mathbb{Z},$$

since $M \cong \mathbb{Z}/\text{im}(t_6) = \mathbb{Z}/(6)$. Notice here we abused notation and wrote 6 instead of the 1×1 matrix [6].

Example 13.3. Let $R = k[x, y]$, where k is a field, and $I = (x, y)$. The R -module $M = R/I$ has 1 generator, $m = 1 + I$, so we can write a presentation for M of the form $F \xrightarrow{p} R$ for some free module F and some R -module homomorphism p . To find such an F , we need to ask about the relations among the generators of M . For any $a \in I$, we have the relation $am = 0$, so I is the **module of relations** for this presentation of M .

How many generators does the module of relations have? In this case, we need 2: the relations $xm = 0$ and $ym = 0$ generate *all* the relations, since for any $a \in I$, we can write $a = rx + sy$ for some $x, y \in R$, and thus $am = 0$ can be rewritten as $r(xm) + s(ym) = 0$, which is a linear combination of the two relations $xm = 0$ and $ym = 0$. Finally, we have the following presentation for M :

$$R^2 \xrightarrow{\begin{bmatrix} x & y \end{bmatrix}} R.$$

Indeed, the image of $\begin{bmatrix} x & y \end{bmatrix}$ is (x, y) , and $M \cong R/(x, y)$.

Conversely, we might be given a matrix and ask about what module it represents; one thing to keep in mind is that some presentations might be inefficient, either by having more generators or more relations than necessary. We want to answer to key questions: given a presentation for a module, how to find a more efficient presentation; and how to decide if two different presentations actually give us isomorphic modules. Keeping these goals in mind, let's try a more elaborate example.

Example 13.4. Consider the matrix

$$A = \begin{bmatrix} 2 & 1 & 0 \\ 3 & 9 & 5 \\ 1 & -2 & 7 \\ 0 & 1 & 2 \end{bmatrix}.$$

What \mathbb{Z} -module M is presented by A ? Formally, M is the quotient module $M = \mathbb{Z}^4/\text{im}(t_A)$, where $t_A : \mathbb{Z}^3 \rightarrow \mathbb{Z}^4$ is defined by $t_A(v) = Av$. Since \mathbb{Z}^4 is generated by its standard basis elements $\{e_1, e_2, e_3, e_4\}$, we deduce as in Lemma 11.43 that $M = \mathbb{Z}^4/\text{im}(t_A)$ is generated by the cosets of the e_i . To keep the notation short, we set $m_i = e_i + \text{im}(t_A)$.

Let $N = \text{im}(t_A)$ and note that N is the submodule of \mathbb{Z}^4 generated by the columns of A :

$$N = R \left\{ \begin{bmatrix} 2 \\ 3 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 9 \\ -2 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 5 \\ 7 \\ 2 \end{bmatrix} \right\} = R\{2e_1 + 3e_2 + e_3, e_1 + 9e_2 - 2e_3 + e_4, 5e_2 + 7e_3 + 2e_4\}.$$

Since N maps to 0 under the quotient map $q : \mathbb{Z}^4 \rightarrow M = \mathbb{Z}^4/N$, the relations of M can be written as

$$\begin{cases} 2m_1 + 3m_2 + m_3 = 0 \\ m_1 + 9m_2 - 2m_3 + m_4 = 0 \\ 5m_2 + 7m_3 + 2m_4 = 0. \end{cases}$$

We can now see that this is a rather inefficient presentation, since we can clearly use the first equation to solve for $m_3 = -2m_1 - 3m_2$. This implies that M can be generated using only m_1, m_2 and m_4 , that is

$$M = R\{m_1, m_2, m_3, m_4\} = R\{m_1, m_2, m_4\}.$$

This eliminates the first equation and the latter two become

$$\begin{cases} 5m_1 + 15m_2 + m_4 = 0 \\ -14m_2 - 16m_2 + 2m_4 = 0 \end{cases}$$

Now we can also eliminate m_4 , i.e leaving just two generators m_1, m_2 that satisfy

$$-24m_1 - 46m_2 = 0.$$

Another way to do this is to look at the matrix A and use elementary row operations to “make zeros” on the 1st and 2nd columns, as follows:

$$A = \begin{bmatrix} 2 & 1 & 0 \\ 3 & 9 & 5 \\ 1 & -2 & 7 \\ 0 & 1 & 2 \end{bmatrix} \rightarrow \begin{bmatrix} 0 & 5 & -14 \\ 0 & 15 & -16 \\ 1 & -2 & 7 \\ 0 & 1 & 2 \end{bmatrix} \rightarrow \begin{bmatrix} 0 & 0 & -24 \\ 0 & 0 & -46 \\ 1 & 0 & 13 \\ 0 & 1 & 0 \end{bmatrix}$$

Eliminating the generators m_3 and m_4 amounts to dropping the first two columns (which are the 3rd and 4th standard basis vectors) as well as the last two rows. As we will prove soon, this shows that the \mathbb{Z} -module presented by A is isomorphic to the \mathbb{Z} -module presented by

$$B = \begin{bmatrix} -24 \\ -46 \end{bmatrix}.$$

We can go further. Set $m'_1 := m_1 + 2m_2$. Then m'_1 and m_2 also form a generating set of M . The relation on m_1, m_2 translates to

$$-24m'_1 + 2m_2 = 0$$

given by the matrix

$$C = E_{2,1}(-2)B = \begin{bmatrix} -24 \\ 2 \end{bmatrix}.$$

Note that we have done a row operation (subtract twice row 1 from row 2) to get from B to C . Continuing in this fashion by adding 12 row 2 to row 1 we also form

$$D = E_{1,2}(12)C = \begin{bmatrix} 0 \\ 2 \end{bmatrix},$$

The last matrix D presents the module $M' = \mathbb{Z}^2/\text{im}(t_D)$ with generators a, b , where

$$a = e_1 + \text{im}(t_D), \quad b = e_2 + \text{im}(t_D)$$

and relation $2a = 0$. This module M' is isomorphic to our original module M . As we will see, this proves $M \cong \mathbb{Z} \oplus \mathbb{Z}/2$. An explicit isomorphism between M' and $\mathbb{Z} \oplus \mathbb{Z}/2$ is given by sending $\mathbb{Z}^2 \rightarrow \mathbb{Z} \oplus \mathbb{Z}/2$ by the unique \mathbb{Z} -module homomorphism defined by

$$e_1 \mapsto (1, 0) \text{ and } e_2 \mapsto (0, [1]_2).$$

Now notice that the kernel of this homomorphism is the submodule $(2e_2)\mathbb{Z} = \text{im}(t_D)$. Then the first isomorphism theorem gives $M' = \mathbb{Z}^2/\text{im}(t_D) \cong \mathbb{Z} \oplus \mathbb{Z}/2$.

Lemma 13.5. *Let R be a commutative ring with $1 \neq 0$, $A \in \text{Mat}_{m,n}(R)$ and $B \in M_{m',n'}(R)$ for some $m, n, m', n' \geq 1$. Then A and B present isomorphic R -modules if B can be obtained from A by any finite sequence of operations of the following form:*

- (1) *an elementary row operation,*
- (2) *an elementary column operation,*
- (3) *deletion of the j th column and i th row of A if $Ae_j = e_i$, that is, if the j th column of A is the vector e_i ,*
- (4) *the reverse of 3: insertion of a row and column satisfying $Ae_j = e_i$,*
- (5) *deletion of a column of all 0's,*
- (6) *the reverse of 5: insertion of a column of all 0s.*

Proof. It is sufficient to show that each individual operation gives an isomorphism, as the composition of isomorphisms is an isomorphism.

For operations (1) and (2), consider matrices A and A' where A' is obtained from A by the given elementary row/column operation, and set $M = R^m/\text{im}(t_A)$ and $M' = R^{m'}/\text{im}(t_{A'})$. We need to prove that there is an isomorphism $M \cong M'$.

In case (1), where we have an elementary row operation, let E be the corresponding elementary matrix. Since $A' = EA$, the isomorphism $E : R^n \rightarrow R^n$ maps $\text{im}(A)$ bijectively onto $\text{im}(A')$. Thus E induces an isomorphism

$$M = R^m/\text{im}(t_A) \xrightarrow{\cong} R^m/\text{im}(t_{A'}) = M'.$$

In case (2), where we have an elementary column operation, let E be the corresponding elementary matrix. Since $A' = AE$ and since E is an isomorphism, we have

$$\text{im}(t_{A'}) = \text{im}(t_{AE}) = \text{im}(t_A \circ t_E) = \text{im}(t_A)$$

and so $m = m'$ and $M = R^m/\text{im}(t_A) = R^{m'}/\text{im}(t_{A'}) = M'$. In fact, note that for this one we get equality, not merely an isomorphism.

For case (3), we have $m' = m - 1$ and $n' = n - 1$. Since R^m is free, by the [UMP for free modules](#) there is a unique R -module homomorphism $p : R^m \rightarrow R^{m-1}$ sending

$$\begin{aligned} e_1 &\mapsto e'_1, \dots, e_{i-1} \mapsto e'_{i-1} \\ e_i &\mapsto 0 \\ e_{i+1} &\mapsto e'_i, \dots, e_m \mapsto e'_{m-1} \end{aligned}$$

Similarly, there is a unique R -module homomorphism $q: R^n \rightarrow R^{n-1}$ sending

$$\begin{aligned} e_1 &\mapsto e'_1, \dots, e_{j-1} \mapsto e'_{j-1}, \\ e_j &\mapsto 0, \\ e_{j+1} &\mapsto e'_j, \dots, e_n \mapsto e'_{n-1}. \end{aligned}$$

Here the elements e_i are part of a standard basis for R^n or for R^m , while the elements e'_i are part of a standard basis for R^{n-1} or for R^{m-1} . Then the diagram

$$\begin{array}{ccc} R^n & \xrightarrow{A} & R^m \\ q \downarrow & & \downarrow p \\ R^{n-1} & \xrightarrow{A'} & R^{m-1} \end{array}$$

commutes by the definition of A' . In particular, $p(\text{im}(t_A)) \subseteq \text{im}(t_{A'})$ and so p induces an R -module homomorphism

$$\bar{p}: M \rightarrow M',$$

and we claim \bar{p} is bijective.

Since p is onto, so is \bar{p} . Suppose $m \in \ker(\bar{p})$. Then $m = v + \text{im}(t_A)$ for some $v \in R^m$ and $p(v) \in \text{im}(t_{A'})$. Say $p(v) = A'w$. Since q is onto, $w = q(u)$ for some u . Then

$$p(v - Au) = p(v) - pA(u) = p(v) - A'q(u) = p(v) - A'w = p(v) - p(v) = 0,$$

and thus $v - Au \in \ker(p)$. Now, the kernel of p is clearly Re_i , so that $v - Au = re_i$ for some r . Finally, since $Ae_j = e_i$, we have $A(re_j) = re_i = v - Au$ and hence $v = A(u + re_j)$, which proves $v = t_A(u + re_j) \in \text{im}(t_A)$ and hence that $m = 0$.

For (5), it is clear that the columns of A' generate the same submodule of R^m as do the columns of A , and thus $M = M'$.

Finally, for operations (4) and (6), since the isomorphism relation is reflexive, the statements of parts (3) and (5) show that parts (4) and (6) are true as well. \square

13.2 Existence of presentations

Which modules have presentations? If we take this in a broad sense, the answer is every module.

Theorem 13.6. *Let R be a ring and M be a module. Then there exist free modules F, G and a homomorphism $\alpha: G \rightarrow F$ such that $M \cong F/\text{im}(\alpha)$.*

Proof. Let S be a generating set for M , and let $F = F_R(S)$ be the free module with basis S . By the UMP for free modules, there is a unique homomorphism ϕ such that for each $s \in S$, $\phi(s) = s$. Since S generates M , this map is surjective by an Exercise from earlier.

Let $K = \ker(\phi)$, and let S' be a generating set for K . Set $G = F_R(S')$ to be the free module with basis S' . By the UMP for free modules again, there is a unique $\alpha: G \rightarrow F$ such that for each $s' \in S'$, $\alpha(s') = s'$. We claim that $\text{im}(\alpha) = K$ (left as an exercise).

Thus, the First Isomorphism Theorem, $M \cong F/K = F/\text{im}(\alpha)$. \square

Suppose that R is commutative. In the setting of the previous Theorem, if F is free of rank m , and G is free of finite rank n , then by picking bases for F and G , we can rewrite $\alpha : G \rightarrow F$ as $t_A : R^n \rightarrow R^m$ for some matrix A .

Since this would yield m generators for M , this can only happen if M is finitely generated. This in general does not suffice to guarantee that there will only be finitely many generators for the submodule of relations.

It might seem like no submodule of a finitely generated module could ever fail to itself be finitely generated, but indeed this happens!

Example 13.7. Let k be a field and $R = k[x_1, x_2, \dots]$ be a polynomial ring in infinitely many variables. When we think of R as a module over itself, it is finitely generated, by the element 1. However, there are submodules of R that are not finitely generated: for example, the ideal (x_1, x_2, \dots) generated by all the variables.

Theorem 13.8. *Let R be a PID. Then every submodule of a finitely generated module is also finitely generated.*

Proof. We will first prove that for each $n \geq 1$, every submodule of R^n is finitely generated. The base case $n = 1$ holds by the definition, since a submodule of R^1 is the same thing as an ideal of R , and every ideal is generated by one element. Assume $n > 1$ and that every submodule of R^{n-1} is finitely generated. Let M be any submodule of R^n . Define

$$\pi : R^n \twoheadrightarrow R^1$$

to be the projection onto the last component of R^n . The kernel of π may be identified with R^{n-1} , and so $N := \ker(\pi) \cap M$ is a submodule of R^{n-1} . By assumption, N is finitely generated. The image $\pi(M)$ is a submodule of R^1 , that is, an ideal of R , and so it too is principal. Furthermore, by the [First Isomorphism Theorem](#) $M/\ker(\pi) \cong \pi(M)$. By Lemma 11.43, we deduce that M is a finitely generated module.

Now let T be any finitely generated R -module and $N \subseteq T$ any submodule. Since T is finitely generated, there exists a surjective R -module homomorphism $q : R^n \twoheadrightarrow T$ for some n . Then $q^{-1}(N)$ is a submodule of R^n and hence it is finitely generated by the case we already proved, say by element $v_1, \dots, v_m \in q^{-1}(N)$. We claim that $q(v_1), \dots, q(v_m)$ generate N . Given any $a \in N$, since q is surjective we can find some $b \in q^{-1}(N)$ such that $q(b) = a$. Since v_1, \dots, v_m generated $q^{-1}(N)$, we can find $c_1, \dots, c_m \in R$ such that

$$b = c_1v_1 + \dots + c_mv_m \implies c_1q(v_1) + \dots + c_mq(v_m) = q(c_1v_1 + \dots + c_mv_m) = q(b) = a. \quad \square$$

Theorem 13.9. *Any finitely generated module M over a PID R has a finite presentation given by an $m \times n$ matrix A , that is, there is an isomorphism*

$$M \cong R^m / \text{im}(t_A),$$

where $t_A : R^n \rightarrow R^m$ is the map on free modules $t_A(v) = Av$ induced by A .

Proof. Let M be a finitely generated module over a PID. We follow the argument of Theorem 13.6. Choose a finite generating set y_1, \dots, y_m of M and obtain an R -module map $\pi : R^m \rightarrow M$ that sends e_i to y_i , by using the [UMP for free modules](#). Since every element in

M is given as a linear combination of the y_i , the map π is surjective. Notice, however, that this representation as a linear combination of the y_i is not necessarily unique, so π might have a nontrivial kernel.

Since R^m is finitely generated and R is a PID, the submodule $\ker(\pi)$ is also finitely generated, say by z_1, \dots, z_n . This too leads to a surjective R -module map $g: R^n \rightarrow \ker(\pi)$ that sends $e_i \mapsto z_i$. The composition of $g: R^n \twoheadrightarrow \ker(\pi)$ followed by the inclusion of $\iota: \ker(\pi) \hookrightarrow R^m$ is an R -module homomorphism $t = \iota \circ g: R^n \rightarrow R^m$ and hence by Theorem 12.29 we know t is given by a $m \times n$ matrix $A = [t]_B^C$ with respect to the standard bases of R^m and R^n respectively, meaning $t = t_A$.

It remains to show that $M \cong R^m/\text{im}(t_A)$. First note that since $t_A = \iota \circ g$ and g is surjective we have

$$\text{im}(t_A) = \text{im}(\iota \circ g) = \iota(\text{im}(g)) = \iota(\ker(\pi)) = \ker(\pi).$$

By the [First Isomorphism Theorem](#) we now have

$$M = \text{im}(\pi) \cong R^m / \ker(\pi) = R^m / \text{im}(t_A). \quad \square$$

13.3 Classification of finitely generated modules over PIDs

We just showed that any finitely generated module M over a PID has a finite presentation matrix A . We will discuss a canonical form for such a matrix A and the consequences it has on determining the isomorphism type of M .

Theorem 13.10 (Smith Normal Form (SNF)). *Let R be a PID and let $A \in \text{Mat}_{m,n}(R)$. Then there exist invertible matrices P and Q such that $M = PAQ = [a_{ij}]$ satisfies the following: all nondiagonal entries of M are 0, meaning $a_{ij} = 0$ if $i \neq j$, and the diagonal entries of M satisfy*

$$a_{11} \mid a_{22} \mid a_{33} \mid \dots.$$

Moreover, the number ℓ of nonzero entries of M is uniquely determined by A , and the nonzero diagonal entries $a_{11}, \dots, a_{\ell\ell}$ are unique up to associates.

Furthermore, if R is a Euclidean domain, then P and Q can be chosen to be a product of elementary matrices.

Elementary row and column operations correspond to multiplication by elementary matrices, which are invertible, and that the composition of invertible matrices is invertible. So whenever we apply elementary row and column operations, we can translate it into multiplication by an invertible matrix on the left or the right, respectively.

To transform a matrix A into its Smith Normal Form, we will use a sequence of steps that all correspond to multiplication by invertible matrices. Many of those steps will actually be elementary row and column operations, which correspond to multiplication by an elementary matrix. Elementary matrices are invertible, and a product of invertible matrices is invertible, and so any finite sequence of elementary row and column operations can be described by multiplication by an invertible matrix. However, in general not every invertible matrix can

be obtained as a product of elementary matrices. In fact, there are examples of PIDs R and matrices A for which the Smith Normal Form cannot be obtained by simply taking a sequence of elementary row and column operations. However, it is not easy to give such an example, in part because when our PID R is nice enough, the Smith Normal Form can in fact be obtained by simply taking a sequence of elementary row and column operations. This is the case for Euclidean domains: over such rings, the Euclidean Algorithm for finding the gcd of two elements works, and it's the key step we will need to find a Smith Normal Form. When R is a general PID, however, we need to work a little harder.

Before we prove Theorem 13.10, let's see how to classify modules over PIDs using the Smith Normal Form for their presentation matrix. First, we need a lemma on how to interpret the module presented by a matrix in Smith Normal Form; we leave the proof as an exercise.

Lemma 13.11. *Let R be a commutative ring with $1 \neq 0$, let $m \geq n$, let $A = [a_{ij}] \in \text{Mat}_{m,n}(R)$ be a matrix such that all nondiagonal entries of A are 0, and let M be the R -module presented by A . Then $M \cong R^{m-n} \oplus R/(a_{11}) \oplus \cdots \oplus R/(a_{nn})$.*

Theorem 13.12 (Classification of finitely generated modules over a PID using invariant factors). *Let R be a PID and let M be a finitely generated module. Then there exist $r \geq 0$, $k \geq 0$, and nonzero nonunit elements d_1, \dots, d_k of R satisfying $d_1 | d_2 | \cdots | d_k$ such that*

$$M \cong R^r \oplus R/(d_1) \oplus \cdots \oplus R/(d_k).$$

Moreover r and k are uniquely determined by M , and the d_i are unique up to associates.

Proof. By Theorem 13.9, M has a presentation matrix A . By Theorem 13.10, A can be put into Smith Normal Form B , where the diagonal entries of B are b_1, \dots, b_ℓ and satisfy $b_1 | b_2 | \cdots | b_\ell$. Moreover, k is unique and the d_i are uniquely determined up to associates (ie, up to multiplication by units) by A , hence by B . By Theorem 13.9, M is isomorphic to the module presented by B . By Lemma 13.11, this is isomorphic to

$$M \cong R^r \oplus R/(b_1) \oplus \cdots \oplus R/(b_\ell).$$

Finally, some of these b_i might be units; let $d_1 | \cdots | d_k$ be the nonunits among the b_i , and note that if u is a unit, then $R/(u) \cong (0)$. We conclude that

$$M \cong R^r \oplus R/(d_1) \oplus \cdots \oplus R/(d_k),$$

as desired. \square

Definition 13.13. Let R be a PID, let $r \geq 0, k \geq 0$, and let d_1, \dots, d_k be nonzero nonunit elements of R satisfying $d_1 | d_2 | \cdots | d_k$. Let M be any R -module such that

$$M \cong R^r \oplus R/(d_1) \oplus \cdots \oplus R/(d_k).$$

We say M has **free rank** r and **invariant factors** d_1, \dots, d_k .

Notice that the invariant factors of M are only defined up to multiplication by units.

Remark 13.14. The classification theorem can be interpreted as saying that M decomposes into a free submodule R^r and a torsion submodule $\text{Tor}(M) = R/(d_1) \oplus \cdots \oplus R/(d_k)$.

Corollary 13.15 (Classification of finitely generated abelian groups). *Let G be a finitely generated abelian group. Then*

$$G \cong \mathbb{Z}^r \oplus \mathbb{Z}/n_1 \oplus \cdots \oplus \mathbb{Z}/n_k$$

for some $r \geq 0$, $k \geq 0$, and $n_i \geq 2$ for all i , satisfying $n_{i+1} \mid n_i$ for all i . Moreover, the integers r , k , and n_1, \dots, n_k are uniquely determined by G .

Example 13.16. Consider the \mathbb{Z} -module M presented by the matrix

$$A = \begin{bmatrix} 1 & 6 & 5 & 2 \\ 2 & 1 & -1 & 0 \\ 3 & 0 & 3 & 0 \end{bmatrix}.$$

We can obtain the Smith Normal Form as follows:

$$\begin{aligned} A &= \begin{bmatrix} 1 & 6 & 5 & 2 \\ 2 & 1 & -1 & 0 \\ 3 & 0 & 3 & 0 \end{bmatrix} \xrightarrow{\substack{R2 \rightarrow R2 - 2R1 \\ R3 \rightarrow R3 - 3R1}} \begin{bmatrix} 1 & 6 & 5 & 2 \\ 0 & -11 & -11 & -4 \\ 0 & -18 & -12 & -6 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -11 & -11 & -4 \\ 0 & -18 & -12 & -6 \end{bmatrix} \\ &\xrightarrow{C2 \leftrightarrow C4} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -4 & -11 & -11 \\ 0 & -6 & -12 & -18 \end{bmatrix} \xrightarrow{\substack{C3 \rightarrow C3 + 2C2 \\ C4 \rightarrow C4 + 3C1}} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -4 & -3 & 1 \\ 0 & -6 & 0 & 0 \end{bmatrix} \xrightarrow{C2 \leftrightarrow C4} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & -3 & -4 \\ 0 & 0 & 0 & -6 \end{bmatrix} \\ &\rightarrow \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -6 \end{bmatrix} \xrightarrow{C3 \leftrightarrow C4} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -6 & 0 \end{bmatrix} \xrightarrow{C3 \rightarrow -C3} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 6 & 0 \end{bmatrix}. \end{aligned}$$

Thus the Smith normal form of A is

$$M = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 6 & 0 \end{bmatrix},$$

with invariant factor $d_1 = 6$. Notice that the two ones are not invariant factors: we only care about nonunits. Therefore we have

$$M \cong \mathbb{Z}/(1) \oplus \mathbb{Z}/(1) \oplus \mathbb{Z}/(6) \cong \mathbb{Z}/(6).$$

Here is a spinoff of the classification theorem.

Theorem 13.17 (Classification of finitely generated modules over a PID using elementary divisors). *Let R be a PID and let M be a finitely generated module. Then there exist $r \geq 0$, $s \geq 0$, prime elements p_1, \dots, p_s of R (not necessarily distinct), and $e_1, \dots, e_s \geq 1$ such that*

$$M \cong R^r \oplus R/(p_1^{e_1}) \oplus \cdots \oplus R/(p_s^{e_s}).$$

Moreover, r and s are uniquely determined by M , and the list $p_1^{e_1}, \dots, p_s^{e_s}$ is unique up to associates and reordering.

Proof. First, write M in invariant factor form $M \cong R^r \oplus R/(d_1) \oplus \cdots \oplus R/(d_k)$. Then write each invariant factor as a product of prime powers

$$d_i := \prod_{j=n_i}^{n_{i+1}} p_j^{e_j},$$

and recall that by the CRT we have

$$R/(d_i) \cong R/(p_{n_i}^{e_{n_i}}) \oplus \cdots \oplus R/(p_{n_{i+1}}^{e_{n_{i+1}}}).$$

Substituting into the invariant factor form gives the desired result. Uniqueness follows from the uniqueness of the invariant factor form and of the prime factorizations of each d_i . \square

Definition 13.18. Let R be a PID, let $r \geq 0$, $s \geq 0$, p_1, \dots, p_s be prime elements of R , and let $e_1, \dots, e_s \geq 1$. Let M be the R -module $M \cong R^r \oplus R/(p_1^{e_1}) \oplus \cdots \oplus R/(p_s^{e_s})$. The elements $p_1^{e_1}, \dots, p_s^{e_s}$ of R are the **elementary divisors** of M .

Careful that a particular prime might appear repeatedly in the elementary divisors of a particular module.

Example 13.19. When $R = \mathbb{Z}$ and $M = \mathbb{Z}/(6)$, we can write $M \cong \mathbb{Z}/(2) \oplus \mathbb{Z}/(3)$, so the elementary divisors are 2 and 3.

Corollary 13.20. Let G be a finitely generated abelian group. Then there exist $r, s \geq 0$, prime integers p_1, \dots, p_s , and positive integers $e_i \geq 1$ such that

$$G \cong \mathbb{Z}^r \oplus \mathbb{Z}/p_1^{e_1} \oplus \cdots \oplus \mathbb{Z}/p_s^{e_s}.$$

Moreover, r , p_i , and e_i are all uniquely determined by G .

13.4 Proof of the Smith Normal Form Theorem

We have yet to show that every matrix over a PID has a Smith Normal Form.

Definition 13.21. Let R be a PID and $A \in \text{Mat}_{m,n}(R)$. For $1 \leq t \leq \min\{m, n\}$ we define $I_t(A)$ to be the ideal generated the $t \times t$ minors of A , i.e., the determinants of $t \times t$ submatrices of A . In particular, $I_1(A)$ is the ideal generated by all entries of A . We write $\gcd(A)$ for a generator of $I_1(A)$.

Note that $\gcd(A)$ is only well-defined up to associates: in a domain, two elements generate the same ideal if and only if they are unit multiples of each other. We will say $\gcd(A) = \gcd(B)$ to mean that there exist equal gcds for A and B .

We will use the following fact that you proved in the Homework.

Lemma 13.22. Let R be a commutative ring. Let $A \in \text{Mat}_{m,n}(R)$ be any matrix and let $P \in \text{Mat}_m(R)$ and $Q \in \text{Mat}_n(R)$ be invertible matrices. Then for any $t \geq 0$, we have $I_t(A) = I_t(PAQ)$, where I_t denotes the ideal generated by all $t \times t$ minors of a matrix.

In particular, $\gcd(A) = \gcd(PAQ)$.

Lemma 13.23. Let R be a PID and $x, y \in R$. Let g be a GCD of x and y .

(1) There exists an invertible 2×2 matrix $P \in \text{Mat}_2(R)$ such that

$$P \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} g \\ 0 \end{bmatrix}.$$

(2) If R is a Euclidean domain, then P can be chosen to be a product of elementary matrices.

Proof. We first prove part (1). By definition of greatest common divisor, $(x, y) = (\gcd(x, y))$, so there exist $a, b \in R$ such that $ax + by = \gcd(x, y)$. Write $g := \gcd(x, y)$ and $h = \gcd(a, b)$. Then $ax + by$ is a multiple of gh , but since $ax + by = g$ and R is a domain, we conclude that h must be a unit, and $(a, b) = (h) = (1)$. In particular, we can find $c, d \in R$ such that $ad - bc = 1$. Finally, $bx + cy \in (x, y) = (g)$, so

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} g \\ eg \end{bmatrix}.$$

Now we can apply the row operation that adds $-e$ times the first row to the second row: by setting

$$P := \begin{bmatrix} 1 & 0 \\ -e & 1 \end{bmatrix} \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} a & b \\ c - ea & b - de \end{bmatrix}.$$

we get

$$P \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} g \\ 0 \end{bmatrix}.$$

Finally, one can easily check that

$$P^{-1} = \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} \begin{bmatrix} 1 & 0 \\ e & 1 \end{bmatrix}.$$

We now prove part (2). Let N be a Euclidean norm for R . We proceed by induction on $M = \min\{N(x), N(y)\}$. If $M = 0$ then either x or y has norm zero and hence is a unit. Swapping rows if necessary (which is an elementary operation), we can assume x is a unit, and hence the GCD of x and y . Then

$$\begin{bmatrix} 1 & 0 \\ -yx^{-1} & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x \\ 0 \end{bmatrix}.$$

For the inductive step, again swapping if necessary, we can assume that $N(x) \geq N(y)$. By the division algorithm, we can write $x = qy + r$ with either $r = 0$ or $N(r) < N(y)$. Note that the GCD g of x and y is equal to the GCD of y and r . We have

$$\begin{bmatrix} 1 & -q \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} r \\ y \end{bmatrix}.$$

If $r = 0$ we are done; otherwise we can apply the induction hypothesis to find some product of elementary matrices P such that

$$P \begin{bmatrix} r \\ y \end{bmatrix} = \begin{bmatrix} g \\ 0 \end{bmatrix}.$$

Then $\begin{bmatrix} 1 & -q \\ 0 & 1 \end{bmatrix} P$ is the product we seek. \square

By transposing the matrices in ??, we can show that there exists an invertible 2×2 matrix Q such that

$$\begin{bmatrix} x & y \end{bmatrix} Q = \begin{bmatrix} g & 0 \end{bmatrix}.$$

We are now finally ready to show that every matrix over a PID can be put into Smith Normal Form.

Theorem 13.24. *Let R be a PID and let $A \in \text{Mat}_{m,n}(R)$. There exist invertible matrices P and Q such that $D = PAQ = [a_{ij}]$ satisfies the following: all nondiagonal entries of M are 0, meaning $a_{ij} = 0$ if $i \neq j$, and the diagonal entries of M satisfy*

$$a_{11} \mid a_{22} \mid a_{33} \mid \dots.$$

Moreover, the number ℓ of nonzero entries of D is uniquely determined by A , and the nonzero diagonal entries $a_{11}, \dots, a_{\ell\ell}$ are unique up to multiplication by units. Furthermore, if R is a Euclidean domain, that P and Q can be chosen to be products of elementary matrices.

Proof. We show the existence of such matrices P, Q in steps.

Proof of existence: To construct P, Q , we note first that it suffices to show the following:

Key claim: Given a matrix A , there exist invertible (products of elementary matrices, if R is Euclidean) P', Q' matrices such that

$$P' A Q' = \begin{bmatrix} d_1 & 0 \\ 0 & B_1 \end{bmatrix}$$

with $d_1 = \gcd(A)$ and the 0's denote a row and column of zeroes.

Indeed, once we have shown this, note that d divides every entry of B , so $d \mid \gcd(B)$. Then we can apply the claim to get

$$P''(P' A Q') Q'' = \begin{bmatrix} d_1 & 0 & 0 \\ 0 & d_2 & 0 \\ 0 & 0 & B_2 \end{bmatrix}$$

where $d_2 = \gcd(B_1)$ and $d_2 \mid \gcd(B_2)$. Repeating this process ends in a matrix of the form we seek.

Proof of Key Claim, case 1: In the setting of the key claim, suppose that $d_1 = \gcd(A)$ occurs as some entry of A . After swapping rows and columns (which corresponds to multiplication by elementary matrices), we can assume that d_1 is the top left entry of A . Now, every element of the first row and the first column (and every entry!) is a multiple

of d_1 , so we can subtract a suitable multiple of row one from each row (which corresponds to multiplication by elementary matrices) to get zeroes in the other entries of the first row. After that, do the same with the columns. Then we are done.

Proof of Key Claim, case 2: Without assuming that $d_1 = \gcd(A)$ does not occur as any entry of A , Choose an entry a of A for which the number of irreducible factors of a/d_1 is minimal. We proceed by induction on the number of such factors; the base case where this number is zero is case 1 above. Using row and column operations, we can put a in the top left entry.

If a does not divide some entry b of the first row, use a column operation to put b in the 1, 2 position. Using the Lemma, we can multiply by a suitable matrix to replace a, b by $\gcd(a, b)$ and 0. Since a does not divide b and $d_1 \mid \gcd(a, b) \mid a$, the number of irreducible factors of $\gcd(a, b)/d_1$ must be strictly less than that of a/b . We can then apply the induction hypothesis.

A similar argument applies if a does not divide some entry of the first column.

If a does divide every entry of the first row and first column, proceed as in case 1 to clear out the rest of the first row and column, and then add a row containing an entry b that does not divide a to the first row. Then proceed as in the first part of this case.

Proof of uniqueness: For a matrix of the form D , the only minors that are nonzero are those where the choices of columns and rows are the same, and hence the only nonzero $t \times t$ minors of M are $d_{s_1} \cdots d_{s_t}$ for some $s_1 < \cdots < s_t$. Since $d_{s_1} \cdots d_{s_t}$ divide each other, it follows that $I_t(A) = I_t(D) = d_1 \cdots d_t$. This proves uniqueness, for it shows that d_1, \dots, d_ℓ are all defined from A directly, without any choices. \square

Example 13.25. Consider the PID $R = k[x]$, where k is any field, and the matrix

$$A = \begin{bmatrix} x-1 & 0 \\ 1 & x-2 \end{bmatrix}.$$

The first row has already been zeroed out, but unfortunately $x-1$ does not divide 1. In this case, though, we can see that $\gcd(A) = 1$, so we can switch the first and second rows to get

$$\begin{bmatrix} 1 & x-2 \\ x-1 & 0 \end{bmatrix}.$$

Now we zero out the rest of the first row and first column using row and column operations:

$$\begin{bmatrix} 1 & x-2 \\ x-1 & 0 \end{bmatrix} \xrightarrow{R2 \rightarrow R2 - (x-1)R1} \begin{bmatrix} 1 & x-2 \\ 0 & -(x-1)(x-2) \end{bmatrix} \xrightarrow{C2 \rightarrow C2 - (x-2)C1} \begin{bmatrix} 1 & 0 \\ 0 & -(x-1)(x-2) \end{bmatrix}.$$

This is a Smith Normal Form. If we prefer to not have that negative sign, we can multiply the second row by -1 , to obtain

$$\begin{bmatrix} 1 & x-2 \\ x-1 & 0 \end{bmatrix} \xrightarrow{R2 \rightarrow R2 - (x-1)R1} \begin{bmatrix} 1 & x-2 \\ 0 & -(x-1)(x-2) \end{bmatrix} \xrightarrow{C2 \rightarrow C2 - (x-2)C1} \begin{bmatrix} 1 & 0 \\ 0 & (x-1)(x-2) \end{bmatrix}.$$

There is only one invariant factor, which is $(x-1)(x-2)$. The $k[x]$ -module M presented by A is

$$M \cong k[x]/((x-1)(x-2)).$$

If we prefer to write this in terms of elementary divisors, our module has two: $x - 1$ and $x - 2$, and it is isomorphic to

$$M \cong k[x]/(x - 1) \oplus k[x]/(x - 2).$$

Chapter 14

Canonical forms for endomorphisms

We will now apply the structure theory for modules over PIDs to study matrices over fields.

14.1 The module associated to an F -linear endomorphism

For any linear transformation from an F -vector space to itself, we can construct a module over a polynomial ring $F[x]$.

Definition 14.1. Let F be a field, and V be an F -vector space. Let $\phi : V \rightarrow V$ be a linear transformation. For any polynomial

$$f(x) = a_nx^n + \cdots + a_1x + a_0,$$

we have an F -linear transformation $f(\phi) : V \rightarrow V$ given by

$$f(\phi) = a_n\phi^n + \cdots + a_1\phi + a_0\text{id}_V.$$

We define a V_ϕ to be the $F[x]$ -module with underlying additive group $(V, +)$ and $F[x]$ -action given by

$$f(x) \cdot v = f(\phi)(v).$$

That is, V_ϕ is the same F -vector space as V , with a bigger action where the action of x on V is given by ϕ . Of course, one has to verify that this definition satisfies the module axioms. We leave this for you as an exercise if you want more practice with the module axioms.

Every $F[x]$ -module can be thought of this way.

Lemma 14.2 ($F[x]$ -modules). *Let F be a field. There is a bijection*

$$\begin{aligned} \{V \mid V \text{ a } F[x]\text{-module}\} &\longleftrightarrow \{(V, \phi) \mid V \text{ a } F\text{-vector space, } \phi \in \text{End}_F(V)\} \\ V &\rightarrow (V, \mu_x : V \rightarrow V) \\ W_\phi &\leftarrow (W, \phi), \end{aligned}$$

where $\mu_x : V \rightarrow V$ is the map of multiplication by x on the $F[x]$ -module V .

Proof. If V is an $F[x]$ -module then V is an F -vector space by restriction of scalars along the inclusion $F \hookrightarrow F[x]$. Let $\mu_x : V \rightarrow V$ be as above. To show that $\mu_x \in \text{End}_F(V)$, note that for any $c \in F$ and $v, v_1, v_2 \in V$ the axioms of the $F[x]$ -module give us

$$\mu_x(v_1 + v_2) = x(v_1 + v_2) = xv_1 + xv_2 = \mu_x(v_1) + \mu_x(v_2) \text{ and } \mu_x(cv) = x(cv) = c\mu_x(v).$$

The construction $(W, \phi) \rightarrow W_\phi$ was discussed above. It remains to check that these are mutually inverse constructions, which we leave for you. \square

Example 14.3. Let $F = \mathbb{R}$, and $V = \mathbb{R}^2$.

- (1) For $\phi = 0$, in the $\mathbb{R}[x]$ -module V_ϕ , we have $x \cdot \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ and more generally
- $$(c_n x^n + \dots + c_0) \cdot \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} = \begin{bmatrix} c_0 \lambda_1 \\ c_0 \lambda_2 \end{bmatrix}.$$

The $\mathbb{R}[x]$ -module V_ϕ is evidently generated by e_1 and e_2 , since any element of V_ϕ can be written $\lambda_1 e_1 + \lambda_2 e_2$. (We didn't even need to use the x 's!). Since x times any vector $v \in V$ is zero, $x \in \text{ann}_{\mathbb{R}[x]}(V_\phi)$. You can check that $\text{ann}_{\mathbb{R}[x]}(V_\phi) = (x)$.

- (2) Let ϕ be the linear transformation of rotation by $\pi/3$ counterclockwise. The standard basis vectors e_1, e_2 generate V_ϕ , since any element is an \mathbb{R} -linear combination of e_1, e_2 . Even better, V_ϕ is generated by e_1 as an $\mathbb{R}[x]$ -module, since e_1 and $xe_1 = \begin{bmatrix} \sqrt{3}/2 \\ 1/2 \end{bmatrix}$ are \mathbb{R} -linearly independent, so any $v \in V_\phi$ can be written as $\lambda_1 e_1 + \lambda_2 xe_1 = (\lambda_2 x + \lambda_1) e_1$. Since ϕ^6 is the identity on V , we have $(x^6 - 1)v = \phi^6(v) - v = 0$ for all $v \in V$. It follows that $x^6 - 1$ annihilates V_ϕ . We will see soon that $\text{ann}_{\mathbb{R}[x]}(V_\phi) = (x^2 - x + 1)$.

We can give a presentation for V_ϕ in general by choosing a basis for V . Recall that given an F -vector space V with $\dim_F(V) = n$ and an ordered basis B for V we have proven that $\text{End}_F(V) \cong \text{Mat}_n(F)$ via the maps $\phi \mapsto [\phi]_B^B$ and $A \mapsto \phi_A$.

Theorem 14.4. *Let F be a field, let V be an F -vector space of dimension n , let $\phi: V \rightarrow V$ be a linear transformation, let B be an ordered basis for V , and let $A = [\phi]_B^B$. Then the matrix $xI_n - A \in \text{Mat}_n(F[x])$ presents the $F[x]$ -module V_ϕ .*

Proof. Let $B = \{b_1, \dots, b_n\}$ be any basis for V , and note that B is a generating set for V_ϕ as a module over $F[x]$. Then V_ϕ can then be written as a quotient of $F[x]^n$. More precisely, let e_1, \dots, e_n denote the standard $F[x]$ -basis for the free $F[x]$ -module $F[x]^n$, and let $\pi: F[x]^n \rightarrow V_\phi$ be the surjective $F[x]$ -module homomorphism sending e_i to b_i . That is,

$$\pi((g_1(x), \dots, g_n(x))) = \pi \left(\sum_{i=1}^n g_i(x) e_i \right) = \sum_{i=1}^n g_i(x) b_i = \sum_{i=1}^n g_i(\phi) b_i.$$

By the First Isomorphism Theorem, we have $V_\phi \cong F[x]^n / \ker(\pi)$. On the other hand, the matrix $xI_n - A$ determines a map

$$\phi_{xI_n - A}: F[x]^n \rightarrow F[x]^n,$$

and to show that $V_\phi \cong F[x]^n/\text{im}(\phi_{xI_n - A})$ it suffices to show that $\text{im}(\phi_{xI_n - A}) = \ker(\pi)$. Now $(\pi \circ \phi_{xI_n - A})(e_i) = \pi((xI_n - A)e_i) = (xI_n - A)\pi(e_i) = (xI_n - A)b_i = xb_i - Ab_i = \phi(b_i) - \phi(b_i) = 0$.

This proves $\text{im}(xI_n - A) \subseteq \ker(\pi)$. It follows by UMP for quotient modules that there is a surjection of $F[x]$ -modules

$$W := F[x]^n/\text{im}(xI_n - A) \twoheadrightarrow V_\phi.$$

We may also regard this as a surjection of F -vector spaces. Since $\dim_F(V_\phi) = n$ and the map above is surjective, we have $\dim_F(W) \geq n$, which follows from the Rank Nullity Theorem. To establish that the map above is an isomorphism, it suffices to show that $\dim_F(W) \leq n$.

Denote by $c_i = e_i + \text{im}(xI_n - A)$ the image of the standard basis of $F[x]^n$ in W . The i th column of $xI_n - A$ gives the relation $xc_i = v_i$ in W , where v_i is the i -th column of A . It follows that $p(x)c_i = p(A)c_i$ in W for any polynomial $p(x)$. Thus a typical element of W , given by $\sum_i g_i(x)c_i$, is equal to $g_1(A)c_1 + \cdots + g_n(A)c_n$. Such an expression belongs to the F -span of c_1, \dots, c_n in W ; that is; c_1, \dots, c_n span W as an F -vector space. Therefore, we have the desired inequality $\dim_F(W) \leq n$, which completes our proof. \square

Corollary 14.5. Suppose F is a field, V is an F -vector space, and $\phi: V \rightarrow V$ is a linear transformation. There exist unique monic polynomials $g_1 | \cdots | g_k \in F[x]$ of positive degree and an $F[x]$ -module isomorphism

$$V_t \cong F[x]/(g_1) \oplus \cdots \oplus F[x]/(g_k).$$

The polynomials g_1, \dots, g_k are both the invariant factors of the $F[x]$ -module V_ϕ and the entries on the diagonal of the Smith normal form of $xI_n - [\phi]_B^B$ for any basis B of V .

Proof. The statement says that $xI_n - [t]_B^B$ presents the $F[x]$ -module V_ϕ , and the remainder of the statement is an immediate application of the Classification of finitely generated modules over PIDs to this special case once we show that there is no free summand. Note that $F[x]$ is an infinite dimensional vector space over F , while V_ϕ is a finite dimensional vector space. If V_ϕ had a free summand, then it would contain an infinite linearly independent set over F , and thus it could not be finite-dimensional. \square

Definition 14.6. The polynomials g_1, \dots, g_k in ?? are called the **invariant factors** of the linear transformation ϕ .

Example 14.7. Let

$$A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \in \text{Mat}_2(\mathbb{Q}).$$

Then

$$xI_2 - A = \begin{bmatrix} x-1 & -1 \\ 0 & x-1 \end{bmatrix}.$$

We could compute the invariant factors of $t: \mathbb{Q}^2 \rightarrow \mathbb{Q}^2$ by appealing to the [Smith Normal Form](#) of $xI_2 - A$, but let us try another way. Let

$$\begin{bmatrix} d_1 & 0 \\ 0 & d_2 \end{bmatrix}$$

be the Smith Normal Form of $xI_2 - A$. Recall from the proof of Theorem 13.10 that d_1 is the gcd of the entries of $xI_2 - A$ and $d_1d_2 = \det(xI_2 - A)$. Thus $d_2 = \det(xI_2 - A) = (x - 1)^2$ and $d_1 = 1$. Therefore the only invariant factor of t_A is $(x - 1)^2$.

We know that for linear transformation $\alpha \neq \beta$ from $V \rightarrow V$, the modules V_α and V_β are not equal. It is natural to ask when they are isomorphic. You will show the following in Problem Set #7:

Exercise 99. Let F be a field and $V = F^n$. Let $A, B \in \text{Mat}_{n \times n}(F)$. Then V_{t_A} and V_{t_B} are isomorphic if and only if A and B are similar matrices, i.e., there is some invertible matrix P such that $B = PAP^{-1}$.

14.2 Rational canonical form

You will show the following lemma in Problem Set #6:

Lemma 14.8. For a monic polynomial $f(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0$ with $n \geq 1$, the classes of $1, x, \dots, x^{n-1}$ form a basis for $F[x]/(f(x))$ regarded as an F -vector space. Relative to this basis, the F -linear operator $\mu_x : F[x]/(f(x)) \rightarrow F[x]/(f(x))$ defined by $\mu_x(v) = xv$ is given by the following matrix:

$$C(f) := \begin{bmatrix} 0 & 0 & \cdots & 0 & -a_0 \\ 1 & 0 & \cdots & 0 & -a_1 \\ 0 & 1 & \ddots & 0 & -a_2 \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & 1 & -a_{n-1} \end{bmatrix} = \begin{bmatrix} 0 & \cdots & 0 & -a_0 \\ & \ddots & & -a_1 \\ & & I_{n-1} & \vdots \\ & & & -a_{n-1} \end{bmatrix}.$$

Definition 14.9. In the setup of ??, the matrix $C(f)$ is called the **companion matrix** of the monic polynomial f .

Definition 14.10. Given square matrices A_1, \dots, A_m with entries in a ring R , not necessarily of the same size, we define $A_1 \oplus \cdots \oplus A_m$ to be the block diagonal matrix

$$\begin{bmatrix} A_1 & 0 & \cdots & 0 \\ 0 & A_2 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & 0 & A_m \end{bmatrix}.$$

Remark 14.11. If $f : V_1 \rightarrow W_1$ and $g : V_2 \rightarrow W_2$ are linear transformations, then the map $f \oplus g : V_1 \oplus V_2 \rightarrow W_1 \oplus W_2$ given by $(f \oplus g)(a, c) = (f(a), g(c))$ is a linear transformation. If B_i is a basis for V_i and C_i is a basis for W_i , and $\iota_i : A_i \hookrightarrow A_1 \oplus A_2$ are the natural inclusions, then $\mathcal{B} = \iota_1(B_1) \cup \iota_2(B_2)$ is a basis for $V_1 \oplus V_2$, $\mathcal{C} = \iota_1(C_1) \cup \iota_2(C_2)$ is a basis for $W_1 \oplus W_2$, and

$$[f \oplus g]_{\mathcal{B}}^{\mathcal{C}} = \begin{bmatrix} [f]_{B_1}^{C_1} & 0 \\ 0 & [g]_{B_2}^{C_2} \end{bmatrix}.$$

Theorem 14.12 (Rational Canonical Form). *Let F be a field, V a finite dimensional F -vector space, and $\phi: V \rightarrow V$ an F -linear transformation. There is a basis B of V such that*

$$[\phi]_B^B = C(g_1) \oplus \cdots \oplus C(g_k) = \begin{bmatrix} C(g_1) & 0 & 0 & \cdots & 0 \\ 0 & C(g_2) & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & & \vdots \\ 0 & 0 & \cdots & 0 & C(g_k) \end{bmatrix}$$

where g_1, \dots, g_k are the invariant factors of t , meaning in particular they are monic polynomials of positive degree such that $g_1 | g_2 | \cdots | g_k$. Moreover, the polynomials g_1, \dots, g_k are unique.

Proof. By ??, $V_\phi \cong \bigoplus_{i=1}^k F[x]/(g_i(x))$ for some unique g_i as in the statement. Set $V_i = F[x]/(g_i(x))$ and note that $V_\phi = V_1 \oplus \cdots \oplus V_k$. The map $\mu_x: V_\phi \rightarrow V_\phi$ given by multiplication by x preserves each summand in this decomposition: $\mu_x(V_i) \subseteq V_i$. Thus if we choose a basis B_i of each summand V_i and set $B = \bigcup_{i=1}^k \iota_i(B_i)$, by ??, B is a basis of V_ϕ and $[\phi]_B^B = [\phi|_{V_1}]_{B_1}^{B_1} \oplus \cdots \oplus [\phi|_{V_k}]_{B_k}^{B_k}$. The result now follows from ?? \square

Definition 14.13. In the setup of ??, the matrix $C(g_1) \oplus \cdots \oplus C(g_k)$ is called the **rational canonical form** (RCF) of the linear transformation ϕ . The rational canonical form of a matrix $A \in \text{Mat}_n(F)$ is defined to be the rational canonical form of the endomorphism t_A represented by A with respect to the standard basis of F^n .

Example 14.14. Let $A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \in \text{Mat}_2(\mathbb{Q})$ as in ???. Because the only invariant factor of $xI_2 - A$ is $(x - 1)^2$, the Rational Canonical Form of t_A is

$$\text{RCF}(A) = C((x - 1)^2) = C(x^2 - 2x + 1) = \begin{bmatrix} 0 & -1 \\ 1 & 2 \end{bmatrix}.$$

The rational canonical form is canonical in that it uniquely characterizes the similarity class of a matrix.

Theorem 14.15. *Let F be a field and let $A, B \in \text{Mat}_n(F)$. The following are equivalent:*

- (1) *A and B are similar matrices.*
- (2) *A and B have the same Rational Canonical Form.*
- (3) *A and B have the same invariant factors.*

Proof. To show (1) \Rightarrow (2), suppose A is similar to B . Then there exists an invertible matrix P such that $B = PAP^{-1}$, and thus

$$xI_n - B = xI_n PAP^{-1} = P(xI_n - A)P^{-1}.$$

Thus the matrices $xI_n - A$ and $xI_n - B$ are also similar. Moreover, similar matrices have the same Smith Normal Form, since they present isomorphic modules and thus A and B

have the Rational Canonical Form. The invariant factors can be read off of the Rational Canonical Form, and thus $(2) \Rightarrow (3)$.

Finally, to show $(3) \Rightarrow (1)$ notice that if A and B have the same invariant factors then by the Structure Theorem of Finitely Generated Modules over PIDs, there is an isomorphism of $F[x]$ -modules $F_{t_A}^n \cong F_{t_B}^n$, which implies by a homework problem in Problem Set #7 that A and B must be similar. \square

14.3 The Cayley-Hamilton Theorem

Definition 14.16. Let F be a field and let $A \in \text{Mat}_n(F)$. The **characteristic polynomial** of A is the polynomial $c_A = \det(xI_n - A)$.

Definition 14.17. Let V be an F -vector space of dimension n , and let $t : V \rightarrow V$ be a linear transformation. The **characteristic polynomial** of t , denoted c_t , is the characteristic polynomial c_A for a matrix $A = [t]_B^B$ with respect to some ordered basis B of V .

Characteristic polynomials are well-defined.

Remark 14.18. We need to check that the characteristic polynomial of a linear transformation is invariant under base changes. More precisely, we need to check that if we choose two different basis B and B' for V , then the matrices $A = [t]_B^B$ and $C = [t]_{B'}^{B'}$ have the same characteristic polynomial. First, recall that A and C are similar matrices, by Theorem 12.40, so $C = PAP^{-1}$ for some invertible matrix P . Moreover, diagonal matrices are in the center of $\text{Mat}_n(R)$, meaning they commute with other matrices, and thus we have the following:

$$\begin{aligned}\det(xI_n - C) &= \det(xI_n - PAP^{-1}) \\ &= \det(P(xI_n - A)P^{-1}) \\ &= \det(P)\det(xI_n - A)\det(P^{-1}) \\ &= \det(xI_n - A).\end{aligned}$$

We conclude that A and B have the same characteristic polynomial.

Remark 14.19. For any matrices A and B , $c_{A \oplus B} = c_A c_B$.

Definition 14.20. Let F be a field and let $A \in \text{Mat}_n(F)$. The **minimal polynomial** of A , denoted m_A , is the unique monic polynomial that generates the principal ideal

$$\{f(x) \in F[x] \mid f(A) = 0\}.$$

Definition 14.21. Let V be an F -vector space of dimension n , and let $t : V \rightarrow V$ be a linear transformation. The **minimal polynomial** of t , denoted m_t , is the unique monic polynomial generating the ideal $\text{ann}_{F[x]}(V_t)$ in the PID $F[x]$.

Lemma 14.22. Let F be a field. Let V be an F -vector space of dimension n with basis B and let $t : V \rightarrow V$ be a linear transformation. The minimal polynomial m_A of $A = [t]_B^B$ satisfies $m_A = m_t$.

Proof. Since m_A and m_t are both monic, it's sufficient to show $\text{ann}_{F[x]}(V_t) = (m_A)$. Indeed,

$$\begin{aligned}
f \in \text{ann}_{F[x]}(V_t) &\iff f(x)v = 0 \text{ for all } v \in V_t \\
&\iff f(A)v = 0 \text{ for all } v \in V_t \\
&\iff \ker(f(A)) = V_t \\
&\iff \text{rank}(f(A)) = 0 && \text{by the Rank-Nulity Theorem} \\
&\iff f(A) = 0 \\
&\iff f \in (m_A) && \text{by definition of } m_A. \quad \square
\end{aligned}$$

Remark 14.23. If $m(x)$ is the minimal polynomial of an endomorphism t and $f(x)$ is another polynomial such that $f(x)$ annihilates V_t , then $f(x) \in \text{ann}(V_t) = (m(x))$, and thus $m(x) | f(x)$.

Similarly, suppose that $m(x)$ is the minimal polynomial of a matrix A and $f(x)$ is another polynomial such that $f(A) = 0$. By ??, we know that $m(x)$ is also the minimal polynomial of the linear transformation $t : v \mapsto Av$, and that $f(x)$ also annihilates V_t . Thus we can also conclude that $m(x) | f(x)$.

Lemma 14.24. Let F be a field, let V be a finite dimensional F -vector space, and $t : V \rightarrow V$ be a linear transformation with invariant factors $g_1 | \cdots | g_k$. Then $c_t = g_1 \cdots g_k$ and $m_t = g_k$.

Proof. The product of the elements on the diagonal of the Smith Normal Form of $xI_n - A$ is the determinant of $xI_n - A$. Thus the product of the invariant factors $g_1 \cdots g_k$ of V_t is the characteristic polynomial c_t of t . Notice here that we chose our invariant factors g_1, \dots, g_k to be monic, so that $g_1 \cdots g_k$ is monic, and thus actually equal to c_t (not just up to multiplication by a unit).

By Problem Set 5, $\text{ann}_{F[x]}(V_t) = (g_k)$, and since g_k is monic we deduce that $m_t = g_k$. \square

We can now prove the famous Cayley-Hamilton theorem.

Theorem 14.25 (Cayley-Hamilton). Let F be a field, and let V be a finite dimensional F -vector space. If $t : V \rightarrow V$ is a linear transformation, then $m_t | c_t$, and hence $c_t(t) = 0$. Similarly, for any matrix $A \in \text{Mat}_n(F)$ over a field F we have $m_A | c_A$ and $c_A(A) = 0$.

Proof. Let $A = [t]_B^B$ for some basis B of V . Note that the statements about A and t are equivalent, since by definition $c_A = c_t$, while $m_A = m_t$ we have $f(A) = 0$ if and only if $f(t) = 0$. So write $m = m_A = m_t$ and $c = c_A = c_t$.

By ??, $m = g_k$ and $c = g_1 \cdots g_k$, so $m | c$. By definition, we $m(A) = 0$. Since $m | c$, we conclude that $c(A) = 0$. \square

Remark 14.26. As a corollary of the Cayley-Hamilton Theorem, we obtain that the minimal polynomial of $t : V \rightarrow V$ has degree at most $n = \dim(V)$, since m_t divides c_t , which is a polynomial of degree n .

Lemma 14.27. Let F be a field and let V be a finite dimensional F -vector space. If $t : V \rightarrow V$ is a linear transformation, then $c_t | m_t^k$.

Proof. Since $g_i | g_k$ for $1 \leq i \leq k$, we have $c_t = g_1 \cdots g_k | g_k^k = m_t^k$. \square

It follows that c_t and m_t have the same roots, not counting multiplicities.

Definition 14.28. Let V be $t: V \rightarrow V$ be a linear transformation over a field F . A nonzero element $v \in V$ satisfying $t(v) = \lambda v$ for some $\lambda \in F$ is an **eigenvector** of t with **eigenvalue** λ . Similarly, given a matrix $A \in \text{Mat}_n(F)$, a nonzero $v \in F^n$ satisfying $Av = \lambda v$ for some $\lambda \in F$ is an **eigenvector** of A with **eigenvalue** λ .

Theorem 14.29. Let $f \in F$. The following are equivalent:

- (1) λ is an eigenvalue of t .
- (2) λ is a root of c_t .
- (3) λ is a root of m_t .

Proof. By the Cayley-Hamilton Theorem, $m_t|c_t$, and thus (3) \Rightarrow (2). On the other hand, by ?? we know that $c_t | m_t^k$, so if $c_t(\lambda) = 0$ then $m_t(\lambda)^k = 0$, and since we are over a field, we conclude that $m_t(\lambda) = 0$. This shows (2) \Rightarrow (3).

Finally, to show that (1) \Leftrightarrow (2), notice that the scalar $\lambda \in F$ is an eigenvalue of A if and only if there is a nonzero solution v to $(\lambda I_n - A)v = 0$. This happens if and only if $\lambda I_n - A$ has a nontrivial kernel, or equivalently if $\lambda I - A$ is not invertible. Thus $\lambda \in F$ is an eigenvalue of A if and only if it is a root of its characteristic polynomial $c_A(x) = \det(xI_n - A)$, meaning $c_A(\lambda) = 0$. \square

Example 14.30. Let us find the minimal and characteristic polynomials of $T : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ given as rotation by 60 degrees counter-clockwise. We could write this down as matrix and compute its characteristic polynomial, but a simpler way is to notice that $T^3 = -I_2$, and so T satisfies the polynomial $x^3 + 1 = (x + 1)(x^2 - x + 1)$. Its minimal polynomial must therefore divide $x^3 + 1$. Since $x^3 + 1 = (x + 1)(x^2 - x + 1)$ and $x^2 - x + 1$ is irreducible in $\mathbb{R}[x]$, we conclude that the minimal polynomial of T , which we know has degree at most 2, must be either $x + 1$ or $x^2 - x + 1$. If $m_T = x + 1$, then T would be $-I_2$, which is clearly incorrect. So the minimal polynomial of T must be $x^2 - x + 1$. By Cayley-Hamilton, this polynomial must divide the characteristic polynomial, and since the latter also has degree two, we conclude that

$$c_T(x) = x^2 - x + 1.$$

Since this is irreducible, in this example we have no choice for how to form the invariant factors: there must just be one of them, $c_T(x)$ itself. So

$$C(x^2 - x + 1) = \begin{bmatrix} 0 & -1 \\ 1 & 1 \end{bmatrix}$$

is the rational canonical form of T .

Example 14.31. Let's find the minimal polynomial of

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

By the Cayley-Hamilton Theorem, $m_A(x) \mid c_A(x)$. The polynomial $c_A(x)$ is easy to compute since this matrix is upper-triangular:

$$c_A(x) = \det(xI_4 - A) = (x - 1)^4.$$

So $m_A(x) = (x - 1)^j$ for some $j \leq 4$. By brute-force, we verify that $(A - I_4)^3 \neq 0$ and thus it must be the case that $m_A(x) = c_A(x) = (x - 1)^4$.

Example 14.32. Let's find the minimal polynomial of

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

As in the previous example, $c_A(x) = (x - 1)^4$ and so by the Cayley-Hamilton Theorem $m_A(x) = (x - 1)^j$ for some $j \leq 4$. This time we notice that $(A - I_4)^2 = 0$ and so, since $(A - I_4) \neq 0$, we have $m_A(x) = c_A(x) = (x - 1)^2$.

14.4 Jordan canonical form

We now turn to the Jordan canonical form. To motivate it, let us do an example.

Example 14.33. Let us consider

$$A = \begin{bmatrix} 0 & 0 & 8 \\ 1 & 0 & -12 \\ 0 & 0 & 6 \end{bmatrix} = C((x - 2)^3) \in \text{Mat}_3(\mathbb{Q}).$$

This means we can interpret this matrix as arising from the linear transformation l_x on

$$V = \mathbb{Q}[x]/(x - 2)^3$$

given by multiplication by x . Recall that the basis that gives the matrix A is

$$B = \{\bar{1}, \bar{x}, \bar{x^2}\}$$

But notice that

$$B' = \{\overline{(x - 2)^2}, \overline{x - 2}, \bar{1}\}$$

is also a basis of V , and indeed seems like a more pleasing one. Let us calculate what the operator T does to this alternative basis. We could work this out by brute force, but a

cleaner way is to first compute what the operator $T' = T - 2\text{id}_V$ does. It is clear that T' is multiplication by $x - 2$, and hence T' sends each basis element to the previous one, except for the first which is sent to 0. That is the matrix of T' is

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

and hence the matrix for T is $T' + 2I_3$:

$$J_3(2) := \begin{bmatrix} 2 & 1 & 0 \\ 0 & 2 & 1 \\ 0 & 0 & 2 \end{bmatrix}.$$

This is a *Jordan Block*.

Definition 14.34. Let F be a field, let $n > 0$, and let $r \in F$. The **Jordan block** $J_n(r)$ is the $n \times n$ matrix over F with entries satisfying the following:

$$a_{ij} = \begin{cases} r & \text{if } i = j \\ 1 & \text{if } j = i + 1 \\ 0 & \text{otherwise.} \end{cases}$$

Thus a Jordan block looks like

$$\begin{bmatrix} r & 1 & & & \\ r & \ddots & & & \\ & \ddots & \ddots & & \\ & & r & 1 & \\ & & & r & \end{bmatrix}.$$

Theorem 14.35 (Jordan Canonical Form Theorem). *Let F be a field, let V be a finite dimensional vector space, and let $t : V \rightarrow V$ be a linear transformation satisfying the property that the characteristic polynomial c_t of t factors completely into linear factors over F . Then there is an ordered basis B for V such that*

$$[t]_B^B = J_{e_1}(r_1) \oplus \cdots \oplus J_{e_s}(r_s) = \begin{bmatrix} J_{e_1}(r_1) & 0 & 0 & \cdots & 0 \\ 0 & J_{e_2}(r_2) & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & & \vdots \\ 0 & 0 & \cdots & 0 & J_{e_s}(r_s) \end{bmatrix}$$

such that each $r_i \in F$ is a root of the characteristic polynomial c_t and each $e_i \geq 1$. Moreover, the polynomials $(x - r_1)^{e_1}, \dots, (x - r_s)^{e_s}$ are the elementary divisors of the $F[x]$ -module V_t , and this expression for $[t]_B^B$ is unique up to ordering of the Jordan blocks.

Proof. The key point is the following: the assumption that c completely factors into linear terms guarantees that the elementary divisors of c are of the form $(x - r)^e$. The proof then follows along the lines of ???. First write V_t in terms of the elementary divisors, as follows

$$V_t \cong F[x]/((x - r_1)^{e_1}) \oplus \cdots \oplus F[x]/((x - r_s)^{e_s}).$$

Then pick bases $B'_i = \{\overline{(x - r_i)^{e_i-1}}, \dots, \overline{x - r_i}, \overline{1}\}$ for each of the summands and set

$$B := \bigcup_{i=1}^s \iota_i(B'_i).$$

All that remains to show is that the matrix representing multiplication by x on each summand is $J_{e_i}(r_i)$. More precisely, we want to compute the matrix representing the F -linear transformation $T: F[x]/((x - r)^e) \xrightarrow{x} F[x]/((x - r)^e)$ in the basis $B = \{\overline{(x - r)^{e-1}}, \dots, \overline{x - r}, \overline{1}\}$. Let $T' := T - r \cdot \text{id}$, and note that

$$T'(\overline{(x - r)^{e-1}}) = 0$$

and

$$T'(\overline{(x - r)^i}) = \overline{(x - r)^{i+1}} \text{ for all } i < e - 1.$$

Thus the first column of $[T']_B^B$ is zero, and each of the remaining ordered basis vectors is taken to the previous basis vector, so that

$$[T']_B^B = \begin{bmatrix} 0 & 1 & & & \\ & 0 & \ddots & & \\ & & \ddots & \ddots & \\ & & & 0 & 1 \\ & & & & 0 \end{bmatrix}.$$

Since $T = T' + r\text{id}$, we conclude that $[T]_B^B$ is indeed given by the Jordan block $J_e(r)$, as desired. \square

Definition 14.36. Let F be a field, V be a finite dimensional vector space, and let $t: V \rightarrow V$ be a linear transformation satisfying the property that the characteristic polynomial c_t of t factors completely into linear factors and has elementary divisors $(x - r_1)^{e_1}, \dots, (x - r_s)^{e_s}$. The matrix $J_{e_1}(r_1) \oplus \dots \oplus J_{e_s}(r_s)$ is a **Jordan canonical form** (JCF) of t .

A **Jordan canonical form** for $A \in \text{Mat}_n(F)$ is a Jordan canonical form for the linear transformation $t_A: F^n \rightarrow F^n$ such that $A = [t]_E^E$ in the standard basis E of F^n .

The same matrix may fail to have a JCF when interpreted as a matrix with entries in a smaller field while it has a JCF when interpreted as a matrix with entries in a larger field.

Example 14.37. We revisit the example of the rotation by 60° but extend scalars to \mathbb{C} . That is, start with a matrix A with $c_A(x) = x^2 - x + 1 = (x - w)(x - \bar{w})$ where $w = \frac{1+\sqrt{3}i}{2}$. Since the minimal polynomial of A is $m_A = x^2 - x + 1$, we deduced in ?? that the only invariant factor of A is $x^2 - x + 1$, and hence the RCF of A is $C(x^2 - x + 1)$. On the other hand, over \mathbb{C} the polynomial m_A factors, say as $x^2 - x + 1 = (x - w)(x - \bar{w})$, and thus by the CRT

$$\mathbb{C}[x]/(x^2 - x + 1) \cong \mathbb{C}[x]/(x - w) \oplus \mathbb{C}[x]/(x - \bar{w}).$$

Therefore,

$$A \sim C(x - w) \oplus C(x - \bar{w}) = \begin{bmatrix} w & 0 \\ 0 & \bar{w} \end{bmatrix}.$$

The latter matrix is the JCF of A , and in this case the JCF is a diagonal matrix. Notice that if we consider $A \in \text{Mat}_2(\mathbb{R})$ then the characteristic polynomials fails to factor into linear factors. Hence $A \in \text{Mat}_2(\mathbb{R})$ does not have a JCF.

Definition 14.38. Let F be a field, let V be a finite dimensional vector space, and let $t : V \rightarrow V$ be a linear transformation. Then t is **diagonalizable** if there is a basis B for V such that the matrix $[t]_B^B$ is a diagonal matrix. Let $A \in \text{Mat}_n(F)$. Then A is **diagonalizable** if A is similar to a diagonal matrix.

Theorem 14.39. Let F be a field, let V be a finite dimensional vector space, and consider a linear transformation $t : V \rightarrow V$. The following are equivalent:

- (1) t is diagonalizable.
- (2) t has a Jordan canonical form A and A is a diagonal matrix.
- (3) t has a Jordan canonical form and all elementary divisors are of the form $x - r$ with $r \in F$.
- (4) Each invariant factor of t is a product of linear polynomials with no repeated linear factors.
- (5) The minimal polynomial of t is a product of linear polynomials with no repeated linear factors.

Proof. Note that a diagonal matrix is an example of a matrix in JCF. By the uniqueness of the JCF, (1) holds if and only if (2) holds. Moreover, the equivalence of (2) and (3) follows by definition. A matrix has a JCF if and only if its invariant factors factor completely. In this case, the elementary divisors are constructed by decomposing each invariant factor into powers of distinct linear polynomials. This gives that (3) holds if and only if (4) holds. Finally, since the minimal polynomial is one of the invariant factors and every other invariant factor divides it, we get the equivalence between (4) and (5). \square

Part IV

Fields and Galois Theory

Chapter 15

Field Extensions

One motivation for studying field extensions is that we want to build fields in which certain polynomials have roots. Here is a classical example going back to Gauss: while over \mathbb{R} the polynomial $f = x^2 + 1 \in \mathbb{R}[x]$ has no roots, if we want a field in which f does have a root we need to consider $\mathbb{C} = \mathbb{R}(i) = \{a + bi \mid a, b \in \mathbb{R}\}$.

Here's another example that has already come up in this class: the polynomial $g = x^2 - x + 1 \in \mathbb{Q}[x]$. We know that this has a root $\omega = \frac{1+\sqrt{3}i}{2} \in \mathbb{C}$. But if we look for the smallest field containing \mathbb{Q} in which $x^2 - x + 1$ has a root we obtain the field $\mathbb{Q}(\omega) = \{a + b\omega \mid a, b \in \mathbb{Q}\}$.

So here's our goal: starting from a smaller field F and an irreducible polynomial $f \in F[x]$, we want to build a larger field L . One way to do this is to take a root a of f and adjoin it to F obtaining the field $L = F(a)$, which is the collection of all expressions that one can build using addition, subtraction, multiplication and division starting from the elements of $F \cup \{a\}$. Another way to build a larger field L from a smaller field F and an irreducible polynomial $f \in F[x]$ is to let $L = F[x]/(f(x))$. We will show below that these two ways of creating larger fields are one and the same.

Throughout, we will need the following results about irreducible polynomials:

Theorem 15.1. *Let F be a field and $f \in F[x]$. An element $\alpha \in F$ is a root of f if and only if $f = (x - \alpha)g$ for some $g \in F[x]$.*

Theorem 15.2 (Eisenstein's Criterion). *Suppose R is a domain and let $n \geq 1$, and consider the monic polynomial*

$$f(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0 \in R[x].$$

If there exists a prime ideal P of R such that $a_0, \dots, a_{n-1} \in P$ and $a_0 \notin P^2$, then f is irreducible in $R[x]$.

Theorem 15.3 (Gauss' Lemma). *Let R be a UFD with field of fractions F . Regard R as a subring of F and $R[x]$ as a subring of $F[x]$ via the induced map $R[x] \hookrightarrow F[x]$. If $f(x) \in R[x]$ is irreducible in $R[x]$, then $f(x)$ remains irreducible as an element of $F[x]$.*

Theorem 15.4. *Let R be a UFD with field of fractions F . Regard R as a subring of F and $R[x]$ as a subring of $F[x]$ via the induced map $R[x] \hookrightarrow F[x]$. If $f(x) \in R[x]$ is irreducible in $F[x]$ and the gcd of the coefficients of $f(x)$ is a unit, then $f(x)$ remains irreducible as an element of $R[x]$.*

15.1 Definition and first properties

Definition 15.5. A **field extension** is an inclusion of one field F into a larger field L , making F into a subfield of L . We sometimes write $F \subseteq L$ and sometimes L/F to signify that L is a field extension of F .

So a field extension is just another name for a subfield, but the emphasis is different. We think of F as coming first and L later.

Remark 15.6. If $F \subseteq L$ is a field extension, then L is in particular an F -vector space. This is a special case of the more general fact that if $\phi: R \rightarrow S$ is a ring homomorphism, then S is a left R -module via $r \cdot s := \phi(r)s$ by restriction of scalars.

Definition 15.7. The **degree** of a field extension L/F is

$$[L : F] := \dim_F(L).$$

A field extension is **finite** if its degree is finite.

Here are some examples.

Example 15.8. Since $\mathbb{C} = \mathbb{R}(i) = \{a + bi \mid a, b \in \mathbb{R}\}$, we have $[\mathbb{C} : \mathbb{R}] = 2$.

Example 15.9. We have $[\mathbb{R} : \mathbb{Q}] = \infty$. In fact, to be more precise we should say that $[\mathbb{R} : \mathbb{Q}]$ is the cardinality of \mathbb{R} , but in general we lump all infinite field extensions together when talking about degree, and just write $[L : F] = \infty$.

Now we show that for any field F and any nonconstant polynomial f with coefficients in F , there exists a field extension of F in which the polynomial f has at least one root.

Theorem 15.10. Let F be a field, $p \in F[x]$ with $\deg(p) \geq 1$, and $L = F[x]/(p)$. If p is irreducible, then

(1) L/F is a field extension via the map

$$\begin{aligned} F &\longrightarrow L \\ f &\longmapsto f + (p). \end{aligned}$$

(2) The degree of the extension is $[L : F] = \deg(p)$.

(3) The element $\bar{x} := x + (p) \in L$ is a root of p in L .

Proof. First, note that (p) is a nonzero principal ideal in $F[x]$. Recall that over a PID, ideals generated by an irreducible element are maximal. Since p is irreducible, we conclude that (p) is maximal, and thus $F[x]/(p)$ is a field.

We regard L as a field extension of F via the canonical map $F \rightarrow L$ sending $f \in F$ to the coset of the constant polynomial f . This map is not technically an inclusion map, but since it is an injective map we can pretend that it is an inclusion by identifying F with its image under this map. Note that injectivity of this map follows from the fact that (p) is a

proper ideal of $F[x]$, and thus every nonzero constant $a \in F$ is taken to a nonzero element in $L = F[x]/(p)$.

You showed in Problem Set 6 that if $\deg(p) = n$, then the classes of $1, x, \dots, x^{n-1}$ modulo (p) form basis for L regarded as an F -vector space. Therefore, $[L : F] = \deg(p)$. Moreover, we can extend the inclusion $F \subseteq L$ to an inclusion $F[x] \subseteq L[x]$, and thus we can regard p as belonging to $L[x]$. Setting $\bar{x} = x + (p) \in L$, the element \bar{x} is a root of $p(x) \in L[x]$ since

$$p(\bar{x}) = p(x) + (p(x)) = 0_L.$$

□

Example 15.11. The polynomial $f(x) = x^2 + 1$ is irreducible over \mathbb{R} . ?? says that f has a root in the extension $\mathbb{R}[x]/(x^2+1)$, and indeed, $\mathbb{R}[x]/(x^2+1) \cong \mathbb{C}$, where f factors completely into linear factors: $f(x) = (x - i)(x + i)$. In fact, $\mathbb{R}[x]/(x^2 + 1) \cong \mathbb{R}[i]$.

Now that we know that there exists a field extension of F in which $p(x)$ has a root, we may wonder about the *smallest* such extension.

Definition 15.12. Let $F \subseteq L$ be a field extension and $\alpha \in L$. We write $F(\alpha)$ for the smallest subfield of L that contains all of F and α .

In contrast with the previous definition, we will also consider the smallest *ring* containing F and α .

Remark 15.13. Since the intersection of any two subfields of L is again a subfield, $F(\alpha)$ exists and is given by

$$F(\alpha) = \bigcap_{\substack{E \text{ field} \\ F \cup \{\alpha\} \subseteq E \subseteq L}} E.$$

Definition 15.14. Let $F \subseteq L$ be a field extension and $\alpha \in L$. We write

$$F[\alpha] := \{f(\alpha) \mid f \in F[x]\}.$$

Remark 15.15. Note that any subring of L containing F and α must contain all products of α and elements of F , and all linear combinations of such things. Thus $F[\alpha]$ is the smallest subring of L containing F and α . Note that our notation does not include L , since in fact $F[\alpha]$ does not actually depend on the choice of L as long as $L \ni \alpha$.

Here is another way to describe this field $F(\alpha)$. We leave the proof for Problem Set 7.

Lemma 15.16. If $F \subseteq L$ is a field extension and $\alpha \in L$, the field $F(\alpha)$ is the fraction field of $F[\alpha] = \{f(\alpha) \mid f \in F[x]\}$: more precisely,

$$F(\alpha) = \left\{ \frac{g(\alpha)}{f(\alpha)} \mid g(x), f(x) \in F[x], f(\alpha) \neq 0 \right\}.$$

Soon we will give an even better description for $F(\alpha)$ in the case where α is the root of a polynomial $p \in F[x]$.

Definition 15.17. A field extension L/F is called **simple** if $L = F(\alpha)$ for some element α of L . We call such an α a **primitive element** for the extension.

If L/F is a simple field extension, note that there might be many different elements $\alpha \in L$ such that $L = F(\alpha)$. Thus primitive elements are not necessarily unique.

Example 15.18. The extension $\mathbb{R} \subseteq \mathbb{C}$ is simple, and i is a primitive element: $\mathbb{C} = \mathbb{R}(i)$. For another choice of primitive element, take $-i$.

We can generalize this to adjoining a subset instead of a single element.

Definition 15.19. If $F \subseteq L$ is a field extension and A is any subset of L , the **subfield generated by A over F** , denoted $F(A)$, is the smallest subfield of L that contains all of F . If $A = \{a_1, \dots, a_n\}$ is a finite set, we write $F(a_1, \dots, a_n)$ for $F(A)$.

Remark 15.20. Again, since the intersection of any two subfields of L is again a subfield, $F(A)$ exists and is given by

$$F(A) = \bigcap_{E \supseteq F, A} E.$$

Example 15.21. Regard \mathbb{Q} as a subfield of \mathbb{C} and let $F = \mathbb{Q}(\sqrt{2}, \sqrt{3})$. Setting $E = \mathbb{Q}(\sqrt{2})$, we can also think of F as $F = E(\sqrt{3})$. We will see shortly that $E = \{a + b\sqrt{2} \mid a, b \in \mathbb{Q}\}$. In other words, E is the set of \mathbb{Q} -linear combinations of 1 and $\sqrt{2}$, so $[E : \mathbb{Q}] = 2$.

Since $\sqrt{3}^2 \in \mathbb{Q} \subseteq E$, every element in F can be rewritten as an E -linear combination of 1 and $\sqrt{3}$:

$$F = \{\alpha + \beta\sqrt{3} \mid \alpha, \beta \in E\} = \{(a + b\sqrt{2}) + (c + d\sqrt{2})\sqrt{3} \mid a, b, c, d \in \mathbb{Q}\}.$$

and On the other hand, $E \neq F$, so we conclude that $[F : E] = 2$.

We claim that F is in fact a simple extension of \mathbb{Q} ; more precisely, that $\mathbb{Q}(\sqrt{2} + \sqrt{3}) = F$. Set $\beta := \sqrt{2} + \sqrt{3}$. Note that $\beta^2 = 5 + 2\sqrt{6}$ and

$$\beta^3 = 5\sqrt{2} + 5\sqrt{3} + 4\sqrt{3} + 6\sqrt{2} = 11\sqrt{2} + 9\sqrt{3}.$$

So $\frac{1}{2}(\beta^3 - 9\beta) = \sqrt{2}$, and hence $\sqrt{2} \in \mathbb{Q}(\beta)$. Likewise, $\sqrt{3} = -\frac{1}{2}(\beta^3 - 11\beta) \in \mathbb{Q}(\beta)$. So $\mathbb{Q}(\beta) = \mathbb{Q}(\sqrt{2}, \sqrt{3})$. This shows that $\mathbb{Q}(\sqrt{2}, \sqrt{3})/\mathbb{Q}$ is simple and $\sqrt{2} + \sqrt{3}$ is a primitive element of this field extension.

This example is an illustration of the Primitive Element Theorem, which we might or might not have time to prove this semester: every finite extension of \mathbb{Q} is simple.

Next we will show that if α is a root of a given polynomial $p(x) \in F[x]$, then $F(\alpha)$ is determined by $p(x)$ up to isomorphism.

Theorem 15.22. Let L/F be a field extension and let $p(x) \in F[x]$ be an irreducible polynomial. If p has a root $\alpha \in L$, then there is an isomorphism ϕ with $\phi|_F = \text{id}_F$ and

$$\begin{aligned} \frac{F[x]}{(p(x))} &\xrightarrow{\quad} F(\alpha) \\ x + (p(x)) &\mapsto \alpha \\ f(x) + (p(x)) &\mapsto f(\alpha). \end{aligned}$$

Proof. Let $\tilde{\phi} : F[x] \rightarrow F(\alpha)$ be the evaluation homomorphism that sends $x \mapsto \alpha$; more precisely, $\tilde{\phi}(f(x)) := f(\alpha)$, and the restriction of this map to F is the identity on F . Since $p(\alpha) = 0$, we have $(p(x)) \subseteq \ker(\tilde{\phi})$, and since $(p(x))$ is a maximal ideal and $\ker(\tilde{\phi}) \neq F[x]$, we conclude that $(p(x)) = \ker(\tilde{\phi})$.

Now by Theorem 11.35 we get an injective ring homomorphism

$$\phi : \frac{F[x]}{(p(x))} \rightarrow F(\alpha)$$

such that $\phi(f(x) + (p(x))) = \tilde{\phi}(f(x)) = f(\alpha)$.

It remains to be shown that ϕ is surjective. We will actually show more, namely that $\text{im}(\phi) = F[\alpha] = F(\alpha)$. Note first that by the definition of ϕ above, the image of $\tilde{\phi}$ on $F[x]$ is $F[\alpha]$. However, since ϕ is injective the image of $\tilde{\phi}$ is a field contained in $F(\alpha)$, and since the smallest field containing $F[\alpha]$ is $F(\alpha)$, we must in fact have $\text{im}(\tilde{\phi}) = F(\alpha)$. \square

Let's formalize the extra information we have obtained in the course of proving the theorem. First we used the following useful fact:

Remark 15.23. If $\phi : F \rightarrow L$ is an injective ring homomorphism and F and L are fields then the image of ϕ is a subfield of L .

Corollary 15.24. Let L/F be a field extension and let $p(x) \in F[x]$ be irreducible having a root $\alpha \in L$. Then $F[\alpha] = F(\alpha)$.

Corollary 15.25 (Uniqueness of $F(\alpha)$). Let $p(x) \in F[x]$ be irreducible and let α and β be two roots of $p(x)$ in some extensions L and K of F . Then $F(\alpha) \cong F(\beta)$, so that the two roots are algebraically indistinguishable.

Example 15.26. Taking $p(x) = x^2 + 1 \in \mathbb{R}[x]$ with roots $\alpha = i$ and $\beta = -i$ in \mathbb{C} , we actually obtain *equal* fields $\mathbb{R}(i) = \mathbb{C} = \mathbb{R}(-i)$. But ?? gives that there is an interesting isomorphism $\phi : \mathbb{C} \xrightarrow{\cong} \mathbb{C}$ that sends i to $-i$. In general, we have $\phi(a + bi) = a - bi$ for $a, b \in \mathbb{R}$.

Example 15.27. Another example illustrating ?? is that $\mathbb{Q}(\sqrt{2})$ and $\mathbb{Q}(-\sqrt{2})$ are isomorphic fields. In fact, they are equal: $\mathbb{Q}(\sqrt{2}) = \mathbb{Q}(-\sqrt{2})$. But again ?? gives that there is an interesting isomorphism $\phi : \mathbb{Q}(\sqrt{2}) \xrightarrow{\cong} \mathbb{Q}(-\sqrt{2}) = \mathbb{Q}(\sqrt{2})$ that sends $\sqrt{2}$ to $-\sqrt{2}$. In general, we have $\phi(a + b\sqrt{2}) = a - b\sqrt{2}$ for $a, b \in \mathbb{Q}$.

The two examples above preview the central idea of Galois theory.

Example 15.28. In ??, we showed that $\mathbb{Q}(\sqrt{2}, \sqrt{3}) = \mathbb{Q}(\sqrt{2} + \sqrt{3})$. We want to find a polynomial $p \in \mathbb{Q}[x]$ such that $\mathbb{Q}(\sqrt{2} + \sqrt{3}) \cong \mathbb{Q}[x]/(p(x))$. Set $\beta = \sqrt{2} + \sqrt{3}$.

Note that we have $\beta^2 = 5 + 2\sqrt{6}$ and $\beta^4 = 49 + 20\sqrt{6}$ and hence $\beta^4 - 10\beta^2 + 1 = 0$. So β is a root of $x^4 - 10x^2 + 1$. It can be shown that this polynomial is irreducible. How? First, Gauss' Lemma says that it is sufficient to show that it is irreducible in $\mathbb{Z}[x]$.

Suppose that f does factor. Then that factorization will be preserved when we go modulo p for any prime p . We will use this to show that f has no linear factors. When we go modulo

3, we claim that f has no roots: indeed, Fermat's Little Theorem says that $a^3 = a$ for all $a \in \mathbb{Z}/(3)$, so our polynomial becomes

$$f(x) = x^4 - x^2 + 1 = x^2 - x^2 + 1 = 1.$$

Since there are no roots modulo 3, we conclude that f has no linear factors over \mathbb{Z} either. Thus if f factors over \mathbb{Z} , it must factor as a product of degree 2 polynomials, which we can assume to be minimal. Suppose

$$f(x) = (x^2 + ax + b)(x^2 + cx + d).$$

These coefficients must satisfy the following system of equations:

$$\begin{cases} a + c &= 0 \\ b + d + ac &= -10 \\ ad + bc &= 0 \\ bd &= 1. \end{cases}$$

The first equation tells us that $a = -c$, so $0 = ad + bc = a(d - b)$, and since \mathbb{Z} is a domain, we conclude that $d = b$. Moreover, $b^2 = 1$, so $b \in \{-1, 1\}$. Finally, we have

$$b + d + ac = -10 \implies a^2 = 10 \pm 2.$$

But neither 8 nor 12 are squares in \mathbb{Z} , so this is impossible.

The previous example partially illustrates a nice trick: to show that a polynomial over \mathbb{Q} is irreducible, we need only to show it is irreducible over \mathbb{Z} , and to do that, it is sufficient to show that the polynomial is irreducible modulo a prime. In what follows, we will be very interested in irreducible polynomials, and we might want to use this type of tricks. Before we move on, let's see another example of this technique.

Example 15.29. Consider the polynomial $f(x) = x^4 - 10x^2 - 19 \in \mathbb{Q}[x]$. We claim it is irreducible, and thanks to Gauss' Lemma it is sufficient to show that f is irreducible over \mathbb{Z} . If f is reducible over \mathbb{Z} , it must also be reducible over $\mathbb{Z}/(p)$ for all primes, since going modulo p will preserve the fact that f factors. Modulo 3, our polynomial becomes

$$f(x) = x^4 + 2x^2 + 2.$$

Repeating the trick from ??, since x^4 and x^2 take the same values over $\mathbb{Z}/(3)$, we see that for any $a \in \mathbb{Z}/(3)$ we have

$$f(a) = a^4 + 2a^2 + 2 = 3a^2 + 2 = 2 \neq 0.$$

Thus f has no roots modulo 3, and thus it has no linear factors. Thus if it is reducible, it must be a product of two degree 2 factors, say

$$f(x) = (x^2 + ax + b)(x^2 + cx + d).$$

These coefficients must satisfy the following system of equations:

$$\begin{cases} a + c &= 0 \\ b + d + ac &= 2 \\ ad + bc &= 0 \\ bd &= 2. \end{cases}$$

Since $a = -c$, we get $a(d - b) = ad + bc = 0 \implies b = d$. Thus the last equation tells us that $b^2 = 2$, but all squares modulo 3 are 0 or 1, so this is impossible.

15.2 Algebraic and transcendental extensions

Definition 15.30. For a field extension $F \subseteq L$ and $\alpha \in L$, we say α is **algebraic** over F if $f(\alpha) = 0$ for some nonconstant polynomial $f \in F[x]$. Otherwise, we say α is **transcendental** over F .

Example 15.31. The element $i \in \mathbb{C}$ is algebraic over \mathbb{R} , since $i^2 + 1 = 0$. In fact, every element of \mathbb{C} is algebraic over \mathbb{R} , and we will soon see why. In contrast, the numbers π and e of \mathbb{R} are transcendental over \mathbb{Q} , though these are both deep facts.

Theorem 15.32. Suppose L/F is a field extension and $\alpha \in L$.

- (1) The set $I := \{f(x) \in F[x] \mid f(\alpha) = 0\}$ is an ideal of $F[x]$.
- (2) $I = 0$ if and only if α is transcendental over F . Equivalently, $I \neq 0$ if and only if α is algebraic over F .
- (3) If α is algebraic over F , meaning $I \neq 0$, then the unique monic generator $m_{\alpha,F}(x)$ of the ideal I is irreducible.
- (4) If α is algebraic over F , then there is an isomorphism of fields

$$F(\alpha) \cong F[x]/(m_{\alpha,F}(x))$$

sending F identically to F and sending x to α .

- (5) The element α is algebraic over F if and only if $[F(\alpha) : F] < \infty$. In this case,

$$[F(\alpha) : F] = \deg(m_{\alpha,F}(x)).$$

- (6) The element α is transcendental over F if and only if $[F(\alpha) : F] = \infty$. In this case, there is an isomorphism of fields between $F(\alpha)$ and the field of fractions of $F[x]$:

$$F(\alpha) \cong F(x) := \left\{ \frac{g(x)}{f(x)} \mid g \neq 0 \right\}$$

sending F identically to F and sending x to α .

Proof. The set I is the kernel of the evaluation homomorphism that maps $x \mapsto \alpha$. This map is a ring homomorphism, and thus I must be an ideal of $F[x]$. The content of (2) follows by definition of algebraic and transcendental elements.

To show (3), assume $I \neq 0$ and let p be its unique monic generator. Suppose $p = fg$. Since $p(\alpha) = 0$ in F and F is a field (and thus a domain), either $f(\alpha) = 0$ or $g(\alpha) = 0$. Therefore, either $f(x) \in I$ or $g(x) \in I$. This proves (3) is a prime ideal and hence p is a prime element. Since $F[x]$ is a PID, it follows that p is irreducible.

The statement of (4) is already ??.

Let's show (5). If α is algebraic over F , then (4) shows that

$$[F(\alpha) : F] = \deg(m_{\alpha,F}(x)) < \infty.$$

For the converse, if $[F(\alpha) : F] < \infty$, then the infinite list $1, \alpha, \alpha^2, \dots$ of elements of $F(\alpha)$ must be F -linearly dependent. Thus $a_0 + a_1\alpha + \cdots + a_n\alpha^n = 0$ for some n and some $a_0, \dots, a_n \in F$ not all zero. This shows α is the root of a nonzero polynomial.

To show (6), the map ϕ defined as in (4) is injective. Since the target is a field L and $F[x]$ is an integral domain, by the UMP of the fraction field ϕ can be extended to the field of fractions of $F[x]$, so there is a homomorphism of fields $\tilde{\phi} : F(x) \rightarrow L$. The image of this field map is

$$\left\{ \frac{g(\alpha)}{f(\alpha)} \mid g, f \in F[x], f(\alpha) \neq 0 \right\},$$

which is precisely $F(\alpha)$ by ???. The map is injective since it is a field homomorphism that is not identically zero. \square

Definition 15.33. Let $F \subseteq L$ be a field extension and $\alpha \in L$, and consider the ideal

$$I = \{f(x) \in F[x] \mid f(\alpha) = 0\}$$

from the previous theorem. The unique monic generator $m_{\alpha,F}(x)$ for I is called the **minimal polynomial** of α over F .

Remark 15.34. Note that the minimal polynomial of α over F , if it exists, must divide every polynomial in $F[x]$ that has α as a root. Also, it can be characterized as the monic polynomial in $F[x]$ of least degree having α as a root.

Example 15.35. Note that the minimal polynomial of i over \mathbb{R} is $m_{i,\mathbb{R}}(x) = x^2 + 1$.

Theorem 15.36 (The Degree Formula). *Suppose $F \subseteq L \subseteq K$ are field extensions. Then*

$$[K : F] = [K : L][L : F].$$

In particular, the composition of two finite extensions of fields is again a finite extension.

Proof. Let $A \subseteq K$ be a basis for K as an L -vector space and let $B \subseteq L$ be a basis for L as an F -vector space. Consider the subset of K given by

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

First, we claim that AB is a basis of K as an F -vector space. For $a \in K$, we have $a = \sum_i l_i a_i$ for some $a_1, \dots, a_m \in A$ and $l_1, \dots, l_m \in L$. For each i , l_i is an F -linear combination of a finite set of elements of B . Combining these gives that a is in the F -span of AB . To prove linear independence, it suffices to prove that if a_1, \dots, a_m and b_1, \dots, b_n be distinct elements of A and B respectively, then the set $\{a_i b_j\}$ is linearly independent. Suppose $\sum_{i,j} f_{i,j} a_i b_j = 0$ for some $f_{i,j} \in F$. Since the b_j are L -linearly independent and

$$\sum_{i,j} f_{i,j} a_i b_j = \sum_j \left(\sum_i f_{i,j} a_i \right) b_j$$

and $f_{i,j} a_i \in L$, we get that, for each j , $\sum_i f_{i,j} a_i = 0$. Using now that the a_i are F -linearly independent, we have that for all j and all i , $f_{i,j} = 0$. This proves that the set

$\{a_i b_j \mid i = 1, \dots, m, j = 1, \dots, n\}$ is linearly independent over F , and hence AB is linearly independent over F .

In particular, this shows that the elements of the form ab with $a \in A$ and $b \in B$ are all distinct, so $|AB| = |A| \cdot |B|$. Since AB is a basis for L over K , we conclude that

$$[K : F] = [K : L][L : F]. \quad \square$$

Example 15.37. In ?? we showed that $\mathbb{Q}(\sqrt{2}, \sqrt{3}) = \mathbb{Q}(\beta)$ with $\beta = \sqrt{2} + \sqrt{3}$. We claim that $m_{\beta, \mathbb{Q}}(x) = x^4 - 10x^2 + 1$. By the Degree Formula, we have

$$[\mathbb{Q}(\beta) : \mathbb{Q}] = [\mathbb{Q}(\beta) : E][E : \mathbb{Q}] = 2 \cdot 2 = 4.$$

Thus $m_{\beta, \mathbb{Q}}(x)$ has degree 4. We already know that β is a root of $x^4 - 10x^2 + 1$, hence this polynomial is divisible by the minimal polynomial of β . Since they are both monic and have degree 4, it must be that $m_{\beta, \mathbb{Q}}(x) = x^4 - 10x^2 + 1$. Arguing this way, there is no need to check this polynomial is irreducible; it must be by ?? (3).

Definition 15.38. A field extension $F \subseteq L$ is **algebraic** if every element $a \in L$ is algebraic over F .

Definition 15.39. We say an extension of fields $F \subseteq L$ is **finite** if it has finite dimension.

Note: this is not a statement about the number of elements in the fields F and L .

Example 15.40. The extension $\mathbb{R} \subseteq \mathbb{C}$ is finite, since $[\mathbb{C} : \mathbb{R}] = 2$.

Lemma 15.41. *Every finite extension of fields is algebraic.*

First proof. Let $K \subseteq L$ be a finite field extension, and let $a \in L$. Since the extension is finite, any infinite set of elements in L must be linearly dependent over F . In particular, the set

$$\{a^n \mid n \geq 0\}$$

is linearly dependent, so there exists n such that

$$\{1, a, a^2, \dots, a^n\}$$

is linearly dependent. Writing an equation of linear dependence, say

$$b_n a^n + \dots + b_1 a + b_0 = 0$$

for some $b_i \in F$, we might as well assume that $b_n \neq 0$ (otherwise, replace n by the largest value of i such that $b_i \neq 0$), and thus after multiplying by b_n^{-1} we conclude that we can write a^n in terms of the lower powers of a . In particular, a is algebraic over F . \square

Proof using the Degree formula. Let $K \subseteq L$ be a finite field extension, and let $a \in L$. By the Degree Formula, we have

$$[L : K] = [L : K(a)][K(a) : K],$$

and thus $K \subseteq K(a)$ must be finite. By ?? (5), a must be algebraic over K . \square

The converse is false, as shown by the following example:

Example 15.42. Let $\overline{\mathbb{Q}}$ denote the set of complex numbers that are algebraic over \mathbb{Q} , which is by definition an algebraic extension of \mathbb{Q} . However, we claim that $\overline{\mathbb{Q}}$ is not finite over \mathbb{Q} .

First, let p any prime integer, $n > 0$ be any integer, and consider the polynomial $x^n - p$ over $\mathbb{Q}[x]$. By applying Eisenstein's Criterion with the prime ideal (p) , we conclude that $x^n - p$ is irreducible over \mathbb{Z} . By Gauss' Lemma, $x^n - p$ is also irreducible over \mathbb{Q} .

Now $\overline{\mathbb{Q}}$ contains the subextension $\mathbb{Q}(a)$, where a is a root of $x^n - p$. Since $x^n - p$ is irreducible over \mathbb{Q} , it is the minimal polynomial of a over \mathbb{Q} , and thus by ?? (5) we conclude that the degree of this extension is $[\mathbb{Q}(a) : \mathbb{Q}] = n$. Thus $\overline{\mathbb{Q}}$ contains subextensions of \mathbb{Q} of arbitrarily large degree. By the Degree Formula applied to $\mathbb{Q} \subseteq \mathbb{Q}(a) \subseteq \overline{\mathbb{Q}}$, if $\overline{\mathbb{Q}}$ had finite degree over \mathbb{Q} then that degree would be divisible by n for all n . We conclude that $[\mathbb{Q} : \mathbb{Q}] = \infty$.

Theorem 15.43. *Given field extensions $F \subseteq L \subseteq E$, L/F and E/L are both algebraic if and only if E/F is algebraic.*

Proof. (\Leftarrow) Suppose $F \subseteq E$ is algebraic. Every element in L is in E as well, and thus it is algebraic over F ; thus $F \subseteq L$ is algebraic. Moreover, any element $\alpha \in E$ is algebraic over F by assumption, so it satisfies a polynomial with coefficients in F . But any polynomial with coefficients in F is also a polynomial with coefficients in L , and thus α is algebraic over L .

(\Rightarrow) Fix $\alpha \in E$. We need to prove α is a root of some monic polynomial with coefficients in F . This is surprisingly hard to prove directly, and in fact the proof we will give is rather nonconstructive.

Since α is algebraic over L , it is a root of some polynomial $a_n x^n + \dots + a_1 x + a_0 \in L[x]$. Note that this polynomial belongs to $F(a_0, \dots, a_n)[x]$ too, and so α is algebraic over $F(a_0, \dots, a_n)$.

Consider the chain of field extensions

$$F \subseteq F(a_0) \subseteq F(a_0, a_1) \subseteq \dots \subseteq F(a_0, a_1, \dots, a_n) \subseteq F(a_0, \dots, a_n, \alpha).$$

Each $a_i \in L$ is algebraic over F for all i , and α is algebraic over $F(a_0, a_1, \dots, a_n)$, so each step in our tower of extensions consists of adding an algebraic element to the previous field. By ??, each step in this chain has finite dimension. By the Degree Formula,

$$[F(a_0, \dots, a_n, \alpha) : F] = [F(a_0, \dots, a_n, \alpha) : F(a_0, \dots, a_n)] \cdots [F(a_0) : F]$$

is finite. Moreover, if we reorder the tower above to start from $F \subseteq F(\alpha)$, by the Degree Formula we have

$$[F(\alpha) : F][F(a_0, \dots, a_n, \alpha) : F(\alpha)] = [F(a_0, \dots, a_n, \alpha) : F] < \infty.$$

Therefore, $[F(\alpha) : F]$ is finite. By ?? (5) again, α is algebraic over F . \square

In the proof of ??, we also showed the following corollary of the Degree Formula:

Corollary 15.44. *If $\alpha_1, \dots, \alpha_n$ are algebraic over F , then $F \subseteq F(\alpha_1, \dots, \alpha_n)$ is a finite algebraic extension.*

15.3 Algebraically closed fields and algebraic closure

Definition 15.45. For any field extension $F \subseteq L$, we define the **algebraic closure** of F in L to be the set

$$\overline{F}_L = \{\alpha \in L \mid \alpha \text{ is algebraic over } F\}.$$

Lemma 15.46. For any field extension $F \subseteq L$, the set \overline{F}_L is a subfield of L that contains F . Moreover, it is the largest subfield of L that is algebraic over F .

Proof. First, note that every element in $a \in F$ satisfies the monic polynomial $x - a$, and thus $F \subseteq \overline{F}_L$, which is in particular nonempty. The claims that $F \subseteq \overline{F}_L$ and that \overline{F}_L is the largest subfield of L that is algebraic over F follow from the definition of \overline{F}_L .

It remains to show that \overline{F}_L is a field: we need to show that \overline{F}_L is closed under addition, multiplication, and taking additive and multiplicative inverses. Let $\alpha, \beta \in \overline{F}_L$. Since α and β are algebraic over F and consequently β is algebraic over $F(\alpha)$, we have that $[F(\alpha) : F] < \infty$ and $[F(\alpha, \beta) : F(\alpha)] < \infty$. Thus by the Degree Formula the extension $F(\alpha, \beta)/F$ is finite, and hence algebraic by ???. It follows that every element of $F(\alpha, \beta)$ is algebraic over F . In particular $\alpha \pm \beta$, $\alpha\beta$, and α^{-1} (if $\alpha \neq 0$) are elements of $F(\alpha, \beta) \subseteq \overline{F}_L$. \square

The notion of algebraic closure is closely related (pun intended) to being algebraically closed.

Definition 15.47. A field L is **algebraically closed** if every polynomial $f(x) \in L[x]$ that is not a constant has a root in L .

This is equivalent to the condition that every nonconstant polynomial splits completely into linear factors.

Example 15.48. The Fundamental Theorem of Algebra says that any polynomial in $\mathbb{C}[x]$ completely factors as a product of linear terms, thus \mathbb{C} is an algebraically closed field.

Lemma 15.49. If $F \subseteq L$ is a field extension with L algebraically closed, then \overline{F}_L is also algebraically closed.

Proof. Let $f \in \overline{F}_L[x]$ be a nonconstant polynomial. Since $\overline{F}_L \subseteq L$, $f \in L[x]$, and thus f has a root in L , say $\alpha \in L$. Since α satisfies a polynomial in $\overline{F}_L[x]$, it must then be algebraic over \overline{F}_L . Thus $\overline{F}_L \subseteq \overline{F}_L(\alpha)$ is an algebraic extension, and $F \subseteq \overline{F}_L \subseteq \overline{F}_L(\alpha)$ is a composition of two algebraic extensions. By ??, $F \subseteq \overline{F}_L(\alpha)$ is algebraic. By definition, this says that α is algebraic over F , and thus $\alpha \in \overline{F}_L$. Therefore, f has a root over \overline{F}_L , and \overline{F}_L is algebraically closed. \square

Remark 15.50. In contrast, if L/F is a field extension with L not algebraically closed, then \overline{F}_L need not be algebraically closed. For example, think of the extremal case when $F = L$, where we must have $\overline{F}_L = F$, which is not algebraically closed by assumption.

Example 15.51. ?? shows that the field $\overline{\mathbb{Q}}$ defined in ?? is algebraically closed.

Definition 15.52. Given a field F , a field L is called an **algebraic closure** of F if $F \subseteq L$ is an algebraic field extension and L is algebraically closed.

Remark 15.53. Let L be an algebraic closure of F . Since L is algebraically closed by definition, by ?? we conclude that \overline{F}_L is algebraically closed. On the other hand, since $F \subseteq L$ is algebraic by definition, we conclude that $\overline{F}_L = L$. This explains why we say L is an *algebraic closure* of F .

Example 15.54.

- 1) Since $[\mathbb{C} : \mathbb{R}] = 2$, the extension $\mathbb{R} \subseteq \mathbb{C}$ is finite, and thus by ?? the extension $\mathbb{R} \subseteq \mathbb{C}$ must also be algebraic. Moreover, \mathbb{C} is algebraically closed by the Fundamental Theorem of Algebra. Thus \mathbb{C} is an algebraic closure of \mathbb{R} .
- 2) By ??, an algebraic closure inside an algebraically closed field is algebraically closed. Thus $\overline{\mathbb{Q}}_{\mathbb{C}} = \{z \in \mathbb{C} \mid z \text{ is algebraic over } \mathbb{Q}\}$ is an algebraic closure of \mathbb{Q} .

Next we will show that every field has a unique algebraic closure, so we can talk about *the* algebraic closure of a field. To do that, we first need a lemma.

Lemma 15.55. *If L/F is an algebraic field extension and every nonconstant polynomial $f(x) \in F[x]$ splits completely into linear factors in $L[x]$, then L is algebraically closed and hence is an algebraic closure of F .*

Proof. Suppose $g(x) \in L[x]$ is not constant. We need to prove g has a root in L . We may form a (possibly trivial) algebraic extension $L \subseteq E$ such that $g(x)$ has a root α in E . Note that E/F is algebraic and hence α is algebraic over F . So α is a root of some $f(x) \in F[x]$. But then $f(x) = \prod_i (x - \beta_i) \in L[x]$ and it follows that α must one of the β_i , and hence belongs to L . \square

We are going to use one more technical result, which will also be helpful to us later.

Theorem 15.56. *Let F be a field, f be an irreducible nonconstant polynomial, and consider a field isomorphism $\theta: F \rightarrow F'$. Consider the isomorphism $\tilde{\theta}: F[x] \rightarrow F'[x]$ induced by θ , and let $f' = \tilde{\theta}(f) \in F'[x]$ be the polynomial corresponding to f . Let α be any root of f in some field extension L of F , and α' be any root of f' in some field extension L' of F' . Then there exists a field isomorphism*

$$\hat{\theta}: F(\alpha) \rightarrow F'(\alpha')$$

that extends the map θ and sends α to α' .

Proof. The key point is that

$$F[x]/(f) \cong F(\alpha)$$

via a map that is the identity on F and sends x to α , as we saw in ???. Thus we have

$$F(\alpha) \cong F[x]/(f) \cong F'[x]/(f') \cong F'(\alpha')$$

with the middle isomorphism induced by θ . Tracking through these maps shows that it extends θ and sends α to α' :

$$\alpha \mapsto x + (f) \mapsto x + (f') \mapsto \alpha'.$$

\square

We are now ready to show that every field has an algebraic closure, and that algebraic closures are unique up to isomorphism.

Theorem 15.57 (Existence and uniqueness of algebraic closures). *For any field F , there exists an algebraic closure of F . If L and L' are two algebraic closures of the same field F , then there exists a field isomorphism $\phi : L \xrightarrow{\cong} L'$ such that $\phi|_F = \text{id}_F$.*

Proof of existence of algebraic closures. First, we reduce the proof of existence to the following:

Claim: There is an algebraic field extension $F \subseteq L$ such that every nonconstant polynomial in $F[x]$ has at least one root in L .

Let us assume the claim holds. By using this fact repeatedly, we may form a tower of field extensions

$$F = F_0 \subseteq F_1 \subseteq F_2 \subseteq \dots$$

such that, for all i , the extension $F_i \subseteq F_{i+1}$ is algebraic and every nonconstant polynomial in $F_i[x]$ has at least one root in F_{i+1} . At each step, we apply the claim to F_i to construct F_{i+1} .

Let $E := \cup_i F_i$. One can show E is a field and $F \subseteq E$ is algebraic (exercise). Given $f \in F[x]$, by assumption f has a root α in F_1 , and hence f factors as $f(x) = (x - \alpha)g(x)$ for $g(x) \in F_1[x]$. But then g has a root in F_2 and hence factors in $F_2[x]$. Repeating this we see f splits completely into linear factors in $F_n[x]$, where $n = \deg(f)$, and thus f splits completely into linear factors in $E[x]$. By ??, E is an algebraic closure of F .

Proof of Claim: Let S be the collection of all nonconstant polynomials with coefficients in F , and for each $f \in S$, pick an indeterminate y_f . Now we form the rather large polynomial ring $R = F[y_f \mid f \in S]$. Let I be the ideal generated by $f(y_f)$. We claim that I is a proper ideal. If not, then $1 \in I$, so we would have an equation of the form

$$1 = g_1 f_1(y_{f_1}) + \dots + g_m f_m(y_{f_m})$$

in R . There exists finite extension E of F in which each f_i has a root α_i : by ??, f_i has a root α_i in some extension of F , and $F(\alpha_1, \dots, \alpha_n)$ is a finite extension of F by ???. Evaluating the above equation by setting $y_{f_i} = \alpha_i$, we get $1 = 0$, which is impossible. This shows that I must be a proper ideal.

Since I is a proper ideal, it is contained in some maximal ideal \mathfrak{m} . The quotient ring $K := R/\mathfrak{m}$ is a field, and the composition $F \hookrightarrow R \twoheadrightarrow K$ is a ring map $F \rightarrow K$ between two fields, and thus must be injective. By a slight abuse of notation, we will think of this map as an actual inclusion. For any $f \in S$, in K we have $f(y_f) = 0$, so the image $\bar{y}_f \in K$ of $y_f \in R$ is a root of f . We have constructed a field extension $F \subseteq K$ such that every $f \in S$ has at least one root in K .

We are not quite done since it is not clear that K is algebraic over F . For each $f \in S$, pick a root $\beta_f \in K$ of f . Let $L = F(\beta_f \mid f \in S) \subseteq K$. Then L is algebraic over F and every member of S has at least one root in L . \square

Proof of uniqueness of algebraic closures. Suppose L and L' are two algebraic closures of F . Let S be the set of pairs (E, i) where E is a subfield of L that contains F and $i : E \hookrightarrow L'$ is a ring homomorphism with $i|_F = \text{id}_F$. Make S into a poset by declaring that $(E, i) \leq (E', i')$ whenever $E \subseteq E'$ and $i'|_E = i$.

One can show (exercise!) that S satisfies the hypotheses of Zorn's Lemma, and hence it has a maximal element (E, i) . We claim E must equal L . If not, we can find $\alpha \in L \setminus E$. Let $p(x) = m_{\alpha, E}$ and set $E' := i(E)$. So i maps E isomorphically onto E' . Let $p'(x)$ be the polynomial in $E'[x]$ corresponding to $p(x)$ via i , and pick any root α' of $p'(x)$ in L' . By ??, there is an isomorphism $E(\alpha) \rightarrow E'(\alpha')$ extending the isomorphism i . Since $E(\alpha) \subseteq L$, and by assumption $\alpha \notin E$, this contradicts the maximality of (E, i) .

Thus we have a field extension $F \subseteq i(L) \subseteq L'$ with $i(L) \cong L$ via an isomorphism that fixes F . It follows that $i(L)$ is also an algebraic closure of F . Since L'/F is algebraic, we must have $i(L) = L'$. Thus i is surjective, and thus an isomorphism. \square

We will then be able to talk about not just *an* algebraic closure of F but *the* algebraic closure of F , so we can simplify our notation a bit.

Definition 15.58. Given a field F , we will write \overline{F} for its algebraic closure inside an algebraically closed field extension of F .

By ??, \overline{F} is defined only up to isomorphism.

Example 15.59. The field \mathbb{C} is the algebraic closure of \mathbb{R} , so we write $\overline{\mathbb{R}} = \mathbb{C}$.

Example 15.60. In ??, we defined $\overline{\mathbb{Q}}$ as the set of complex numbers that are algebraic over \mathbb{Q} . In our notation from this chapter, this is what we denote by $\overline{\mathbb{Q}}_{\mathbb{C}}$, the algebraic closure of \mathbb{Q} in \mathbb{C} . Since \mathbb{C} is an algebraically closed field, this is *the* algebraic closure of \mathbb{Q} , which explains our notation $\overline{\mathbb{Q}}$ from ?. This field $\overline{\mathbb{Q}}$ is sometimes called the **field of algebraic numbers**.

Remark 15.61. Earlier we used the notation \overline{F}_L to denote the algebraic closure of F in L . If L is an algebraically closed field, then by ?? we know that \overline{F}_L is also algebraically closed, and since it is by definition algebraic over F , then \overline{F}_L is *an* algebraic closure of F , so in fact this is *the* algebraic closure of F (defined only up to isomorphism).

In contrast, if L is not an algebraically closed field, we saw in ?? that \overline{F}_L is not necessarily algebraically closed. In particular, \overline{F}_L is not necessarily an algebraic closure of F , and thus \overline{F}_L and \overline{F} might denote completely different things.

From now on, many constructions will happen inside an algebraic closure of a given field, and we will use the notation \overline{F} .

Remark 15.62. Every algebraically closed field is infinite. Indeed, if F is a finite field, say $F = \{a_1, \dots, a_n\}$, then the polynomial

$$(x - a_1) \cdots (x - a_n) + 1 \in F[x]$$

has no roots in F . In particular, the algebraic closure of any finite field is infinite.

15.4 Splitting fields

Definition 15.63. Let F be a field and let $f \in F[x]$ be a nonconstant polynomial. A **splitting field** of f over F is a field extension $F \subseteq L$ such that f splits completely into linear factors in $L[x]$, and f does not split completely into linear factors over any proper subfield of L that contains F .

A splitting field of f is given by adjoining all the roots of f .

Lemma 15.64. If $F \subseteq E$ is a field extension such that $f \in F[x]$ factors in $E[x]$ as

$$f = c \prod_{i=1}^n (x - \alpha_i)$$

for some $c, \alpha_1, \dots, \alpha_n \in E$, then $F(\alpha_1, \dots, \alpha_n)$ is a splitting field for f over F .

Proof. Note that c is just the coefficient of f in degree n , and thus $c \in F$. Thus $f(x)$ also factors as

$$f(x) = c \prod_{i=1}^n (x - \alpha_i)$$

in $F(\alpha_1, \dots, \alpha_n)[x]$. Hence, given some splitting field L of f over F , by the minimality condition in the definition, we must have $L \subseteq F(\alpha_1, \dots, \alpha_n)$. However, the splitting field L must contain all roots of f in order for f to split completely in $L[x]$, so we also have $F(\alpha_1, \dots, \alpha_n) \subseteq L$. \square

Remark 15.65. Note that there may be repetitions in the list $\alpha_1, \dots, \alpha_n$, but that does not affect the validity of anything here.

Theorem 15.66 (Existence of splitting fields). *Let F be a field and $f \in F[x]$ a nonconstant polynomial. There exists a splitting field L for f over F .*

Proof. Let \overline{F} be an algebraic closure of F , which exists by ???. Let $\alpha_1, \dots, \alpha_m$ be the roots of f in \overline{F} . By construction, $F(\alpha_1, \dots, \alpha_m)$ is a splitting field of f . \square

Example 15.67.

- a) As a silly example, if f already splits into linear factors over $F[x]$, then F itself is the splitting field of f over F .
- b) The splitting field of $f = x^2 + 1$ over \mathbb{R} is \mathbb{C} : the roots of f are i and $-i$, and $\mathbb{R}(i, -i) = \mathbb{C}$.
- c) Let q be any irreducible quadratic polynomial in $\mathbb{R}[x]$. You will show in problem set 10 that the splitting field of q is \mathbb{C} .

Remark 15.68. In general, to form a field extension given by adjoining all the roots of two polynomials g_1 and g_2 amounts to forming a splitting field of their product $g_1 g_2$. This naturally generalizes to any number of polynomials g_1, \dots, g_n : to adjoin all the roots of g_1, \dots, g_n is the same as forming the splitting field of $g_1 \cdots g_n$.

It seems intuitive that by adjoining all the roots of $f \in F[x]$ to F , we will get a *unique* field (up to isomorphism). That is, it seems intuitive that splitting fields are unique up to isomorphism. This is indeed true, but the proof is a bit technical. We will actually show something a bit stronger.

Theorem 15.69. *Let F be a field, f be a nonconstant polynomial, and a field isomorphism $\theta: F \rightarrow F'$. Consider the isomorphism $\tilde{\theta}: F[x] \rightarrow F'[x]$ induced by θ , and let $f' = \tilde{\theta}(f) \in F'[x]$ be the polynomial corresponding to f . Suppose L is a splitting field of f over F and L' is a splitting field of f' over F' . Then there is a field isomorphism $\hat{\theta}: L \rightarrow L'$ extending θ .*

Proof. We proceed by induction on the degree n of f . If f is linear, then so is f' , and in this case $L = F$ and $L' = F'$. We have shown this already in ??.

Let p be any irreducible factor of f , and let $\alpha \in L$ be any one of the roots of p . Let $p' = \tilde{\theta}(p)$ be the irreducible polynomial in $F'[x]$ that corresponds to p , and let α' be any one of the roots of p' . By ??, there is an isomorphism

$$\phi: F(\alpha) \rightarrow F'(\alpha')$$

extending θ and sending α to α' .

In $F(\alpha)$, f factors as $f = (x - \alpha)g$, and in $F'(\alpha')$. Moreover, since ϕ extends θ and $\phi(\alpha) = \alpha'$, it follows that $\phi(x - \alpha) = x - \alpha'$. Therefore, the corresponding polynomial $f' = \phi(f)$ factors as $f' = (x - \alpha')g'$, and we have

$$(x - \alpha')\phi(g) = \phi(x - \alpha)\phi(g) = \phi((x - \alpha)g) = \phi(f) = f' = (x - \alpha')g'.$$

Since F' is a domain, we conclude that $\phi(g) = g'$.

Since L is a splitting field of f over F , f factors completely over L and thus so must g . Moreover, any other field $E \supseteq F(\alpha)$ containing all the roots of g would also contain α , and thus all the roots of f , and thus $E = L$. Thus L is a splitting field for g over $F(\alpha)$, and L' is a splitting field of g' over $F(\alpha')$: Since $\deg(g) < \deg(f) = n$, it follows by induction that there is a field isomorphism $\hat{\theta}: L \rightarrow L'$ that extends ϕ and hence extends θ . \square

Corollary 15.70 (Uniqueness of the splitting field of $f(x)$ over the base field F). *Any two splitting fields L and L' of $f(x) \in F[x]$ over F are isomorphic via an isomorphism $\phi: L \rightarrow L'$ that fixes F , i.e. $\phi|_F = \text{id}_F$.*

Proof. Apply part (2) of ?? to $\theta = \text{id}_F$. \square

We will now be referring to *the* splitting field of F , rather than *a* splitting field, thanks to the uniqueness result above.

Corollary 15.71. *If L is the splitting field over F of an irreducible polynomial $f(x) \in F[x]$ and if $\alpha, \beta \in L$ are any two roots of f , then there is a field automorphism $s: L \rightarrow L$ such that $s|_F = \text{id}_F$ and $s(\alpha) = \beta$.*

Proof. We basically already proved this, but since it is of large importance, let's do so again:

Since α, β are roots of the same irreducible polynomial, by ?? there is an isomorphism $\tau: F(\alpha) \rightarrow F(\beta)$ such that $\tau|_F = \text{id}_F$ and $\tau(\alpha) = \beta$. We have two field maps, the inclusion $F(\alpha) \hookrightarrow L$ and the composition of $F(\alpha) \xrightarrow{\tau} F(\beta) \hookrightarrow L$, and realize L as the splitting field of f over $F(\alpha)$ in two different ways. Since splitting fields are unique, by ??, an isomorphism such as s exists. \square

Example 15.72. Let L be the splitting field of $x^3 - 2$ over \mathbb{Q} , so $L = \mathbb{Q}(\sqrt[3]{2}, e^{2\pi i/3}\sqrt[3]{2}, e^{4\pi i/3}\sqrt[3]{2})$. ?? says that there is a field automorphism s of L such that

$$s(e^{2\pi i/3}\sqrt[3]{2}) = e^{4\pi i/3}\sqrt[3]{2}.$$

In fact, complex conjugation gives such an isomorphism.

?? also says that there is a field automorphism τ of L such that

$$\tau(\sqrt[3]{2}) = e^{2\pi i/3}\sqrt[3]{2},$$

but it is not as clear what map this τ is.

Example 15.73. The splitting field of $f(x) = x^4 - 5x^2 + 6 = (x^2 - 2)(x^2 - 3)$ is

$$\mathbb{Q}(\sqrt{2}, -\sqrt{2}, \sqrt{3}, -\sqrt{3}) = \mathbb{Q}(\sqrt{2}, \sqrt{3}) = \mathbb{Q}(\sqrt{2} + \sqrt{3}).$$

Note that we have shown the last equality in ???. This is an example where the splitting field of $f \in F[x]$ is not the algebraic closure of F : we showed in ?? that $[\mathbb{Q}(\sqrt{2} + \sqrt{3}) : \mathbb{Q}] = 4$, while in ?? we showed that $[\overline{\mathbb{Q}} : \mathbb{Q}] = \infty$. Thus $\mathbb{Q}(\sqrt{2} + \sqrt{3}) \subsetneq \overline{\mathbb{Q}}$.

Lemma 15.74. *For every field F and every nonconstant polynomial $f \in F[x]$ of degree $n \geq 1$, every splitting field L for f over F satisfies $[L : F] \leq n!$.*

Proof. By ??, splitting fields are unique up to isomorphism, so we just need to show that there exists a splitting field L for f over F with $[L : F] \leq n!$.

Intuitively, we just need to adjoin all the roots of f , which is possible since we already know we can adjoin a root of any polynomial. More formally, we start by showing that there is a field extension E/F such that f splits completely in $E[x]$, but without the minimality condition. Proceed by induction on the degree n of f . In the base case, $n = 1$, so f is linear and so $E = F$ works.

Assume f has degree $n > 1$. We proved in ?? that there exists a field extension K of F such that f has a root α . In $K[x]$ we have $f = (x - \alpha)g$ with $\deg(g) = \deg(f) - 1 = n - 1$. By induction, there is a field extension E of K with $[E : K] \leq (n - 1)!$ in which g splits completely. Then f also splits completely in E and by the Degree Formula

$$[E : F] = [E : K][K : F] \leq (n - 1)!n = n!.$$

Finally, let

$$f(x) = \prod_i (x - \alpha_i)$$

be the factorization of f in $E[x]$, and set $L = F(\alpha_1, \dots, \alpha_n)$. By ??, L is a splitting field of f over F . By the Degree Formula,

$$[E : F(\alpha_1, \dots, \alpha_n)][F(\alpha_1, \dots, \alpha_n) : F] = [E : F] \leq n! \implies [F(\alpha_1, \dots, \alpha_n) : F] \leq n!. \quad \square$$

The degree of the splitting field of f can be $n!$, but it can also be much smaller.

Example 15.75. Let us find the splitting field L of $x^3 - 2$ over \mathbb{Q} , and the degree of this field. Its roots in \mathbb{C} are $\sqrt[3]{2}$, $\zeta_3 \sqrt[3]{2}$, and $\zeta_3^2 \sqrt[3]{2}$, where $\zeta_3 = e^{\frac{2\pi i}{3}}$. So

$$L = \mathbb{Q}(\sqrt[3]{2}, \zeta_3 \sqrt[3]{2}, \zeta_3^2 \sqrt[3]{2}).$$

It is useful to simplify this a bit, by noting that

$$\zeta_3 = \frac{\zeta_3^2 \sqrt[3]{2}}{\zeta_3 \sqrt[3]{2}} \in L$$

and thus

$$L = \mathbb{Q}(\sqrt[3]{2}, \zeta_3).$$

We know from ?? above that $[L : \mathbb{Q}] \leq 3! = 6$. We claim it is exactly 6. First, we have

$$\mathbb{Q} \subseteq \mathbb{Q}(\sqrt[3]{2}) \subseteq L.$$

Moreover, $x^3 - 2$ is irreducible over \mathbb{Q} , and $\sqrt[3]{2}$ satisfies this polynomial, so it must be the minimal polynomial of $\sqrt[3]{2}$ over \mathbb{Q} . Thus $[\mathbb{Q}(\sqrt[3]{2}) : \mathbb{Q}] = 3$. Note that $\mathbb{Q}(\sqrt[3]{2}) \subseteq \mathbb{R}$ but ζ_3 is not real, so $\mathbb{Q}(\sqrt[3]{2}) \subseteq L$ has degree at least two. The Degree Formula shows that

$$[L : \mathbb{Q}] = [L : \mathbb{Q}(\sqrt[3]{2})][\mathbb{Q}(\sqrt[3]{2}) : \mathbb{Q}] \geq 3 \cdot 2 = 6.$$

By ??, $[L : \mathbb{Q}] \leq 6$, so we conclude that $[L : \mathbb{Q}] = 6$.

Example 15.76. Let $f(x) = x^n - 1 \in \mathbb{Q}[x]$. Then f splits completely in $\mathbb{C}[x]$, and its n many roots are the n th roots of 1. One of these is $\zeta_n := e^{2\pi i/n}$. Notice that every other n th root of 1 is a power of this one. Thus $\mathbb{Q}(\zeta_n)$ is the splitting field of $x^n - 1$ over \mathbb{Q} . This is a somewhat special example: upon adjoining one of the roots of f we got all the others for free. This happens in other examples too, but it is certainly *not* a general principle.

In particular, we see that the degree of $\mathbb{Q} \subseteq \mathbb{Q}(\zeta_n)$ is at most n , far less than the bound of $n!$ given by ?.?. In fact, it is at most $n - 1$, since f factors as $(x - 1)(x^{n-1} + \dots + x + 1)$, and hence the minimum polynomial of ζ_n is a divisor of $x^{n-1} + \dots + x + 1$.

When $n = p$ is prime, then one can show that $x^{p-1} + \dots + x + 1$ is irreducible, and hence it must equal the minimum polynomial of ζ_p . So, in this case, the degree of $\mathbb{Q} \subseteq \mathbb{Q}(\zeta_p)$ is exactly $p - 1$. However, the degree of $\mathbb{Q} \subseteq \mathbb{Q}(\zeta_n)$ can be smaller than $n - 1$ in general; for example, when $n = 4$, $\zeta_4 = i$ and $[\mathbb{Q}(i) : \mathbb{Q}] = 2$. Note that $x^3 + x^2 + x + 1$ factors as

$$x^3 + x^2 + x + 1 = (x^2 + 1)(x + 1)$$

and $m_{i, \mathbb{Q}}(x) = x^2 + 1$.

Example 15.77. In Problem Set 9, you showed that the splitting field F of $f = x^4 + 5 \in \mathbb{Q}[x]$ is a degree 8 extension of \mathbb{Q} , and again $8 < 4! = 24$. This is also an example where adding one root does not give us the entire splitting field: since f is irreducible over \mathbb{Q} , which one can show via Eisenstein's Criterion, then for any root α of f we must have $[\mathbb{Q}(\alpha) : \mathbb{Q}] = 4$, but since $[F : \mathbb{Q}] = 8$, then $\mathbb{Q}(\alpha)$ must not contain at least one of the other roots of f .

Here is another interesting feature of this example: let

$$\alpha_1 = e^{\pi i/4} \sqrt[4]{5}, \quad \alpha_2 = e^{3\pi i/4} \sqrt[4]{5}, \quad \alpha_3 = e^{5\pi i/4} \sqrt[4]{5}, \quad \alpha_4 = e^{7\pi i/4} \sqrt[4]{5},$$

and note that these are the four roots of f . You showed in Problem Set 9 that $\mathbb{Q}(\alpha_1 + \alpha_4) \subseteq \mathbb{R}$, but since $\alpha_1, \dots, \alpha_4 \notin \mathbb{R}$, this says that none of the roots (including α_1 and α_2) is in $\mathbb{Q}(\alpha_1 + \alpha_4)$. This is in stark contrast with the example $\mathbb{Q}(\sqrt{2} + \sqrt{3}) = \mathbb{Q}(\sqrt{2}, \sqrt{3})$ from before.

15.5 Separability

Definition 15.78. Let R be a commutative ring. The **characteristic** $\text{char}(R)$ of R is defined to be the smallest positive integer n such that

$$n \cdot 1_R = \underbrace{1_R + \dots + 1_R}_n = 0_F$$

if such an integer exists, and 0 otherwise.

Example 15.79. Here are some familiar examples: $\text{char}(\mathbb{Z}) = 0$ and $\text{char}(\mathbb{Z}/n) = n$.

Definition 15.80. Given a field F , its **prime field** is the subfield of F generated by 1_F . More precisely, the field

$$\text{Frac}(\{k1_F \mid k \in \mathbb{Z}\}).$$

You proved the following lemma in Problem Set 8:

Lemma 15.81. *Let F be a field.*

- a) *The prime field of F is isomorphic to exactly one of the fields \mathbb{Q} or \mathbb{Z}/p .*
- b) *The characteristic $\text{char}(F)$ is either 0 or a prime number p .*

In prime characteristic p , the most important tool we have at our disposal is the Frobenius endomorphism $x \mapsto x^p$. This is a simple but very powerful tool. The fact that the p th power map is a ring homomorphism is known as the **Freshman's Dream**.

Lemma 15.82 (Freshman's Dream). *If R is a commutative ring of prime characteristic p , then the function*

$$\begin{aligned} F: R &\longrightarrow R \\ c &\longmapsto F(c) = c^p \end{aligned}$$

is a ring homomorphism.

Proof. Since

$$(a+b)^p = \sum_{k=0}^p \binom{p}{k} a^k b^{p-k}$$

and the binomial coefficients $\binom{p}{k}$ are divisible by p for any $1 \leq k \leq p-1$, it follows that

$$(a+b)^p = a^p + b^p.$$

Because we also have $(ab)^p = a^p b^p$ by commutativity of R , and $F(1) = 1^p = 1$, the function F is a ring homomorphism, as desired. \square

Remark 15.83. Let R be a commutative ring of prime characteristic p . Since $\text{End}(R)$ is closed under composition, the e th iterate of the Frobenius endomorphism F^e , given by

$$\begin{aligned} F^e = \underbrace{F \circ \dots \circ F}_{e \text{ times}}: R &\longrightarrow R \\ x &\longmapsto F^e(x) = x^{p^e} \end{aligned}$$

is also a ring homomorphism.

We are now ready to talk about separability.

Definition 15.84. Let F be a field, $f \in F[x]$ be a monic polynomial, and α be a root of f in some field extension L of F . The **multiplicity** of α in f is the number of times $x - \alpha$ appears in the factorization

$$f = \prod_i (x - \beta_i)$$

of f in some (equivalently, any) splitting field of f .

Definition 15.85. A polynomial $f \in F[x]$ is **separable** if the multiplicity of every root of f in \overline{F} is 1.

Example 15.86. The polynomial $x^3 - 1$ is separable in $\mathbb{R}[x]$ because it has 3 distinct roots in \mathbb{C} , namely 1, ζ_3 , and ζ_3^2 , but not in $\mathbb{Z}/3[x]$, since $x^3 - [1]_3 = (x - [1]_3)^3$.

Definition 15.87. For any field F and $f = a_n x^n + \dots + a_1 x + a_0 \in F[x]$, define its **derivative** to be

$$f' = n a_n x^{n-1} + (n-1) a_{n-1} x^{n-2} + \dots + 2 a_2 x + a_1.$$

Remark 15.88. The derivative is an F -linear map $F[x] \rightarrow F[x]$: indeed, for any $f, g \in F[x]$, we have

$$(f + g)' = f' + g' \quad \text{and} \quad (af)' = af'$$

for all $a \in F$.

Remark 15.89. If F is a field of characteristic 0, then every nonconstant polynomial $f \in F[x]$ has $f' \neq 0$; in fact, $\deg(f') = \deg(f) - 1$. In contrast, in prime characteristic p the condition $f' = 0$ does not imply f is constant. For example, $(x^p)' = px^{p-1} = 0$.

Lemma 15.90 (Criteria for separability). *Let F be a field and $f \in F[x]$.*

- a) *Given a root α of f in some field extension L of F , the multiplicity of α in f is at least 2 if and only if $f'(\alpha) = 0$.*
- b) *A polynomial f is separable if and only if $\gcd(f, f') = 1$ in $F[x]$.*
- c) *If f is irreducible in $F[x]$, then f is separable if and only if $f' \neq 0$.*

Proof. Let L be the splitting field of f .

- a) If $f = (x - \alpha)^2 g(x)$ in $L[x]$, then $f'(x) = 2(x - \alpha)g(x) + (x - \alpha)^2 g'(x)$, so $f'(\alpha) = 0$. Conversely, if $f = (x - \alpha)h(x)$ and $h(\alpha) \neq 0$, then $f'(x) = h(x) + (x - \alpha)h'(x)$ does not have α as a root.
- b) By 1), f is separable if and only if f has no common roots with f' . By a problem in Problem Set 9, we have $\gcd(f, f') = 1$ if and only if f and f' have no common roots in \overline{F} .
- c) Since the degree of f' is strictly less than the degree of f and f is irreducible, we have that $\gcd(f, f') \neq 1$ if and only if $f' = 0$. \square

Definition 15.91. An algebraic field extension L/F is **separable** if for every $\alpha \in L$ the minimal polynomial $m_{\alpha,F}$ of α over F is separable. An extension that is not separable is sometimes called **inseparable**.

Exercise 100. Let $\alpha_1, \dots, \alpha_n$ be algebraic elements over F , and let $L := F(\alpha_1, \dots, \alpha_n)$. Show that the extension $F \subseteq L$ is separable if and only if $m_{\alpha_i,F}$ is separable for every i .

Definition 15.92. A field F is **perfect** if every algebraic extension of F is separable.

Remark 15.93. Every irreducible polynomial in $F[x]$ is separable if and only if every algebraic field extension L/F is separable.

Corollary 15.94 (Every field of characteristic zero is perfect). *Let F be a field of characteristic zero. Every irreducible polynomial in $F[x]$ is separable and every algebraic field extension L/F is separable.*

Proof. For every $\alpha \in L$, its minimal polynomial $m_{\alpha,F}$ is nonconstant. Since $\text{char}(F) = 0$, $m'_{\alpha,F} \neq 0$. Since $m_{\alpha,F}$ is irreducible in $F[x]$, ?? implies $m_{\alpha,F}(x)$ is separable. \square

Example 15.95. The characteristic zero extension $\mathbb{Q} \subseteq \mathbb{Q}(\sqrt{2}, \sqrt{3})$ is algebraic, and thus separable by ??.

Lemma 15.96. *Let F be a field with prime characteristic $\text{char}(F) = p$.*

- a) *If b is an element of F that is not a p th power of an element of F and L is an algebraic extension of F that contains a root of $x^p - b$, then $F \subseteq L$ is not separable.*
- b) *If every element of F is the p th power of another element of F , then every algebraic extension $F \subseteq L$ is separable.*

Proof.

- a) In general, for such F and b , let α be a root of $x^p - b$ in some field extension of F and let $L := F(\alpha)$. We claim that $F \subseteq L$ is not separable; specifically, we claim $m := m_{\alpha,F}$ is not separable. Since α is a root of $x^p - b$, we have $m \mid x^p - b$. In $L[x]$, by the Freshman's Dream we have

$$(x - \alpha)^p = x^p - \alpha^p = x^p - b.$$

It follows that m must divide $(x - \alpha)^p$ in $L[x]$ and hence m must have the form $(x - \alpha)^i$ for some $1 \leq i \leq p$. But $i \neq 1$ since $\alpha \notin F$. Thus α is a multiple root of m and m is irreducible in $F[x]$.

- b) Given an irreducible polynomial $q \in F[x]$, if $q' = 0$, then we must have that q is a sum of terms of the form bx^{mp} , for some $m \geq 0$ and $b \in F$. By assumption, for each such term, we have $b = c^p$ for some $c \in F$, and thus each term of q has the form $(cx^m)^p$. By the Freshman's Dream, $q = g^p$ for some polynomial $g \in F[x]$. But this is impossible since q is irreducible. We conclude that $q' \neq 0$, which by ?? implies that q is separable. This shows that every irreducible polynomial over F is separable, and thus every algebraic extension over F is separable. \square

Remark 15.97. Let F be a field of prime characteristic p . The condition that every element in F is a p th power of another element in F is equivalent to saying that the Frobenius map is surjective. We can write this more succinctly as $F^p = F$. ?? says that a field of prime characteristic p is perfect if and only if $F^p = F$.

Example 15.98. Let p be a prime and t be a variable, let $F = \mathbb{Z}/p(t)$, and consider the field $L = F[z]/(z^p - t)$. The element $t \in F$ is not a p th power of any element in F , and $F \subseteq L$ is an extension containing a root of the polynomial $x^p - t$. By ??, $F \subseteq L$ is an inseparable extension. Moreover, the field F is not perfect.

Theorem 15.99 (Finite fields are perfect). *Every algebraic field extension of a finite field is separable.*

Proof. Problem Set 10. □

We have shown that fields of characteristic 0 and fields K of characteristic p such that $K = K^p$ are separable.

Chapter 16

Galois theory

An approximate definition of Galois Theory is the study of the symmetries enjoyed by the roots of a polynomial. As a simple example, the polynomial $x^2 + 1 \in \mathbb{R}[x]$ has two roots, and there are essentially indistinguishable from an algebraic point of view — which root is $\sqrt{-1}$ and which is the negative of it? It makes no difference, really.

For another example, consider $p(x) = x^3 - 2 \in \mathbb{Q}[x]$, which has three roots. As we will soon learn, these roots of $x^3 - 2$ are as symmetric as possible over \mathbb{Q} . On the other hand, $q(x) = x^4 - 2 \in \mathbb{Q}[x]$ has four roots, and we will soon see that these four root are not as symmetric as possible over \mathbb{Q} .

Before starting the chapter, you might want a reminder of group actions. Below we include some of the definitions we will need for your convenience, though it is highly recommended that you read through the relevant portion of the 817 notes.

Definition 16.1. For a group (G, \cdot) and a set S , an **action** of G on S is a function

$$G \times S \rightarrow S,$$

typically written as $(g, s) \mapsto g \cdot s$, such that

- $g \cdot (g' \cdot s) = (gg') \cdot s$ for all $g, g' \in G$ and $s \in S$, and
- $e_G \cdot s = s$ for all $s \in S$.

Let $\text{Aut}(S)$ denote the set of automorphisms of the set S , which is a group under composition of functions. A group action of G on S is a group homomorphism $G \rightarrow \text{Aut}(S)$.

Definition 16.2. An action of a group G on a set S is called **faithful** if the associated group homomorphism is injective. Equivalently, an action is faithful if and only if for a given $g \in G$, whenever $g \cdot s = s$ for all $s \in S$, it must be that $g = e_G$.

Definition 16.3. A group action of (G, \cdot) on S is **transitive** if for all $p, q \in S$ there is a $g \in G$ such that $q = g \cdot p$. Equivalently, an action is transitive if $\text{Orb}_G(p) = S$ for any $p \in S$.

Definition 16.4. Let G be a group acting on a set S . The equivalence relation on S induced by the action of G , written \sim_G , is defined by $s \sim_G s'$ if and only if there is a $g \in G$ such that $s' = g \cdot s$. The equivalence classes of \sim_G are called **orbits**, specifically the equivalence class

$$\text{Orb}_G(s) = \{g \cdot s \mid g \in G\}$$

is the orbit of S . The set of equivalence classes with respect to \sim_G is written S/G .

16.1 Group actions on field extensions

Definition 16.5. Let K be a field. The **automorphism group** of K , denoted $\text{Aut}(K)$, is the collection of field automorphisms of K , with the binary operation of composition.

Definition 16.6. Let K/F be a field extension. The **automorphism group** of K/F , denoted $\text{Aut}(K/F)$, is the collection of field automorphisms of K that restrict to the identity on F , with the binary operation of composition.

Exercise 101. Let K/F be a field extension. Then $\text{Aut}(K)$ is a group under composition of maps, and $\text{Aut}(K/F)$ is a subgroup of $\text{Aut}(K)$.

Some books write $\text{Gal}(L/F)$ for $\text{Aut}(L/F)$, and call it the Galois group of L over F . We will reserve that notation for a special type of finite extensions – those that are Galois – and use only $\text{Aut}(L/F)$ for the general case.

Example 16.7. The automorphism group $\text{Aut}(\mathbb{C}/\mathbb{R})$ has two elements: the identity map and the map given by complex conjugation. The fact that each of these is an element of $\text{Aut}(\mathbb{C}/\mathbb{R})$ amounts to the fact that complex conjugation commutes with addition and multiplication of complex numbers. To see these are all the automorphisms, suppose $\tau \in \text{Aut}(\mathbb{C}/\mathbb{R})$. Since $\tau|_{\mathbb{R}} = \text{id}_{\mathbb{R}}$, then for any $z = a + ib \in \mathbb{C}$ we have $\tau(z) = a + b\tau(i)$. Moreover,

$$-1 = \tau(-1) = \tau(i \cdot i) = \tau(i) \cdot \tau(i),$$

and so $\tau(i) = \pm i$.

Example 16.8. For any squarefree integer d , the group $\text{Aut}(\mathbb{Q}(\sqrt{d})/\mathbb{Q})$ also has two elements: the identity and the map defined by $a + b\sqrt{d} \mapsto a - b\sqrt{d}$. The details are similar to the previous example, so we leave them as an exercise.

Remark 16.9. Let L be a field and let $\sigma \in \text{Aut}(L)$. Then the UMP of polynomial rings gives that there is an induced ring homomorphism $(-)^{\sigma} : L[x] \rightarrow L[x]$ such that for each $q = a_nx^n + \cdots + a_0 \in K[x]$, we have

$$q^{\sigma}(x) = \sigma(a_n)x^n + \cdots + \sigma(a_0).$$

If $\sigma \in \text{Aut}(L/K)$ and $q \in K[x]$, then $q^{\sigma} = q$.

Lemma 16.10. Let K/F be a field extension, $\sigma \in \text{Aut}(K/F)$, and $q \in F[x]$.

a) For all $c \in K$, we have $\sigma(q(c)) = q(\sigma(c))$.

b) If $\alpha \in K$ is a root of q , then $\sigma(\alpha)$ also is a root of q .

Proof.

a) By assumption, σ is a homomorphism and it restricts to the identity on F . Thus for any polynomial $q = a_nx^n + \cdots + a_0 \in F[x]$, we have

$$\sigma(q(c)) = \sigma(a_nc^n + \cdots + a_0) = \sigma(a_n)\sigma(c)^n + \cdots + \sigma(a_0) = a_n\sigma(c)^n + \cdots + a_0 = q(\sigma(c))$$

b) If α is a root of f , then

$$\begin{aligned} 0 &= \sigma(0) = \sigma(q(\alpha)) \\ &= q(\sigma(\alpha)) \quad \text{by a)} \end{aligned}$$

showing that $\sigma(\alpha)$ is also a root of q . \square

We now come to the main idea connecting field extensions and groups. It concerns the *action* of the group of automorphisms of a splitting field of a polynomial on the set of roots of that polynomial.

Theorem 16.11. *Let L be the splitting field of a polynomial $f \in F[x]$. Let S be the set of distinct roots of f in L , and let $n := |S|$.*

a) *The group $\text{Aut}(L/F)$ acts faithfully on S , via*

$$\sigma \cdot b := \sigma(b)$$

for all $\sigma \in \text{Aut}(L/F)$ and $b \in S$, and hence $\text{Aut}(L/F)$ is isomorphic to a subgroup of S_n .

b) *If $f \in F[x]$ is an irreducible polynomial, then $\text{Aut}(L/F)$ acts transitively on S .*

Proof.

a) Let $G = \text{Aut}(L/F)$. To see that the action above is well-defined, notice that if $b \in S$ then $\sigma(b) \in S$ by ???. Now we have

$$\sigma \cdot (\sigma' \cdot b) = \sigma(\sigma'(b)) = (\sigma \circ \sigma')(b) \quad \text{for all } \sigma, \sigma' \in G, b \in S,$$

$$1_G \cdot b = \text{id}_K(b) = b \quad \text{for all } \sigma \in G \text{ and } b \in S,$$

so the given formula does indeed define an action of G on S .

This action is faithful: if σ fixes all the roots $\alpha_1, \dots, \alpha_n$ of f , then it fixes every element of $F(\alpha_1, \dots, \alpha_n) = L$. Thus the corresponding group homomorphism $\text{Aut}(L/F) \rightarrow \text{Aut}(S)$ is injective. On the other hand, the group of automorphisms on a set of n elements is isomorphic to S_n , so we have an inclusion of $\text{Aut}(L/F)$ into S_n , and thus $\text{Aut}(L/F)$ is isomorphic to a subgroup of S_n .

b) Let α, β be any two roots of f . By ??, there is an isomorphism $\theta : F(\alpha) \rightarrow F(\beta)$ that fixes F .

Our polynomial factors both as $f = (x - \alpha)g$ and $f = (x - \beta)h$. Since $f^\theta = f$ and $(x - \alpha)^\theta = x - \beta$, we must have $g^\theta = h$. ?? applies, showing there is an automorphism $\sigma : L \rightarrow L$ that extends θ . In particular, σ fixes F , since σ extends θ and $\theta|_F = \text{id}_F$, so $\sigma \in \text{Aut}(L/F)$. Moreover, since σ extends θ we have $\sigma(\alpha) = \theta(\alpha) = \beta$. This proves the action is transitive on the set of roots of any irreducible polynomial. \square

Soon we will show that if $f \in F[x]$ is separable but not necessarily irreducible, and L is the splitting field of f , then the orbits of the action of $\text{Aut}(L/F)$ on the set of roots of f are precisely the sets of roots of the same irreducible factor of f . But to do so, we will need a little bit of Galois theory.

Corollary 16.12. Let L be the splitting field of a polynomial $f \in F[x]$ with n distinct roots. Then $|\text{Aut}(L/F)| \leq n!$.

Proof. We showed in ?? that $\text{Aut}(L/F)$ is isomorphic to a subgroup of S_n , and thus it has at most $|S_n| = n!$ elements. \square

We will give an improved version of this result soon.

Exercise 102. Let F be a field and $L = F(a_1, \dots, a_n)$, where a_1, \dots, a_n are elements in some extension of F that are algebraic over F . Each element $\sigma \in \text{Aut}(L/F)$ is uniquely determined by $\sigma(a_1), \dots, \sigma(a_n)$.

A typical question that arises from ?? is to explicitly identify the automorphisms of a splitting field extension as a subgroup of the symmetric group.

Example 16.13. Let L be the splitting field of $f = x^3 - 2 \in \mathbb{Q}[x]$ and $G := \text{Aut}(L/\mathbb{Q})$. Recall from ?? that $L = \mathbb{Q}(\sqrt[3]{2}, \zeta)$, where $\zeta = e^{2\pi i/3}$, and that $[L : \mathbb{Q}] = 6$. Let us write the roots of f as

$$\alpha_1 = \sqrt[3]{2}, \alpha_2 = \zeta\alpha_1, \alpha_3 = \zeta^2\alpha_1.$$

From ??, G acts transitively on $\{\alpha_1, \alpha_2, \alpha_3\}$, and hence G is isomorphic to a subgroup of S_3 .

The restriction of complex conjugation to L determines an element s of G of order 2, since L is closed under complex conjugation. We have

$$s(\alpha_1) = \alpha_1, s(\alpha_2) = \alpha_3, s(\alpha_3) = \alpha_2$$

and so s corresponds to the permutation $(23) \in S_3$.

Since the action of G on the roots of f is transitive, there is also an element $\tau \in G$ such that $\tau(\alpha_1) = \alpha_2$. Such a τ corresponds to either (12) , (123) of S_3 . Either way, τ and s generate all of S_3 .

We conclude that $|G| = 6$, the maximum possible, and G is isomorphic to S_3 . You should think of this as saying that the roots of $x^3 - 2$ are as interchangeable as possible, since $\text{Aut}(L/\mathbb{Q})$ is as large as possible.

Example 16.14. Let L be the splitting field of $f = x^4 - 2$ over \mathbb{Q} . The roots of f are

$$\alpha_1 = \sqrt[4]{2}, \quad \alpha_2 = i\alpha_1, \quad \alpha_3 = -\alpha_1, \quad \alpha_4 = -i\alpha_1,$$

and $L = \mathbb{Q}(\alpha_1, i)$. Let us start by computing $[L : \mathbb{Q}]$. Consider the chain of extensions

$$\mathbb{Q} \subseteq \mathbb{Q}(\alpha_1) \subseteq L = \mathbb{Q}(\alpha_1)(i).$$

The extension $\mathbb{Q} \subseteq \mathbb{Q}(\alpha_1)$ has degree 4, since $x^4 - 2$ is irreducible,¹ and the extension $\mathbb{Q}(\alpha_1) \subseteq L$ has degree at most 2, since i is a root of the degree 2 polynomial $x^2 + 1$. Since $\mathbb{Q}(\alpha_1) \subseteq \mathbb{R}$ and L contains elements that are not real, the extension $\mathbb{Q}(\alpha_1) \subseteq L$ cannot be trivial, and thus it must have degree exactly 2. We conclude that $[L : \mathbb{Q}] = 8$.

¹By using Eisenstein's Criterion with the prime 2 to show f is irreducible over \mathbb{Z} , and Gauss' Lemma to show that f must then also be irreducible over \mathbb{Q} .

Set $G := \text{Aut}(L/\mathbb{Q})$. We know G is isomorphic to a subgroup of S_4 . Since $L = \mathbb{Q}(\alpha_1, i)$, by ?? any $\tau \in G$ is uniquely determined by what it does to α_1 and i . Such a τ must send α_1 to a root of f , and thus to one of $\alpha_1, \dots, \alpha_4$. Moreover, since i is a root of $x^2 + 1$, so is $\tau(i)$, by ??, and thus τ must send i to $\pm i$. By combining the possibilities for $\tau(\alpha_1)$ and $\tau(i)$, we have at most 8 possibilities, so $|G| \leq 8$. In particular, G corresponds to a *proper* subgroup of S_4 , and so the roots of $x^4 - 2$ do not have as many symmetries as are conceivable.

Claim: $|G| = 8$ and G is isomorphic to the subgroup of S_4 generated by $(2\ 4)$ and $(1\ 2\ 3\ 4)$.

Let s be the map obtained by restricting complex conjugation to L , and note that indeed $s \in \text{Aut}(L/\mathbb{Q})$. This map s corresponds to $(2\ 4) \in S_4$.

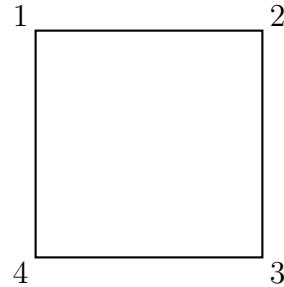
Now consider the field extension $\mathbb{Q}(i) \subseteq L$. Since $[L : \mathbb{Q}] = 8$ and $[\mathbb{Q}(i) : \mathbb{Q}] = 2$, by the Degree Formula we must have $[L : \mathbb{Q}(i)] = 4$. Since $L = \mathbb{Q}(i)(\alpha_1)$, the degree of $m_{\alpha_1, \mathbb{Q}(i)}$ must be 4. In particular, this shows that $x^4 - 2$ remains irreducible as a polynomial in $\mathbb{Q}(i)[x]$. So L is the splitting field of the irreducible polynomial $x^4 - 2$ over $\mathbb{Q}(i)$, and we may thus apply ?? to get that there is an element $\tau \in \text{Aut}(L/\mathbb{Q}(i))$ such that $\tau(\alpha_1) = \alpha_2$. We may regard τ as an element of $\text{Aut}(L/\mathbb{Q})$ too. Such a τ satisfies $\tau(i) = i$, so

$$\tau(\alpha_2) = \tau(i\alpha_1) = i\tau(\alpha_1) = i\alpha_2 = \alpha_3.$$

A key point here is that if we had merely taken τ to be an element of $\text{Aut}(L/\mathbb{Q})$ sending α_1 to α_2 , we would have no idea what τ does to α_2 : it was key to define $\tau \in \text{Aut}(L/\mathbb{Q}(i))$ as we did. We also get $\tau(\alpha_3) = \alpha_4$ and $\tau(\alpha_4) = \alpha_1$, so τ corresponds to the permutation $(1\ 2\ 3\ 4)$.

The proves that G is isomorphic to a subgroup of S_4 that contains $(2\ 4)$ and $(1\ 2\ 3\ 4)$. Since the subgroup generated by these two elements has order 8 and $|G| \leq 8$, then G must be the subgroup $\langle (2\ 4), (1\ 2\ 3\ 4) \rangle$ of S_n , and $|G| = 8$.

Finally, we claim that this subgroup of S_4 is isomorphic to D_4 . Indeed, consider a square with



Let $\rho \in D_4$ be the clockwise rotation by $\frac{\pi}{2}$ and $\tau \in D_4$ is the reflection across the line determined by the vertices 1 and 3. On the vertices of the square, the element ρ sends $1 \mapsto 2$, $2 \mapsto 3$, $3 \mapsto 4$, and $4 \mapsto 1$. Similarly, τ switches the vertices 2 and 4. One can check that the map

$$\begin{aligned} G &\longrightarrow D_4 \\ (1\ 2\ 3\ 4) &\longmapsto \rho \\ (2\ 4) &\longmapsto \tau \end{aligned}$$

determines an isomorphism.

16.2 Automorphism groups of finite field extensions

We will now focus on finite field extensions and their automorphism groups. We start by giving a much better upper bound on the order of the automorphism group of a finite field extension.

Theorem 16.15. *Let L/F be a finite field extension. Then*

$$|\text{Aut}(L/F)| \leq [L : F].$$

If L is the splitting field of a separable polynomial in $F[x]$, then

$$|\text{Aut}(L/F)| = [L : F].$$

Proof. We proceed by induction on $[L : F]$. In the base case, $[L : F] = 1$, and thus $L = F$, so $\text{Aut}(L/F)$ is the trivial group, and both statements hold.

Now let $n \geq 1$ and suppose that $|\text{Aut}(L/F)| \leq [L : F]$ holds for all L and F such that $[L : F] < n$. Let $[L : F] = n$. Pick $\alpha \in L \setminus F$ and let $m = m_{\alpha, F}$, and consider $F(\alpha)/F$.

Note that $H = \text{Aut}(L/F(\alpha))$ is a proper subgroup of $G = \text{Aut}(L/F)$. By induction, we have $|H| \leq [L : F(\alpha)]$. We claim that it suffices to prove $[G : H] \leq [F(\alpha) : F]$. Indeed, using the Degree Formula and the fact that $|G| = |H| \cdot [G : H]$, if $[G : H] \leq [F(\alpha) : F]$ then

$$|G| = |H| \cdot [G : H] \leq [L : F(\alpha)][F(\alpha) : F] = [L : F].$$

To show that $[G : H] \leq [F(\alpha) : F]$, consider the function

$$\begin{aligned} G/H &= \{\text{cosets of } H \text{ in } G\} \xrightarrow{\Psi} \{\text{roots of } m \text{ in } L\} \\ gH &\longmapsto g(\alpha). \end{aligned}$$

By ??, for any $g \in G$ the element $g(\alpha)$ is also a root of m . For any $h \in H$, we have

$$gh(\alpha) = g(h(\alpha)) = g(\alpha).$$

Thus Ψ is well-defined. Moreover, for any $g_1, g_2 \in G$ we have

$$\Psi(g_1H) = \Psi(g_2H) \iff g_1(\alpha) = g_2(\alpha) \iff g_2^{-1}g_1(\alpha) = \alpha$$

which is equivalent to saying that $g_2^{-1}g_1$ fixes $F(\alpha)$, and equivalently

$$g_2^{-1}g_1 \in H \iff g_1H = g_2H.$$

This proves that the function Ψ is injective.

By ??, $\deg(m) = [F(\alpha) : F]$. Thus Ψ is an inclusion of G/H into a set with at most $[F(\alpha) : F]$ many elements. Therefore,

$$[G : H] = |G/H| \leq [F(\alpha) : F].$$

Now suppose that f is separable, so that

$$f = c \prod_{i=1}^n (x - \alpha_i) \in L[x]$$

with $\alpha_i \neq \alpha_j$ for $i \neq j$ and $L = F(\alpha_1, \dots, \alpha_n)$.

Set $\alpha = \alpha_1$ and let m be the irreducible factor of f that has α as a root. Notice $m = m_{\alpha, F}$. As before, we consider $F(\alpha)$ and set $H = \text{Aut}(L/F(\alpha)) \leq \text{Aut}(L/F) = G$. Note that L is also the splitting field of

$$g = \prod_{i=2}^n (x - \alpha_i) \in F(\alpha)[x]$$

over $F(\alpha)$, and g is also separable. By induction $|H| = [L : F(\alpha)]$, and it remains to show that

$$[G : H] = [F(\alpha) : F] = \deg(m).$$

Since f is separable, so is m , so $\deg(m)$ is exactly the number of distinct roots of m . Showing that $[G : H] = \deg(m)$ amounts to the assertion that the injective map Ψ is also surjective. This holds since G acts transitively on the roots of m , as shown in ??.

The finite field extensions whose automorphism group is as large as possible are very important, and are the main subject of this final chapter.

Definition 16.16. A **finite** field extension $F \subseteq L$ is **Galois** if $|\text{Aut}(L/F)| = [L : F]$. In this case we write $\text{Gal}(L/F)$ for $\text{Aut}(L/F)$, and say $\text{Gal}(L/F)$ is the **Galois group** of L over F .

Example 16.17 (a nonexample). We claim that field extension $\mathbb{Q} \subseteq \mathbb{Q}(\sqrt[3]{2})$ is not Galois. Indeed, suppose $s \in \text{Aut}(L/\mathbb{Q})$. Then s is entirely determined by where it sends $\sqrt[3]{2}$ and it must send this element to another root of $x^3 - 2$. But the other two roots of this polynomial are not real, and hence not in L . So $s(\sqrt[3]{2}) = \sqrt[3]{2}$ and $s = \text{id}$.

This shows $\text{Aut}(L/\mathbb{Q})$ is the trivial group, so $\text{Aut}(L/\mathbb{Q}) = 1 < 3 = [L : \mathbb{Q}]$. In particular, the extension is not Galois.

?? tells us how to construct Galois extensions:

Corollary 16.18 (First construction of Galois extensions from splitting fields). *If L is the splitting field of a separable polynomial $f \in F[x]$, then L/F is Galois.*

Definition 16.19. Let $f \in F[x]$ be a separable polynomial with splitting field L . The **Galois group** of f is $\text{Gal}(L/F)$.

We will need the following notation.

Definition 16.20. If G is subgroup of $\text{Aut}(L)$, the **subfield of L fixed by G** , denoted L^G , is by definition

$$L^G := \{\alpha \in L \mid s(\alpha) = \alpha, \text{ for all } s \in G\}.$$

Note that the textbook writes this as L_G .

Exercise 103. If G is subgroup of $\text{Aut}(L)$, show that L^G is a subfield of L .

Example 16.21. Let $G = \text{Aut}(\mathbb{C}/\mathbb{R})$. Then \mathbb{C}^G is the subgroup of complex numbers fixed by all the elements in $\text{Aut}(\mathbb{C}/\mathbb{R})$, which we saw in ?? has only two elements, the identity and the conjugation map s . Therefore, \mathbb{C}^G is the set of complex numbers fixed by conjugation, and thus $\mathbb{C}^{\text{Aut}(\mathbb{C}/\mathbb{R})} = \mathbb{R}$.

16.3 The Fundamental Theorem of Galois Theory

The following is an important theorem with many corollaries. In fact, the Fundamental Theorem of Galois Theory, which we will state shortly, will follow from this result.

Theorem 16.22 (Artin's Theorem). *Let L be any field. If G is a finite subgroup of $\text{Aut}(L)$, then L^G is a subfield of L , the extension L/L^G is finite and Galois, and $\text{Gal}(L/L^G) = G$.*

Note that we really do mean equality here: both G and $\text{Gal}(L/L^G)$ are subgroups of $\text{Aut}(L)$, and the theorem states that they coincide. The containment $G \subseteq \text{Gal}(L/L^G)$ is clear: if $\sigma \in G$, then by construction σ fixes every element of L^G and hence $\sigma \in \text{Gal}(L/L^G)$. The point of the theorem is that the extension $L^G \subseteq L$ is always Galois and that if $\sigma \in \text{Aut}(L)$ fixes every element of L^G then σ must belong to G .

We will not prove Artin's Theorem right away. Instead, we will first deduce some of its consequences, including the Fundamental Theorem of Galois Theory. We will then illustrate the Fundamental Theorem with many examples and give some consequences of it too. Only then will we circle back to prove Artin's Theorem.

Example 16.23. The group $G = \{\text{id}_{\mathbb{C}}, \sigma\}$, where σ is complex conjugation, is a finite subgroup of $\text{Aut}(\mathbb{C})$. Artin's Theorem tells us that $\mathbb{C}^G \subseteq \mathbb{C}$ is finite and Galois with Galois group G . It follows that $[\mathbb{C} : \mathbb{C}^G] = |G| = 2$. We already knew all this, since $\mathbb{C}^G = \mathbb{R}$.

As we head towards the Fundamental Theorem of Galois Theory, we start by stating a few helpful corollaries of Artin's Theorem. These will also allow us to show that finite Galois extensions are precisely the splitting fields of separable polynomials.

Corollary 16.24. *Let L/F be any Galois extension. Then $F = L^{\text{Gal}(L/F)}$.*

Proof. Note that $F \subseteq L^{\text{Gal}(L/F)}$ holds by definition, and so

$$[L : F] = [L : L^{\text{Gal}(L/F)}][L^{\text{Gal}(L/F)} : F]$$

by the Degree Formula. But Artin's Theorem gives

$$[L : L^{\text{Gal}(L/F)}] = |\text{Gal}(L/F)|,$$

and we also know that $[L : F] = |\text{Gal}(L/F)|$. Therefore, $[L^{\text{Gal}(L/F)} : F] = 1$ and thus $F = L^{\text{Gal}(L/F)}$. \square

Example 16.25. We know from ?? that $L = \mathbb{Q}(\sqrt[4]{2}, i)$ is Galois over \mathbb{Q} with Galois group D_4 . More precisely, this identification is given by writting

$$\alpha_1 = \sqrt[4]{2}, \quad \alpha_2 = i\sqrt[4]{2}, \quad \alpha_3 = -\sqrt[4]{2}, \quad \alpha_4 = -i\sqrt[4]{2}$$

and labelling the four corners of a square with $\alpha_1, \dots, \alpha_4$, counter-clockwise. Consider

$$\beta := \alpha_1 + \cdots + \alpha_4$$

and $\gamma = \alpha_1 \cdots \alpha_4$. Then each of β and γ are fixed by every Galois automorphism and hence by ?? β and γ must be rational. In fact, one can easily see that $\beta = 0$ and $\gamma = 2$, but notice that the exact same reasoning would apply in general to the sum of roots and the product of roots in the splitting field of any separable polynomial.

Corollary 16.26. *Let $F \subseteq L$ be a Galois extension. For every $\alpha \in L$, $m_{\alpha,F}$ is separable and all of its roots belong to L . Moreover, $\text{Gal}(L/F)$ acts transitively on the set of roots of $m_{\alpha,F}$.*

Proof. Let $\alpha \in L$ and consider the orbit $\alpha_1 = \alpha, \dots, \alpha_m$ of α under the action of $\text{Gal}(L/F)$. Set

$$f := (x - \alpha_1) \cdots (x - \alpha_m).$$

For any $\tau \in \text{Gal}(L/F)$, since τ permutes the elements of any orbit then

$$f^\tau = (x - \tau(\alpha_1)) \cdots (x - \tau(\alpha_m)) = f.$$

This proves that f has all its coefficients in the field $F^{\text{Gal}(L/F)}$, which by ?? coincides with the field F . Thus $f \in F[x]$. Moreover, by construction f is separable. Since α is a root of f , the minimal polynomial $m_{\alpha,F}$ must divide f , and thus $m_{\alpha,F}$ is also separable and has all its roots in L .

Finally, this also shows that all the roots of $m_{\alpha,F}$ are on the orbit of α with respect to the action of $\text{Gal}(L/F)$, and thus $\text{Gal}(L/F)$ acts transitively on the set of roots of $m_{\alpha,F}$. \square

Remark 16.27. Note that any irreducible polynomial over F with a root in L is the minimal polynomial of some element in L . So ?? says in particular that if F/L is Galois and $f \in F$ is any irreducible polynomial, then $\text{Gal}(F/L)$ acts transitively on the set of roots of f .

Corollary 16.28. *A finite field extension L/F is Galois if and only if L is the splitting field of some separable polynomial with coefficients in F .*

Proof. We already proved before in ?? that if L is the splitting field of some separable polynomial $f \in F[x]$, then $F \subseteq L$ is Galois.

For the reverse direction, suppose that $F \subseteq L$ is a Galois extension. In particular, it is a finite extension, so $L = F(\beta_1, \dots, \beta_n)$ for some $\beta_1, \dots, \beta_n \in L$; for example, the β_i could be chosen to be an F -basis of L .

By ??, each $m_{\beta_i,F}$ is separable and all of its roots belong to L . Moreover, if γ_i is a root of $m_{\beta_i,F}$ and $i \neq j$, then we claim that γ_i is not a root of $m_{\beta_j,F}$. Indeed, if $m_{\beta_j,F}(\beta_i) = 0$, then $m_{\alpha_i,F}|m_{\alpha_j,F}$, but they are both monic irreducible polynomials over F , so we must have $m_{\alpha_i,F} = m_{\alpha_j,F}$.

Let m_1, \dots, m_s be the distinct polynomials among the $m_{\beta_i, F}$, and set

$$g := \prod_{i=1}^s m_i.$$

Since distinct m_i do not share any common roots and all the m_i are separable, their product g is also separable. Moreover, all of the roots of g belong to L , and hence the splitting field of g is contained in L . Since β_i is a root of g for all i , $L = F(\beta_1, \dots, \beta_n)$ must be precisely the splitting field of g . \square

Theorem 16.29. *Let L be the splitting field of a separable polynomial $f \in F[x]$. The orbits of the action of $\text{Aut}(L/F)$ on the set S of roots of f are the subsets of S that are the roots of the same irreducible factor of f .*

Proof. For each $b \in S$, the orbit of b is

$$\text{Orb}_{\text{Aut}(L/F)}(b) = \{\sigma(b) \mid \sigma \in \text{Aut}(L/F)\}.$$

Since b is a root of f , there exists an irreducible factor $p \in F[x]$ of f such that b is a root of p . Since $p \in F[x]$, by ?? we know that $\sigma(b)$ is a root of p for any $\sigma \in \text{Aut}(L/F)$. Thus the orbit of b is contained in the set of roots of p in L .

By ??, $F \subseteq L$ is a Galois extension. By ??, $\text{Gal}(L/F)$ acts transitively on the set of roots of p . Thus every root of p is in the orbit of b under the action of $\text{Aut}(L/F)$. We conclude that the orbit of b is precisely the set of roots of p . \square

Definition 16.30. Given a field extension $F \subseteq L$, an **intermediate field** is a subfield E of L that contains F , so that $F \subseteq E \subseteq L$.

Corollary 16.31. *If $F \subseteq L$ is Galois, then for any intermediate field E the extension $E \subseteq L$ is Galois.*

Proof. This follows from ??: if L is the splitting field over F of a separable polynomial $f \in F[x]$, then L is also the splitting field over E of the same polynomial f , but now viewed as a polynomial in E . \square

Remark 16.32 (Warning!). In the setting of ??, note that E need not be Galois over the original field F . For example, $L = \mathbb{Q}(\sqrt[3]{2}, e^{2\pi i/3})$ is Galois over $F = \mathbb{Q}$ but we saw in ?? that $E = \mathbb{Q}(\sqrt[3]{2})$ is not Galois over \mathbb{Q} . Nevertheless, ?? says that L is Galois over E .

Definition 16.33. Let E_1 and E_2 be two subfields of K . The **composite** of E_1 and E_2 , denoted E_1E_2 , is the smallest subfield of K containing both E_1 and E_2 ; more precisely, it is the intersection of all the subfields of K that contain both E_1 and E_2 .

Example 16.34. Let $E = \mathbb{Q}(\sqrt{2})$ and $F = \mathbb{Q}(\sqrt[3]{2})$. We claim that the composite of E and F is the field $L = \mathbb{Q}(\sqrt[6]{2})$. On the one hand, $\sqrt{2} = \sqrt[6]{2}^3 \in L$ and $\sqrt[3]{2} = \sqrt[6]{2}^2 \in L$, so $E, F \subseteq L$. On the other hand, any subfield of L containing both E and F must contain

$$\frac{\sqrt{2}}{\sqrt[3]{2}} = 2^{\frac{1}{2}-\frac{1}{3}} = 2^{\frac{3-2}{6}} = 2^{\frac{1}{3}} = \sqrt[3]{2}.$$

Thus $L = EF$.

Remark 16.35. Let $F \subseteq L$ be a field extension and consider two intermediary fields E_1 and E_2 . If $E_1 = F(\alpha_1, \dots, \alpha_n)$ and $E_2 = F(\beta_1, \dots, \beta_m)$, then $E_1 E_2 = F(\alpha_1, \dots, \alpha_n, \beta_1, \dots, \beta_m)$. If $\alpha_1, \dots, \alpha_n$ is a basis for E_1/F and β_1, \dots, β_m is a basis for E_2/F , then $\alpha_i \beta_j$ is a generating set for $E_1 E_2$ over F , so

$$[E_1 E_2 : F] \leq [E_1 : F][E_2 : F].$$

Notice, however, that the inequality might be strict.

We are finally ready for the Fundamental Theorem of Galois Theory:

Theorem 16.36 (Fundamental Theorem of Galois Theory). *Suppose L/F is a finite Galois extension. There is a bijection*

$$\begin{array}{ccc} \{\text{intermediate fields } E, \text{ with } F \subseteq E \subseteq L\} & \xleftarrow{\Psi} & \{\text{subgroups } H \text{ of } \text{Gal}(L/F)\} \\ E \mapsto & & \Psi(E) = \text{Gal}(L/E) \\ \Psi^{-1}(H) = L^H & \xrightarrow{\quad} & H. \end{array}$$

Moreover, this bijective correspondence enjoys the following properties:

(a) Ψ and Ψ^{-1} each reverse the order of inclusion.

(b) Ψ and Ψ^{-1} convert between degrees of extensions and indices of subgroups:

$$[\text{Gal}(L/F) : H] = [L^H : F] \iff [\text{Gal}(L/F) : \text{Gal}(L/E)] = [E : F].$$

(c) Normal subgroups correspond to intermediate fields that are Galois over F :

- If $N \trianglelefteq G$ then L^N/F is Galois.
- If E/F is Galois, then $\text{Gal}(L/E)$ is a normal subgroup of $\text{Gal}(L/F)$.

(d) If $E = L^N$ for a normal subgroup $N \trianglelefteq \text{Gal}(L/F)$, then $\text{Gal}(E/F) \cong \text{Gal}(L/F)/N$.

(e) If H_1, H_2 are subgroups of G with fixed subfields $E_1 = L^{H_1}$ and $E_2 = L^{H_2}$, then

- $E_1 \cap E_2 = L^{<H_1, H_2>}$ and $\text{Gal}(L/E_1 \cap E_2) = \langle H_1, H_2 \rangle$.
- $E_1 E_2 = L^{H_1 \cap H_2}$ and $\text{Gal}(L/E_1 E_2) = H_1 \cap H_2$.

Proof. First, we need to check that both functions are well-defined. For each intermediary field E , we know from ?? that L/E is also Galois, and hence it makes sense to write $\text{Gal}(L/E)$; moreover, any $\sigma \in \text{Gal}(L/E)$ is an automorphism of L that fixes E , and thus $F \subseteq E$, so $\sigma \in \text{Gal}(L/F)$. This shows that Ψ is well-defined. Conversely, given a subgroup H of $\text{Gal}(L/F)$, L^H is a subfield of L by ??.

Next, we need to check that Ψ and Ψ^{-1} are indeed inverse functions. Given a subgroup H of $\text{Gal}(L/F)$, we have $\text{Gal}(L/L^H) = H$ by Artin's Theorem. Thus

$$\Psi \circ \Psi^{-1}(H) = \Psi(L^H) = \text{Gal}(L/L^H) = H.$$

Conversely, given an intermediate field E , L/E is Galois by ??, and hence $L^{\text{Gal}(L/E)} = E$ by ???. Thus

$$\Psi^{-1} \circ \Psi(E) = \Psi(\text{Gal}(L/E)) = L^{\text{Gal}(L/E)} = E.$$

This establishes the fact that Ψ is indeed a bijective correspondence.

Now we check that Ψ satisfies the given list of properties. For brevity, set $G := \text{Gal}(L/F)$.

- (a) The fact that the correspondence is order reversing follows from the definitions. Given intermediate fields $E_1 \subseteq E_2$, any automorphism of L that preserves E_2 must also preserve E_1 , thus $\text{Gal}(L/E_2) \supseteq \text{Gal}(L/E_1)$. Conversely, if $H_1 \leq H_2 \leq \text{Gal}(L/E)$, then every $x \in L$ that is fixed by every $\sigma \in H_2$ must also be fixed in particular by every element of H_1 , so $L^{H_2} \supseteq L^{H_1}$.
- (b) By definition of Galois extension, $[L : F] = |G|$. By Artin's Theorem, for any subgroup $H \leq G$ the extension $L^H \subseteq L$ is also Galois, and thus by definition $[L : L^H] = |H|$. Using the Degree Formula, we have

$$[L^H : F] = \frac{[L : F]}{[L : L^H]} = \frac{|G|}{|H|} = [G : H].$$

So if $H = \Psi(E) = \text{Gal}(E/F)$, then $L^H = E$ and the formula above can be rewritten as

$$[\text{Gal}(L/F) : \text{Gal}(L/E)] = [E : F].$$

- (c) Suppose E is an intermediate field that is Galois over F . Fix $\sigma \in G$ and $\alpha \in E$. Since E/F is Galois, by ?? the polynomial $m_{\alpha,F}$ is separable and all of its roots are in E . By ??, $\sigma(\alpha)$ is also a root of $m_{\alpha,F}$, and thus $\sigma(\alpha) \in E$.

Suppose $\tau \in \text{Gal}(L/E)$. For any $\alpha \in E$ we have $\sigma(\alpha) \in E$, so $\tau(\sigma(\alpha)) = \sigma(\alpha)$. Thus

$$\sigma^{-1}(\tau(\sigma(\alpha))) = \sigma^{-1}(\sigma(\alpha)) = \alpha.$$

This proves that $\sigma^{-1}\tau\sigma \in \text{Gal}(L/E)$ and hence that $\text{Gal}(L/E) \trianglelefteq G$. We have shown that if E is Galois over F , then the corresponding subgroup $\text{Gal}(L/E)$ of G is normal.

For the converse, consider a normal subgroup $N \trianglelefteq G$ and the corresponding intermediate field $E = L^N$, so that $N = \text{Gal}(L/E)$. We will show that E is the splitting field over F of a separable polynomial in $F[x]$, and hence is Galois over F by ??.

Pick any $\alpha \in E$ and set $m := m_{\alpha,F}$. By ??, m is separable and all of its roots belong to L . We claim that all the roots must in fact belong to E . Since m is irreducible and L/F is Galois, by ?? G acts transitively on the set of roots of m . Thus, given be any other root $\beta \in L$ of m , there is a $\sigma \in G$ with $\sigma(\alpha) = \beta$. Since N is normal, for any $\tau \in N$ we have $\sigma\tau' = \tau\sigma$ for some $\tau' \in N$. But $\tau' \in N$ fixes E , so $\tau'(\alpha) = \alpha$. Therefore,

$$\beta = \sigma(\alpha) = \sigma\tau'(\alpha) = \tau\sigma(\alpha) = \tau(\beta)$$

which shows that β is also fixed by N . But then $\beta \in L^N = E$. Therefore, E contains all the roots of $m_{\alpha,F}$, and thus E must contain the splitting field of $m_{\alpha,F}$.

We have $E = F(\alpha_1, \dots, \alpha_l)$ for some $\alpha_1, \dots, \alpha_l \in E$. If m_1, \dots, m_n are the distinct polynomials among $m_{\alpha_1, F}, \dots, m_{\alpha_l, F}$, then E is the splitting field of the separable polynomial $m_1 \cdots m_n$. By ??, E is Galois over F .

If $E = L^N$ for a normal subgroup $N \trianglelefteq \text{Gal}(L/F)$, then $\text{Gal}(E/F) \cong \text{Gal}(L/F)/N$.

- (d) Let $E = L^N$ for a normal subgroup N of G . We want to show that $\text{Gal}(E/F)$ is isomorphic to G/N .

For each $\sigma \in G$, we claim that $\sigma(E) \subseteq E$. By ??, for all $\alpha \in E$ the element $\sigma(\alpha)$ is also a root of $m_{\alpha, F}$. But since E/F is Galois, it must contain all of the roots of $m_{\alpha, F}$, by ??, so $\sigma(\alpha) \in E$. Thus $\sigma(E) \subseteq E$, so the restriction of σ to E determines an injective field homomorphism $\sigma|_E : E \rightarrow E$. Since $\sigma|_F = \text{id}_F$, this map is also a linear transformation of vector spaces over F . But E is a finite vector space over F , and any injective linear transformation $E \rightarrow E$ must be bijective. We conclude that $\sigma|_E$ is an automorphism of E . We thus have a well-defined function

$$\begin{aligned}\phi : G &\longrightarrow \text{Gal}(E/F) \\ \sigma &\longmapsto \phi(\sigma) = \sigma|_E.\end{aligned}$$

Moreover, this map is a group homomorphism by construction. The kernel is the subgroup of G of automorphisms that restrict to the identity on E , which is precisely N . Hence we have an induced injective group homomorphism

$$\bar{\phi} : G/N \rightarrow \text{Gal}(E/F).$$

But $|N| = |\text{Gal}(E/F)| < \infty$, so this map $\bar{\phi}$ must be an isomorphism.

- (e) Let H_1 and H_2 be subgroups of G with fixed fields $E_1 = L^{H_1}$ and $E_2 = L^{H_2}$.

First, we will show that $E_1 \cap E_2 = L^{\langle H_1, H_2 \rangle}$. Given any $\alpha \in E_1 \cap E_2$, $\sigma(\alpha) = \alpha$ for all $\sigma \in H_1$ and all $\sigma \in H_2$, since $\alpha \in E_1$ and $\alpha \in E_2$, so $\alpha \in L^{\langle H_1, H_2 \rangle}$, and $E_1 \cap E_2 \subseteq L^{\langle H_1, H_2 \rangle}$. Conversely, if $\alpha \in L^{\langle H_1, H_2 \rangle}$, then $\sigma(\alpha) = \alpha$ for all $\sigma \in \langle H_1, H_2 \rangle$, so in particular $\sigma(\alpha) = \alpha$ for all $\sigma \in H_i$ and thus $\alpha \in L^{H_i} = E_i$. We conclude that $E_1 \cap E_2 = L^{\langle H_1, H_2 \rangle}$.

Now let us show that $E_1 E_2 = L^{H_1 \cap H_2}$. Since L/F is a finite extension, then by the Degree Formula both of the extensions E_1/F and E_2/F are finite. Let $E_1 = F(\alpha_1, \dots, \alpha_n)$ and $E_2 = F(\beta_1, \dots, \beta_m)$, so that $E_1 E_2 = F(\alpha_1, \dots, \alpha_n, \beta_1, \dots, \beta_m)$. For any $\sigma \in H_1 \cap H_2$ we have $\sigma(\alpha_i) = \alpha_i$ and $\sigma(\beta_j) = \beta_j$ for each i . Since $\sigma|_E$ is completely determined by its values on the α_i and β_j , by ??, but since σ restricted to $E_1 E_2$ agrees with the identity map on the generators, $\sigma|_{E_1 E_2} = \text{id}_{E_1 E_2}$. We conclude that $E_1 E_2 \subseteq L^{H_1 \cap H_2}$.

Moreover, given any $\sigma \in \text{Gal}(L/E_1 E_2)$, its restriction to $E_1 E_2$ is by definition the identity. Thus its restriction to the subfields E_1 and E_2 of $E_1 E_2$ must also be the identity, and therefore we have $\sigma \in \text{Gal}(L/E_1) = H_1$ and $\sigma \in \text{Gal}(L/E_2) = H_2$. We conclude that $\sigma \in H_1 \cap H_2$. Thus $\text{Gal}(L/E_1 E_2) \leq H_1 \cap H_2$. Since Ψ is order reversing, we conclude that $E_1 E_2 \supseteq L^{H_1 \cap H_2}$, giving us the desired equality. \square

This bijection Ψ in ?? is sometimes called the **Galois correspondence**.

Corollary 16.37. *The Galois correspondence induces a lattice isomorphism between the lattice of intermediate fields of a Galois extension L/F and the dual of the lattice of subgroups of $\text{Gal}(L/F)$.*

This is just a fancy way to rephrase the fact that intermediate fields correspond to subgroups in an order-reversing bijection.

Example 16.38. Let L be the splitting field of $x^4 - 2$ over \mathbb{Q} . Let us use the Fundamental Theorem of Galois Theory to list all intermediate fields for L/\mathbb{Q} and to determine which are Galois over \mathbb{Q} . By ??, we know $G := \text{Gal}(L/\mathbb{Q})$ corresponds to the 8 element subgroup of S_4 generated by $\sigma = (2\ 4)$ and $\tau = (1\ 2\ 3\ 4)$, where we number the roots of $x^4 - 2$ as

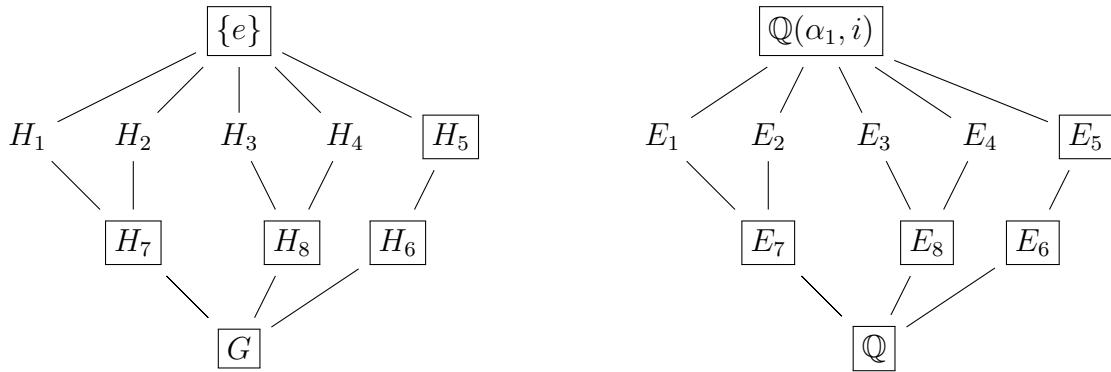
$$\alpha_1 = \sqrt[4]{2}, \quad \alpha_2 = i\alpha_1, \quad \alpha_3 = -\alpha_1, \quad \alpha_4 = -i\alpha_1.$$

We saw in ?? that this group is isomorphic to D_4 , and we can make this isomorphism explicit by labeling the four corners of a square 1, 2, 3, 4 counterclockwise, so that τ is rotation by 90 degrees and σ is reflection about the line joining vertices 1 and 3.

The subgroup lattice and intermediate field lattice are represented below, with normal subgroups and Galois extensions highlighted by boxes. The subgroups are

$\{e\}$	$G = \langle(2\ 4), (1\ 2\ 3\ 4)\rangle$
$H_1 = \langle(2\ 4)\rangle$	$H_5 = \langle(1\ 3)(2\ 4)\rangle$
$H_2 = \langle(1\ 3)\rangle$	$H_6 = \langle(1\ 2\ 3\ 4)\rangle$
$H_3 = \langle(1\ 2)(3\ 4)\rangle$	$H_7 = \langle(1\ 3), (2\ 4)\rangle$
$H_4 = \langle(1\ 4)(2\ 3)\rangle$	$H_8 = \langle(1\ 2)(3\ 4), (1\ 4)(2\ 3)\rangle$

and the lattices are of subgroups of G and intermediate fields of $\mathbb{Q} \subseteq \mathbb{Q}(\alpha_1, i)$ are



The intermediate fields are the fixed subfields of L associated to each of these subgroups. The group G corresponds to \mathbb{Q} and e corresponds to $L = \mathbb{Q}(\alpha_1, i)$. Set $E_i = L^{H_i}$.

The field E_1 has degree $4 = [G : H_1]$ over \mathbb{Q} . Since α_1 and α_3 belong to E_1 and $[\mathbb{Q}(\alpha_1) : \mathbb{Q}] = 4$, we must have $E_1 = \mathbb{Q}(\alpha_1)$. Likewise, $E_2 = \mathbb{Q}(\alpha_2)$.

The field E_3 also has degree 4 over \mathbb{Q} . Let

$$\beta := \alpha_1 + \alpha_2 = (1+i)\sqrt[4]{2} \in E_3.$$

If $[\mathbb{Q}(\beta) : \mathbb{Q}] = 2$, then β would be fixed by a subgroup of index 2 that contains $(1\ 2)(3\ 4)$, and the only possibility is H_8 . But $(1\ 4)(2\ 3)$ sends β to $\alpha_4 + \alpha_3 = -\beta \neq \beta$. So we must have

$[\mathbb{Q}(\beta) : \mathbb{Q}] = 4$ and hence $E_3 = \mathbb{Q}(\beta)$. We claim that $E_4 = \mathbb{Q}((1-i)\alpha_1)$, $E_5 = \mathbb{Q}(\sqrt{2}, i)$, and $E_7 = \mathbb{Q}(\sqrt{2})$, and leave it as an exercise.

The field E_6 has degree $[G : H_6] = 2$ over \mathbb{Q} , and so we merely need to find a single nonrational element of L fixed by τ . Since $\tau(i) = i$, we get $E_6 = \mathbb{Q}(i)$. Similarly, the field E_8 also has degree 2 over \mathbb{Q} and so we just need to find a single nonrational element fixed by the two generators of H_8 . Note that

$$\alpha_1\alpha_2 = \alpha_3\alpha_4 = i\sqrt{2},$$

and so $i\sqrt{2}$ is fixed by both $(12)(34)$ and $(14)(23)$. Thus $E_8 = \mathbb{Q}(i\sqrt{2})$.

Finally, we note that $G, \{e\}, H_5, H_6, H_7, H_8$ are normal subgroups of D_4 , since H_5 is the center of D_4 and each of H_6, H_7 , and H_8 has index two. Some messy checking reveals these to be the only normal subgroups of G . It follows from the Fundamental Theorem that $\mathbb{Q}, L, E_5, E_6, E_7, E_8$ are the only intermediate fields that are Galois over \mathbb{Q} .

As an example, to see directly that E_3 is not Galois over \mathbb{Q} , note that $(1+i)\sqrt[4]{2}$ is a root of $x^4 + 4$, which is irreducible; but $(1-i)\sqrt[4]{2}$ is also a root of this polynomial and it is not in E_3 .

Remark 16.39. Let $F \subseteq L$ be a Galois extension and consider an intermediate field E such that E/F is Galois, with corresponding normal subgroup $N := \text{Gal}(E/F) \trianglelefteq \text{Gal}(L/F)$. Part (d) of the Fundamental Theorem says there is an isomorphism $\text{Gal}(E/F) \cong \text{Gal}(L/F)/N$. In fact, the proof shows that the map

$$\begin{aligned}\phi : G &\longrightarrow \text{Gal}(E/F) \\ \sigma &\longmapsto \phi(\sigma) = \sigma|_E.\end{aligned}$$

is surjective. This says that every $\tau \in \text{Gal}(E/F)$ can be lifted to some $\sigma \in \text{Gal}(L/F)$, so that $\tau = \sigma|_E$.

Notice that while the proof shows that for every $\tau \in \text{Gal}(E/F)$ there exists $\sigma \in \text{Gal}(L/F)$ such that $\tau = \sigma|_E$, the proof is very much nonconstructive. In specific examples, one can sometimes determine σ explicitly by using some of the other tricks we have discussed.

16.4 Solvable polynomials and solvable groups

Galois' theory has many fun applications. There are some famous questions in geometry whose impossibility is shown via Galois theory methods:

- Trisecting an angle: is it possible to trisect any given angle using only a compass and a straightedge?
- Doubling the cube: using only a compass and straightedge, and given a cube, is it possible to construct a cube whose volume is exactly twice the volume of the original cube?
- Squaring the circle: Using only a compass and straightedge, is it possible to construct a square with the same area of a given circle?

All these challenges turn out to be impossible, which one can show using Galois theory. Unfortunately, we will not be proving these here. Instead, we will focus on another famous classical question which is behind the origins of Galois theory:

Question 16.40. Does there exist a formula for the roots of a polynomial of degree n with rational coefficients using only addition, subtraction, multiplication, division, and taking radicals?

The formulas for the roots of polynomials of degree 2, 3, and 4 have been known for hundreds of years, and they involve only sums, products, quotients, and square roots, cube roots, and fourth roots. It turns out that there are polynomials for which no such formula exists; in fact, there is no general formula for polynomials of degree 5.

The first ingredient we need is to better understand the process of taking roots. With that in mind, we will now discuss the Galois groups of the splitting fields of polynomials of the form $x^n - a$. These calculations will be used to prove what Galois himself sort of proved: if the roots of a polynomial can be expressed using iterated radicals, then the Galois group of its splitting field must be a solvable group.

Definition 16.41. A **primitive n th root of unity** over an arbitrary field F is an element ζ in the splitting field K of $x^n - 1$ over F (or in the algebraic closure \overline{F}) such that ζ generates the (multiplicative) subgroup of K :

$$\mu_n(K) := \{\alpha \in K \mid \alpha^n = 1\} \leq (K^\times, \cdot).$$

Exercise 104. Show that for every field K , every finite subgroup of K^\times is cyclic.

Remark 16.42. In particular, the subgroup $\mu_n(K)$ of K^\times of roots of $x^n - 1$ is a cyclic group. As a consequence, a primitive n th root of unity always exists.

Remark 16.43. Let F be any field and let K be the splitting field of $x^n - 1$ over F . Note that if ζ is a generator of $\mu_n(K)$, then in particular $F(\zeta) \subseteq K$ contains all the roots of $x^n - 1$, and thus $F(\zeta)$ must be the splitting field of $x^n - 1$.

Note also that $\mu_n(K)$ is a cyclic group of finite order, say $\mu_n(K) \cong \mathbb{Z}/d$. Moreover, ζ is a generator of $\mu_n(K)$. By a result from Math 817, the complete list of primitive n th roots of unity is

$$\zeta^i \quad \text{such that } \gcd(i, d) = 1.$$

Example 16.44. When $F = \mathbb{Q}$, the element $e^{2\pi i/n} \in \overline{\mathbb{Q}}$ is a primitive n th root of unity. Moreover, the primitive n th roots of unity over \mathbb{Q} are precisely the elements of the form $e^{2\pi ij/n}$ with for any j with $\gcd(n, j) = 1$.

Remark 16.45. Note that if $\text{char}(F) \nmid n$, the polynomial $x^n - 1 \in F[x]$ is separable by ??, since its derivative is $nx^{n-1} \neq 0$ and hence $\gcd(nx^{n-1}, x^n - 1) = 1$. In this case, $|\mu_n(K)| = n$ and so

$$\mu_n(K) = \{1, \zeta, \zeta^2, \dots, \zeta^{n-1}\}.$$

However, if $\text{char}(F) \mid n$ then $\mu_n(K)$ can have fewer than n elements. For example, if F is any field of characteristic 2, the polynomial $x^2 - 1 = (x - 1)^2$ has a unique root over F : the unique second root of unity is 1.

Theorem 16.46. Let F be a field and n a positive integer with $\text{char}(F) \nmid n$, and let $\zeta \in \overline{F}$ be a primitive n th root of unity. The extension $F \subseteq F(\zeta)$ is Galois and $\text{Gal}(F(\zeta)/F)$ is isomorphic to a subgroup of $(\mathbb{Z}/n)^\times$. In particular, $\text{Gal}(F(\zeta)/F)$ is an abelian group.

Proof. By ??, $F(\zeta)$ is the splitting field of $x^n - 1$ over F . As observed above, the polynomial $x^n - 1$ is separable, and thus $F(\zeta)/F$ is Galois by ??.

Given any $\sigma \in \text{Gal}(F(\zeta)/F)$, by ?? $\sigma(\zeta)$ is also an n th root of unity. One can check this explicitly by noting that

$$\sigma(\zeta)^n = \sigma(\zeta^n) = \sigma(1) = 1.$$

Moreover, we claim that $\sigma(\zeta)$ must also be a primitive n th root of unity. Notice that since $1, \zeta, \zeta^2, \dots, \zeta^{n-1}$ are distinct, then the elements $1, \sigma(\zeta), \sigma(\zeta)^2, \dots, \sigma(\zeta)^{n-1}$ must also be distinct, since $\sigma(\zeta^l) = \sigma(\zeta)^l$ for all l .

This proves that $\sigma(\zeta) = \zeta^j$ for an integer j (unique modulo n) such that $\gcd(j, n) = 1$. Thus we have a well-defined function

$$\begin{aligned} \Phi: \text{Gal}(F(\zeta)/F) &\longrightarrow (\mathbb{Z}/n)^\times \\ \sigma &\longmapsto \Phi(\sigma) = j \quad \text{where } \sigma(\zeta) = \zeta^j. \end{aligned}$$

Given any other $\tau \in \text{Gal}(F(\zeta)/F)$, let $\tau(\zeta) = \zeta^i$. Then

$$(\tau \circ \sigma)(\zeta) = \tau(\zeta^j) = \tau(\zeta)^j = \zeta^{ij}.$$

This proves that $\Phi(\tau \circ \sigma) = \Phi(\tau) \cdot \Phi(\sigma)$, so Φ is a group homomorphism.

If $\Phi(\sigma) = 1$, then σ fixes ζ and hence must be the trivial automorphism. This shows that Φ is injective. \square

Corollary 16.47. For any $n > 1$, $\text{Gal}(\mathbb{Q}(e^{2\pi i/n})/\mathbb{Q}) \cong (\mathbb{Z}/n)^\times$.

Sketch of proof. Consider the injective group homomorphism

$$\begin{aligned} \Phi: \text{Gal}(F(\zeta)/F) &\longrightarrow (\mathbb{Z}/n)^\times \\ \sigma &\longmapsto \Phi(\sigma) = j \quad \text{where } j \text{ satisfies } \sigma(\zeta) = \zeta^j. \end{aligned}$$

we constructed in the proof of ???. We claim that Φ is an isomorphism.

To show that homomorphism must be surjective, one shows that the degree of the minimal polynomial of $e^{2\pi i/n}$ is precisely the number of elements of $(\mathbb{Z}/n)^\times$, and thus $|\text{Gal}(\mathbb{Q}(e^{2\pi i/n})/\mathbb{Q})| = |(\mathbb{Z}/n)^\times|$. We skip this detail. \square

We now cover the Galois groups of polynomials of the form $x^n - a$ in the case where the base field contains all the n th roots of unity.

Theorem 16.48. Let F be a field, $a \in F$, and consider a positive integer n such that F contains a primitive n th root of unity and $\text{char}(F) \nmid n$. Let L be the splitting field of $x^n - a$ over F . Then L/F is Galois and $\text{Gal}(L/F)$ is isomorphic to a subgroup of \mathbb{Z}/n , and hence it is cyclic.

Proof. If $a = 0$, then $L = F$ and $\text{Gal}(L/F)$ is the trivial group, so the result is trivially true. If $a \neq 0$, then

$$\gcd(x^n - a, nx^{n-1}) = 1,$$

and hence $x^n - a$ is separable by [??](#). By [??](#), we conclude that $F \subseteq L$ is Galois.

Let α be a root of $x^n - a$ in L , and let $\zeta \in F$ be a primitive n th root of unity. Then the roots of $x^n - a$ are $\zeta^j\alpha$ for $j = 0, \dots, n-1$, and $L = F(\alpha)$. Also, the elements $\zeta^j\alpha$ for $j = 0, \dots, n-1$ are all distinct, and thus for each $\sigma \in \text{Gal}(L/F)$ we have $\sigma(\alpha) = \zeta^j\alpha$ with j well-defined modulo n . Define

$$\begin{aligned} \Phi: \text{Gal}(L/F) &\longrightarrow \mathbb{Z}/n \\ \sigma &\longmapsto \Phi(\sigma) = j \quad \text{where } \sigma(\alpha) = \zeta^j\alpha. \end{aligned}$$

Notice that such integer j is unique modulo n . Let $\tau(\alpha) = \zeta^i\alpha$. Note that $\zeta \in F$ and hence that it is fixed by τ . Then

$$(\tau \circ \sigma)(\alpha) = \tau(\zeta^j\alpha) = \zeta^j\zeta^i\alpha = \zeta^{j+i}\alpha,$$

so Ψ is a group homomorphism. It is injective since $\Psi(\sigma) = 0$ implies that $\sigma(\alpha) = \zeta^0\alpha = \alpha$, so σ fixes α and hence all of L . \square

We are interested in understanding when we can write a formula for all the roots of a given polynomial using only the usual elementary operations and radicals. We formalize this idea as follows:

Definition 16.49. For a field F of characteristic 0, we say $f \in F[x]$ is **solvable by radicals** over F if there exists a finite chain of field extensions

$$F = F_0 \subseteq F_1 \subseteq F_2 \subseteq \dots \subseteq F_m$$

such that f splits completely in F_m , and for each i the field extension $F_i \subseteq F_{i+1}$ is the splitting field of an polynomial of the form $x^{n_i} - a_i$ for some positive integer n_i and some element $a_i \in F_i$.

Note that $a_i = 1$ is allowed here, so that some of the steps may involve adjoining n th roots of unity. Roughly speaking, f is solvable by radicals if each of its roots can be written by an expression involving sums, products, and iterated n th roots of elements of F . Granted, such an expression may perhaps be extremely complicated.

Example 16.50. The polynomial $f = ax^2 + bx + c \in \mathbb{Q}[x]$ is solvable by radicals over \mathbb{Q} since its roots are

$$\frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.$$

Explicitly, take F_1 to be the splitting field of $x^2 - (b^2 - 4ac)$; indeed, f splits completely in F_1 .

Example 16.51. The polynomial $f = x^4 + bx^2 + c \in \mathbb{Q}[x]$ is solvable by radicals over \mathbb{Q} since its roots are

$$\pm \sqrt{\frac{-b \pm \sqrt{b^2 - 4c}}{2}}.$$

Explicitly, we could set F_1 to be the splitting field of $x^2 - (b^2 - 4c)$ over \mathbb{Q} , F_2 to be the splitting field of $x^2 - \left(\frac{-b+\sqrt{b^2-4c}}{2}\right)$ over F_1 , and F_3 to be the splitting field of $x^2 - \left(\frac{-b-\sqrt{b^2-4c}}{2}\right)$ over F_2 . It's not clear if $F_3 = F_2$ or $F_3 \subsetneq F_2$, but either way the tower given shows that f is solvable by radicals.

The notion of solvable polynomial has a group theoretic counterpart.

Definition 16.52. A group G is **solvable** if there is a sequence of subgroups

$$\{e\} = N_0 \trianglelefteq N_1 \trianglelefteq \dots \trianglelefteq N_k = G$$

such that for all $0 \leq i \leq k-1$ we have $N_i \trianglelefteq N_{i+1} \leq G$ and the quotient groups N_{i+1}/N_i are abelian.

Example 16.53. One can show that every group G with $|G| < 60$ is solvable.

Remark 16.54. Suppose that G is a finite simple group. Recall that this means that G has no nontrivial normal subgroups. Then the only sequence of normal subgroups of G is

$$\{e\} \trianglelefteq G,$$

and thus G is solvable if and only if G is abelian. But the only simple abelian groups are $\mathbb{Z}/(p)$ for p a prime, so $G \cong \mathbb{Z}/(p)$ for some prime p .

Example 16.55. The groups A_5 and S_5 are not solvable. To see why, recall that A_5 is a finite simple group and it is not abelian, and thus by ?? we conclude that A_5 is not solvable. As for S_5 , its only nontrivial normal subgroup is A_5 , but both A_5 and S_5 are not abelian, so S_5 is not solvable.

Example 16.56. We claim that the group S_4 is solvable. To see that, consider the subgroup

$$V = \{e, (12)(34), (14)(23), (13)(24)\}$$

and the following sequence of subgroups:

$$\{e\} \trianglelefteq V \trianglelefteq A_4 \trianglelefteq S_4.$$

Since V has order 4 and any group of order 4 is abelian, then V is abelian. In fact, one can show that $V \cong \mathbb{Z}/2 \times \mathbb{Z}/2$. Moreover, the quotients S_4/A_4 and A_4/V have order 2, and thus are abelian.

It turns out that there is a close relationship between solvable groups and solvable polynomials. In what follows, $\text{char}(F) = 0$ is not a necessary assumption, but we included it to make both the statement and the proof simpler.

Theorem 16.57. *Assume F is a field of characteristic 0. If $f \in F[x]$ is solvable by radicals, then the Galois group of the splitting field of $f \in F[x]$ is a solvable group.*

Sketch of proof. For a suitable n , we may assume there is a tower

$$F = F_0 \subseteq F_1 \subseteq \cdots \subseteq F_m$$

such that

- The splitting field L of f satisfies $L \subseteq F_m$.
- The splitting field of $x^n - 1$ over F is F_1 .
- For each $i \geq 1$, F_{i+1} is the splitting field of $x^{d_i} - a_i \in F_i[x]$, where $a_i \in F_i$ and $d_i \mid n$.

Note that $d_i \mid n$ means that F_i contains all the d_i th roots of 1, and thus ?? applies to the extension F_{i+1}/F_i for each $i \geq 1$.

It turns out that there is an extension E such that

$$F = F_0 \subseteq F_1 \subseteq \cdots \subseteq F_m \subseteq E$$

and E/F is Galois with a chain of normal subgroup inclusions

$$\text{Gal}(E/F_m) \trianglelefteq \text{Gal}(E/F_{m-1}) \trianglelefteq \text{Gal}(E/F_{m-2}) \trianglelefteq \cdots \trianglelefteq \text{Gal}(E/F_1) \trianglelefteq \text{Gal}(E/F).$$

The key point is that by ?? and ??, the groups

$$\text{Gal}(F_{i+1}/F_i) \cong \text{Gal}(E/F_i)/\text{Gal}(E/F_{i+1}) \text{ for } i = 0, \dots, m-1$$

are all abelian. These properties imply that $\text{Gal}(E/F)$ is a solvable group, and in turn this implies that $\text{Gal}(L/F)$ is solvable. \square

In characteristic 0, the converse of ?? is also true: if the Galois group of f is solvable, then f is solvable by radicals.

Theorem 16.58. *Every polynomial $f \in \mathbb{Q}[x]$ of degree at most 4 is solvable by radicals.*

The main point is that if L is the splitting field of a polynomial of degree at most 4, then $\text{Gal}(L/\mathbb{Q}) \trianglelefteq S_4$, and that every subgroup of S_4 is solvable. Indeed, formulas for the roots of polynomials of degree 2, 3, and 4 have been known for hundreds of years, and they involve only sums, products, quotients, and square roots, cube roots, and fourth roots.

We can now prove a theorem that has fascinated mathematicians and nonmathematicians alike for many centuries: the fact that the general quintic cannot be solved. More precisely, there is no formula involving only radicals, sums, and products for the zeroes of a general polynomial of degree 5 with rational coefficients. The key point turns out to lie in group theory: S_5 is not a solvable group, and there are polynomials $f \in \mathbb{Q}[x]$ of degree 5 such that the Galois group of the splitting field of f over \mathbb{Q} is S_5 .

Theorem 16.59. *If $f \in \mathbb{Q}[x]$ is any degree 5 irreducible polynomial with exactly 3 real roots, then f is not solvable by radicals.*

Proof. Let L be the splitting field of f . By ??, it suffices to prove $\text{Gal}(L/\mathbb{Q})$ is not a solvable group. In fact, we show it is isomorphic to S_5 .

Let $\alpha_1, \alpha_2, \alpha_3$ be the three real roots of f and let α_4, α_5 the two complex ones. Note that $\overline{\alpha_4} = \alpha_5$. Using this ordering of the roots, we identify $\text{Gal}(L/\mathbb{Q})$ as a subgroup of S_5 , and will identify an element of $\text{Gal}(L/\mathbb{Q})$ sending α_i to α_j with a permutation sending i to j .

Let σ denote complex conjugation. Note that $\sigma \in \text{Gal}(L/\mathbb{Q})$, since σ preserves \mathbb{Q} and α_1, α_2 , and α_3 , and it switches α_4 and α_5 . Thus this element of $\text{Gal}(L/\mathbb{Q})$ corresponds to the transposition $(4\ 5) \in S_5$.

Since $[\mathbb{Q}(\alpha_1) : \mathbb{Q}] = 5$, by the Degree Formula $5 \mid [L : \mathbb{Q}]$. But the extension is Galois, so $5 \mid |\text{Gal}(L/\mathbb{Q})|$. Since 5 is prime, there is an element $\tau \in \text{Gal}(L/\mathbb{Q})$ of order 5 by Cauchy's Theorem. Such an element is necessarily a 5-cycle. The result follows since any 5-cycle and any transposition necessarily generate all of S_5 (exercise).

Finally, we claim that S_5 is not solvable. First, S_5 is not abelian, so the series

$$H_0 = \{e\} \leq S_5$$

does not work since the unique quotient is not abelian. Moreover, as proven in Math 817, the only nontrivial normal subgroup of S_5 is A_5 and A_5 has no nontrivial normal subgroups. Hence the only possible composition series for S_5 would be

$$H_0 = \{e\} \leq A_5 \leq S_5,$$

but in this series the quotient $A_5/\{e\} \cong A_5$ is not abelian. \square

Example 16.60. The polynomial $f(x) = x^5 - 4x + 2$ is not solvable by radicals over \mathbb{Q} . One can show it is irreducible in $\mathbb{Q}[x]$ by the usual combination of Eisenstein's Criterion and Gauss' Lemma. Moreover, we claim that this polynomial has exactly 3 distinct real roots. Unfortunately, we cannot show this directly by *presenting* such roots, exactly because f is *not* solvable by radicals. One could check this informally by graphing f and checking that indeed it crosses the x -axis three times, and noting that all irreducible polynomials over a field of characteristic zero (thus perfect) are separable. If we wanted to check this more carefully, we could use some elementary calculus: $f'(x) = 5x^4 - 4$ has precisely two roots and changes signs at these roots. It follows that f must have exactly 3 real roots. By ??, f is not solvable by radicals.

However, there *are* irreducible polynomials of degree 5 that are solvable by radicals.

Example 16.61. The polynomial $x^5 - 1 \in \mathbb{Q}[x]$ is solvable by radicals; indeed, its roots are all 5th roots of unity over \mathbb{Q} .

We can check this using Galois Theory. By ??, the splitting field L of $x^5 - 1$ over \mathbb{Q} satisfies $\text{Gal}(L/\mathbb{Q}) \cong (\mathbb{Z}/5\mathbb{Z})^\times$. In particular, $\text{Gal}(L/\mathbb{Q})$ is abelian and thus solvable.

More details on the other applications of Galois Theory we mentioned in the beginning of this section can often be found in any standard book on the subject.

16.5 The primitive element theorem

Let $F \subseteq L$ be a field extension. Recall that an element θ so that $L = F(\theta)$ is called a **primitive element** for the a **simple extension** $F \subseteq L$.

Lemma 16.62. *If L/F is a finite extension with F infinite, then $L = F(\theta)$ if and only if there are only finitely many subfields of L containing F .*

Proof. First we show if there are only finitely many subfields of L containing F then L is simple. It's sufficient to show $F(\alpha, \beta)$ is simple for any $\alpha, \beta \in L$ and then the statement about L will follow by induction on the dimension of L . Consider the intermediate fields $E_c = F(\alpha + c\beta)$ for $c \in F$. Since there are only finitely many intermediate subfields, but infinitely many $c \in F$ we have

$$F(\alpha + c\beta) = F(\alpha + c'\beta) =: E \text{ for some } c \neq c'.$$

Then

$$\alpha + c\beta - (\alpha + c'\beta) = (c - c')\beta \in E,$$

so $\beta \in E$ and similarly $\alpha \in E$, thus $E = F(\alpha + c\beta) = F(\alpha, \beta)$.

For the converse, suppose $L = F(\theta)$ is simple and let $f = m_{\theta, F}$. Let E be an intermediate field and $g(x) = m_{\theta, E}$. Then $g \mid f$ in $E[x]$, so g is an irreducible factor of f . Consider E' to be the field obtained by adjoining the coefficients of g to F . Since $g = m_{\theta, E} = m_{\theta, E'}$, we have

$$[F(\theta) : E] = [F(\theta) : E'] = \deg(g).$$

Since $E' \subseteq E$, the Degree Formula gives $E = E'$. So all intermediate fields are generated by the coefficients of the irreducible factors of f . \square

Definition 16.63. Let L/F be a finite separable extension. The **Galois closure** of L over F is the smallest (with respect to containment) Galois extension of F containing L , meaning

$$L^{\text{Gal}} = \bigcap_{\substack{F \subseteq K \text{ Galois} \\ F \subseteq L \subseteq K}} K.$$

Remark 16.64. We should check that every finite field extension has a Galois closure. For example, one can pick a basis $\{\beta_1, \dots, \beta_n\}$ for L over F and take K to be the splitting field of the product of the minimal polynomials of β_1, \dots, β_n . Then $L \subseteq K$ will be the splitting field of a separable polynomial, hence Galois. This shows that the set indexing the intersection above is not empty, so the Galois closure exists as defined.

Theorem 16.65 (Primitive Element Theorem). *If $F \subseteq L$ is a finite and separable extension, then L is simple over F , meaning $L = F(\theta)$ for some $\theta \in L$.*

Proof. If F is a finite field then so is L . Since L is finite, then (L^\times, \cdot) is a cyclic group by a homework problem. Let θ be a generator for this multiplicative group. Then $L = F(\theta)$.

Now suppose F is infinite. Let K be the Galois closure of L over F . Then $G = \text{Gal}(K/F)$ is finite and has finitely many subgroups, thus by the Galois correspondence there are finitely many subfields of K , hence also of L , containing F . By ?? it follows that $F \subseteq L$ is simple. \square

Corollary 16.66. *If $F \subseteq L$ is a finite extension with F perfect, then $L = F(\theta)$ for some $\theta \in L$.*

Proof. By definition, if F is a perfect field then every finite extension of F is separable. By the Primitive Element Theorem, F is simple over F . Also, recall that every field of characteristic zero is perfect, by ??.

In particular, if L/F is a finite extension of fields of characteristic zero, then L is simple over F . The most important special case of this, or at least the one we keep encountering, is that every finite extension of \mathbb{Q} is simple over \mathbb{Q} . On the other hand, our proof of the Primitive Element Theorem is not constructive, and it doesn't tell us how to find a primitive element for a given finite extension. We saw in ?? that $\mathbb{Q}(\sqrt{2}, \sqrt{3}) = \mathbb{Q}(\sqrt{2} + \sqrt{3})$; nevertheless, it is *not* always true that $\mathbb{Q}(\alpha, \beta) = \mathbb{Q}(\alpha + \beta)$, as we saw in ??.

Theorem 16.67. *Every finite Galois extension is simple.*

Proof. Let $F \subseteq L$ be a Galois extension. By ??, for every $\alpha \in L$ the minimal polynomial $m_{\alpha, F}$ is separable, so $F \subseteq L$ is separable. By the Primitive Element Theorem, L is simple over F . \square

For extensions of perfect fields of characteristic zero, this is just a very special case of ???. In this case, we can also prove that the extension is simple directly from ???: if $F \subseteq L$ is Galois, then $\text{Gal}(L/F)$ is finite, and thus it has finitely many subgroups, which by the Galois correspondence says that $F \subseteq L$ has finitely many intermediate fields. By ???, this implies that $F \subseteq L$ is simple.

To construct an example of a finite field extension that is not simple, we need an infinite field of prime characteristic that is not perfect.

Example 16.68. Let $L = \mathbb{Z}/p(s, t)$ be the fraction field of the polynomial ring in two variables $\mathbb{Z}/p[s, t]$, and consider the subfield $K = \mathbb{Z}/p(s, t)$. We claim that this is a finite extension that is not simple. First, note that

$$\{s^i t^j \mid 0 \leq i, j \leq p-1\}$$

is a basis for L over K , so $[L : K] = p^2$. Now let $\alpha \in L$, meaning a rational function

$$\alpha = \frac{f(s, t)}{g(s, t)}$$

for some polynomials $f, g \in \mathbb{Z}/p[s, t]$. For any $a \in \mathbb{Z}/p$ we have $a^p = a$ by Fermat's Little Theorem, so by the Freshman's Dream we have $f(s, t)^p = f(s^p, t^p)$ and $g(s, t)^p = g(s^p, t^p)$. Therefore,

$$\alpha^p = \frac{f(s, t)^p}{g(s, t)^p} = \frac{f(s^p, t^p)}{g(s^p, t^p)} \in K.$$

Thus $x^p - \alpha^p \in K[x]$, and since α is a root of this polynomial we conclude that

$$[K(\alpha) : K] \leq \deg(x^p - \alpha^p) = p.$$

Thus $K(\alpha) \neq L$ for all $\alpha \in L$.

However, not every finite separable extension is Galois.

Example 16.69. We showed in ?? that $\mathbb{Q} \subseteq \mathbb{Q}(\sqrt[3]{2})$ is not Galois. However, the minimal polynomial of $\sqrt[3]{2}$ over \mathbb{Q} is $x^3 - 2$, which is separable, and thus by ?? the extension $\mathbb{Q} \subseteq \mathbb{Q}(\sqrt[3]{2})$ is separable.

16.6 The proof of Artin's Theorem

We now embark on a proof of Artin's Theorem. A key ingredient is the “linear independence of characters”, which is useful in other contexts as well, such as representation theory.

Definition 16.70. For a group G and field F , a **character** of G with values in F is a group homomorphism of the form

$$\chi : G \rightarrow F^\times.$$

Example 16.71.

- 1) If $G = C_n$, cyclic of order n , with generator x , then the UMP for cyclic groups says there is a unique group homomorphism $G \rightarrow \mathbb{C}^\times$ sending $x \mapsto \zeta_n$, and hence $x^i \mapsto \zeta_n^i$. This is an example of a character.
- 2) If K and F are two fields and $\phi : K \rightarrow F$ is a field map, then ϕ restricts to a character $\phi' : K^\times \rightarrow F^\times$.

Note that the set $\text{Fun}(G, F)$ of all functions from G to F is an F -vector space and that the characters of G are elements of this vector space. Therefore it makes sense to talk about linear independence for sets of characters. A point to observe here is that arbitrary linear combinations $\sum_i l_i \chi_i$ are not, in general, group homomorphisms.

Definition 16.72. For G and F and characters χ_1, \dots, χ_n , we say these characters are *linear independent* if whenever $\sum_{i=1}^n l_i \chi_i = 0$ (the constant map 0), we must have $l_i = 0$ for all i . Making this even more explicit: χ_1, \dots, χ_n , are linear independent if given $l_i \in F$ such that $\sum_{i=1}^n l_i \chi_i(g) = 0$ for all $g \in G$, we must have $l_i = 0$ for all i .

Theorem 16.73 (Linear Independence of Characters). *Let G be a group, F be a field, and let $\chi_j : G \rightarrow F^\times$ for $j = 1, \dots, m$ be a finite list of distinct characters, meaning that for all $i \neq j$ we have $\chi_i(g) \neq \chi_j(g)$ for at least one $g \in G$. Then χ_1, \dots, χ_m are linearly independent.*

The Theorem is sort of a *Sophomore's dream*, since it is saying that if a list of a certain sort of vectors in a certain vector space has no repetitions, then the vectors are linearly independent.

Proof. We proceed by induction on m .

Base case: When $m = 1$, since $\chi_1(g) \neq 0$ for all g then $l_1 \chi_1 = 0$ iff $l_1 = 0$.

Induction Step: Suppose $m > 1$ and that $\sum_{i=1}^m l_i \chi_i(g) = 0$ for all $g \in G$ for some $l_i \in F$.

Suppose

$$\sum_{i=1}^m l_i \chi_i = 0. \quad (16.6.1)$$

Evaluating (??) at hg for $g, h \in G$ and using that χ_i are group homomorphisms gives

$$0 = \sum_{i=1}^m l_i \chi_i(hg) = \sum_{i=1}^m l_i \chi_i(h) \chi_i(g) \quad \text{for all } g, h \in G. \quad (16.6.2)$$

Multiplying (??) by $\chi_1(h)$ gives

$$0 = \chi_1(h) \left(\sum_{i=1}^m l_i \chi_i(g) \right) \quad \text{for all } g, h \in G. \quad (16.6.3)$$

Subtracting (??) from (??) we get we get

$$0 = \chi_1(h) \left(\sum_{i=1}^m l_i \chi_i(g) \right) - \sum_{i=1}^m l_i \chi_i(h) \chi_i(g) = \sum_{i=2}^m (\chi_1(h) l_i - \chi_i(h) l_i) \chi_i(g) \quad \text{for all } g, h \in G.$$

Fixing h , the equation above gives a linear dependence between χ_2, \dots, χ_m . Using the induction hypothesis we conclude that

$$\chi_1(h) l_i - \chi_i(h) l_i = 0 \quad \text{for all } h \in G$$

for all i , including $i = m$. Since $\chi_1(h) \neq \chi_m(h)$, we get $l_m = 0$, and hence (??) reduces to

$$\sum_{i=1}^{m-1} l_i \chi_i(g) = 0, \quad \text{for all } g \in G.$$

Using the induction hypothesis again it follows that $l_i = 0$ for all i . □

Example 16.74. Let $G = C_n$, generated by x , and define

$$\begin{aligned} \chi_j: G &\longrightarrow \mathbb{C} \\ x &\longmapsto \chi_j(x) = \zeta_n^j = e^{2\pi j/n} \end{aligned}$$

for $j = 0, \dots, n-1$ by $\chi_i(x) = \zeta_n^i = e^{2\pi i/n}$. These are distinct, and hence by ?? they must be linearly independent.

We now restate Artin's theorem:

Theorem (Artin's Theorem). *Let L be any field and G any finite subgroup of $\text{Aut}(L)$. Then L^G is a subfield of L , L/L^G is a finite Galois extension and $\text{Gal}(L/L^G) = G$.*

Proof of Artin's Theorem. Let G be a finite subgroup of $\text{Aut}(L)$ for a field L . In ??, we left proving that L^G is a subfield of L as an exercise. We will prove the remaining statements. We need to prove L/L^G is a finite extension and that $[L : L^G] = |\text{Aut}(L/L^G)|$.

We start by showing that it suffices to show that $[L : L^G] = |G|$. If $[L : L^G] = |G|$ does indeed hold, then in particular L/L^G is a finite extension. By Theorem ??, since $L^G \subseteq L$ is finite then $|\text{Aut}(L/L^G)| \leq [L : L^G]$. Since $[L : L^G] = |G|$, then we obtain $|\text{Aut}(L/L^G)| \leq |G|$.

On the other hand, any element in G fixes L^G by definition, so $G \leq \text{Aut}(L/L^G)$. But $|\text{Aut}(L/L^G)| \leq |G|$ and these are both finite groups, so $G = \text{Aut}(L/L^G)$. From the inequality before we now conclude that $|\text{Aut}(L/L^G)| = [L : L^G]$, and thus the extension is Galois. Finally, this gives $\text{Gal}(L/L^G) = \text{Aut}(L/L^G) = G$.

It remains to prove that $[L : L^G] = |G|$. Let $n = |G|$ and $G = \{\sigma_1, \dots, \sigma_n\}$ with $\sigma_1 = \text{id}_L$. By ??, we know that $[L : L^G] \geq n$. We want to show that equality holds. If $[L : L^G] > n$, then we can find $n+1$ many L^G -linearly independent elements $\omega_1, \dots, \omega_{n+1}$ in L . Consider the system of n equations with $n+1$ unknowns

$$\left\{ \begin{array}{l} \sigma_1(\omega_1)x_1 + \dots + \sigma_1(\omega_{n+1})x_{n+1} = 0 \\ \sigma_2(\omega_1)x_1 + \dots + \sigma_2(\omega_{n+1})x_{n+1} = 0 \\ \vdots \\ \sigma_n(\omega_1)x_1 + \dots + \sigma_n(\omega_{n+1})x_{n+1} = 0. \end{array} \right.$$

Since there are fewer equations than unknowns, this system has a nontrivial solution. Among these, choose the solution that has the least number r of nonzero components; by reordering the ω_i we may assume this solution has the form $(a_1, \dots, a_r, 0, \dots, 0)$ with $a_i \neq 0$ for all i . By scaling, we may assume $a_r = 1$. Since $\sigma_1 = \text{id}_G$, the first equation says that

$$a_1\omega_1 + \dots + a_{r-1}\omega_{r-1} + \omega_r = 0.$$

If all the a_i belong to L^G then this equation of linear dependence would contradict the linear independence of $\omega_1, \dots, \omega_{n+1}$. Thus $a_i \notin L^G$ for some i . Reordering again, we may assume $a_1 \notin L^G$. Since $a_r = 1$, note in particular that this shows $r > 1$. We thus have

$$\left\{ \begin{array}{l} \sigma_1(\omega_1)a_1 + \dots + \sigma_1(\omega_{r-1})a_{r-1} + \sigma_1(\omega_r) = 0 \\ \sigma_2(\omega_1)a_1 + \dots + \sigma_2(\omega_{r-1})a_{r-1} + \sigma_2(\omega_r) = 0 \\ \vdots \\ \sigma_n(\omega_1)a_1 + \dots + \sigma_n(\omega_{r-1})a_{r-1} + \sigma_n(\omega_r) = 0 \end{array} \right.$$

Now, since $a_1 \notin L^G$, there is a k with $\sigma_k(a_1) \neq a_1$. Apply σ_k to the j th row to obtain

$$\sigma_k\sigma_j(\omega_1)\sigma_k(a_1) + \dots + \sigma_k\sigma_j(\omega_{r-1})\sigma_k(a_{r-1}) + \sigma_k\sigma_j(\omega_r) = 0$$

Since G is a group, as j ranges over all possibilities, $\sigma_k\sigma_j$ ranges over all elements of G . Thus

$$\sigma_i(\omega_1)\sigma_k(a_1) + \dots + \sigma_i(\omega_n)\sigma_k(a_{r-1}) + \sigma_i(\omega_r) = 0 \quad \text{for all } 1 \leq i \leq n.$$

For each i , subtracting this equation from the i th equation in the previous system yields

$$\sigma_i(\omega_1)(a_1 - \sigma_k(a_1)) + \dots + \sigma_i(\omega_{r-1})(a_{r-1} - \sigma_k(a_{r-1})) = 0 \quad \text{for all } 1 \leq i \leq n.$$

Since $a_1 - \sigma_k(a_1) \neq 0$, this is a nontrivial solution to original system of equations with fewer than r nonzero components, which is a contradiction. \square

Index

- $(-)^{\sigma}$, 222
- (S) , 108
- $A + B$, 144
- A_n , 44
- $C_G(a)$, 63
- C_{∞} , 37
- C_n , 37
- D_n , 12
- EF , 230
- $F(A)$, 202
- $F(a)$, 201
- $F(a_1, \dots, a_n)$, 202
- G' , 47
- G/\sim , 39
- G^{ab} , 47
- $H \leq G$, 27
- $H < G$, 27
- HK , 50
- Hg , 39
- IM , 144
- L^{Gal} , 242
- $M \cong N$, 146
- $N \trianglelefteq G$, 42
- $N_G(a)$, 63
- Q_8 , 17
- R -algebra, 147
- R -module, 140
- R -module homomorphism, 145
- R -module isomorphism, 146
- R -module presented by A , 172
- R -submodule, 143
- R^{\times} , 102
- $[G : H]$, 41
- $[G, G]$, 47
- $[L : F]$, 200
- $[f]_B^C$, 162
- $[g, h]$, 47
- $[n]$, 6
- $\text{Gal}(L/F)$, 227
- $\text{char}(R)$, 217
- $\ker(f)$, 20, 146
- $\text{Aut}(G)$, 18
- $\text{Aut}(K)$, 222
- $\text{Aut}(K/F)$, 222
- $\text{Mat}_n(R)$, 100
- $\text{Orb}_G(s)$, 24, 221
- $\text{Syl}_p(G)$, 76
- $\text{Z}(R)$, 105
- $\text{im}(f)$, 146
- $\mu_n(K)$, 236
- \sim_G , 221
- \sim_H , 39
- $b \mid a$, 125
- f^{-1} (for a homomorphism f), 29
- gH , 39
- m -cycle, 6
- n_p , 76
- p -subgroup, 76
- abelian group, 4
- abelianization, 47
- action, 23, 221
- action by conjugation, 26
- action of a group on a set, 23
- action via automorphisms, 88
- algebraic, 205
- algebraic closure, 209
- algebraic element, 205
- algebraic extension, 207
- algebraically closed, 209
- alternating group, 44

- associates, 125
 automorphism, 18
 automorphism group of a field, 222
 automorphism group of a field extension, 222
 basis, 152
 binary operation, 2
 c, 26
 cancellation rule, 103
 canonical (quotient) map, 46
 canonical map, 113, 149
 canonical projection, 46
 canonical quotient map, 149
 canonical surjection, 46, 113
 Cayley's Theorem, 30
 center of a group, 5
 center of a ring, 105
 central element, 105
 centralizer, 63
 change of basis matrix, 165
 character, 244
 characteristic, 217
 characteristic polynomial, 191
 commutative ring, 99
 commutator, 47
 commutator subgroup, 47
 companion matrix, 189
 compatible with multiplication (for an equivalence relation), 38
 composite of two fields, 230
 conjugacy class, 62
 conjugate elements, 61
 conjugate subgroups, 72
 conjugation action, 26
 cycle, 6
 cycle type, 9
 cyclic, 144, 151
 cyclic group, 5, 33
 cyclic group of order n , 37
 cyclic subgroup generated by an element, 29
 degree of a field extension, 200
 degree of a polynomial, 122
 derivative, 218
 derived subgroup, 47
 diagonalizable, 197
 dihedral group, 12
 dimension, 158
 direct product, 155
 direct product (of groups), 82
 direct sum, 155
 direct sum (of groups), 82
 direct sum of matrices, 189
 distributivity, 98
 division ring, 99
 divisor, 125
 domain, 103
 eigenvalue, 193
 eigenvector, 193
 elementary basis change operation, 166
 elementary column operation, 166
 elementary divisor decomposition, 93
 elementary divisors, 93, 181
 elementary matrix, 167
 endomorphism ring, 146
 endomorphisms, 146
 Euclidean domain, 121
 Euclidean function, 121
 Euler φ function, 88
 even permutation, 11
 expansion of an ideal, 112
 external direct product, 84
 faithful action, 25, 221
 field, 99
 field extension, 200
 field of algebraic numbers, 212
 finite dimensional, 157
 finite field extension, 200, 207
 finitely generated, 151
 finitely generated group, 5
 finitely generated ideal, 109
 First Isomorphism Theorem, 49
 fixed point, 58
 free, 152
 free group, 55
 free module, 143

- free rank, 179
 Freshman's Dream, 217
 Galois closure, 242
 Galois correspondence, 233
 Galois extension, 227
 Galois group, 227
 Galois group of a Galois extension, 227
 Galois group of a polynomial, 227
 Gaussian integers, 105
 gcd, 125
 generated by, 144, 151
 generators (of an ideal), 109
 generators for a group, 5
 greatest common divisor, 125
 group, 2
 group action, 221
 group action via automorphisms, 88
 group homomorphism, 18
 group isomorphism, 18
 homomorphism (of groups), 18
 ideal, 106
 ideal generated by, 108
 idempotent element, 104
 identity, 2
 identity element, 2
 image, 20, 146
 image of a homomorphism, 146
 index, 41
 infimum, 35
 infinite cyclic group, 37
 infinite dihedral group, 45
 inseparable, 219
 inseparable extension, 219
 integral domain, 103
 intermediate field, 230
 internal direct product, 84
 internal semidirect product, 91
 invariant factor decomposition, 93
 invariant factors, 93, 179, 188
 inverse, 2, 102
 irreducible element, 126
 isometry, 12
 isomorphic, 146
 isomorphic groups, 18
 isomorphic modules, 146
 isomorphism, 18
 isomorphism (of groups), 18
 isomorphism invariant, 21
 Jordan block, 195
 Jordan canonical form, 196
 kernel, 20, 110, 146
 kernel of a group homomorphism, 20
 kernel of a homomorphism, 146
 Lagrange's Theorem, 31
 lattice, 35
 lattice isomorphism, 36
 lcm, 125
 leading coefficient, 122
 least common multiple, 125
 left R -module, 140
 left action of a subgroup, 39
 left coset, 39
 left ideal, 107
 left inverse, 3
 left regular action, 26
 length of a cycle, 6
 linear combination, 151
 linear transformation, 145
 linearly dependent, 152
 linearly independent, 152
 lower bound, 35
 Main Theorem of Sylow Theory, 77
 matrix of the linear transformation, 162
 matrix ring, 100
 maximal ideal, 119
 minimal polynomial, 191, 206
 module of relations, 173
 monoid, 3
 monomials, 102
 multiple, 125
 multiplicity, 218
 multiplicity of a root, 218
 nilpotent element, 104
 noncommutative ring, 99

- nontrivial subgroup, 27
- norm, 100
- norm function, 121
- normal subgroup, 42
- normalizer, 63
- nullspace, 161
- orbit (of an action), 24
- Orbit Formula, 59
- Orbit-Stabilizer Theorem, 59
- orbits of a group action, 221
- order of a group, 2
- order relation, 35
- parity of a permutation, 11
- partially ordered set, 35
- perfect field, 219
- permutation group of a set X , 6
- permutation on n symbols, 6
- permutation representation, 24
- PID, 124
- polynomial ring, 102
- poset, 35
- power set, 35
- preimage of a homomorphism, 29
- presentation (of a group), 55
- presentation of a group, 5
- prime element, 126
- prime field, 217
- prime ideal, 119
- primitive n th root of unity, 236
- primitive element, 201, 242
- principal ideal, 109
- principal ideal domain, 124
- proper ideal, 107
- quaternion group, 17
- quaternion ring, 100
- quotient group, 39, 45
- quotient map, 113
- quotient ring, 113
- rank, 93, 154, 161
- rank of a group, 93
- rational canonical form, 190
- reflections of D_n , 13
- relations for a group, 5
- represent, 162
- restriction of scalars, 144
- right R -module, 140
- right coset, 39
- right ideal, 107
- right inverse, 3
- ring, 98
- ring homomorphism, 110
- ring isomorphism, 111
- ring map, 110
- ring of scalars, 144
- rng, 98
- rotations of D_n , 13
- Second Isomorphism Theorem, 51
- semidirect product, 85
- semigroup, 3
- separable extension, 219
- separable polynomial, 218
- similar, 166
- simple extension, 242
- simple field extension, 201
- simple group, 69
- simple ring, 107
- solvable by radicals, 238
- solvable group, 239
- span, 156
- spanned by, 151
- special linear group, 29
- splitting field, 213
- stabilizer, 58
- subfield, 104
- subfield generated by A over F , 202
- subfield of L fixed by G , 227
- subgroup, 27
- subgroup generated by a set, 29
- submodule generated by, 144
- subring, 104
- sum of modules, 144
- supremum, 35
- Sylow p -subgroup, 76
- symmetry, 12
- transcendental, 205

- transcendental element, 205
- transitive action, 25
- transitive group action, 221
- transposition, 7
- trivial action, 26
- trivial center, 5
- trivial group, 4
- trivial homomorphism, 18
- trivial subgroups, 27
- trivial submodules, 143
- two sided ideal, 106
- UFD, 128
- unique factorization domain, 128
- unit, 102
- unital ring, 98
- upper bound, 35
- vector of B -coordinates, 164
- vector space, 141
- zero module, 143
- zero ring, 98
- zerodivisor, 102
- Zorn's Lemma, 120