Introduction to Modern Algebra II

Math 818 Spring 2023

April 11, 2023

Contents

Chapter 1

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.

1.1 Basic assumptions

In this class, all rings have a multiplicative identity, written as 1 or 1_R is we want to emphasize that we are referring to the ring R. This is what some authors call *unital rings*; since for us all rings are unital, we will omit the adjective. Moreover, we will think of 1 as part of the structure of the ring, and thus require it be preserved by all natural constructions. As such, a subring S of R must share the same multiplicative identity with R, meaning $1_R = 1_S$. Moreover, any ring homomorphism must preserve the multiplicative identity. To clear any possible confusion, we include below the relevant definitions.

Definition 1.1. A **ring** is a set R equipped with two binary operations, $+$ and $\cdot$, satisfying:

(1) $(R, +)$ is an abelian group with identity element denoted 0 or 0_R.

(2) The operation $\cdot$ is associative, so that $(R, \cdot)$ is a semigroup.

(3) For all $a, b, c \in R$, we have

$$a \cdot (b + c) = a \cdot b + a \cdot c \quad \text{and} \quad (a + b) \cdot c = a \cdot c + b \cdot c.$$

(4) there is a multiplicative identity, written as 1 or 1_R, such that $1 \cdot a = a = a \cdot 1$ for all $a \in R$.

To simplify notation, we will often drop the $\cdot$ when writing the multiplication of two elements, so that ab will mean $a \cdot b$.

Definition 1.2. A ring R is a **commutative ring** if for all $a, b \in R$ we have $a \cdot b = b \cdot a$.

Definition 1.3. A ring R is a **division ring** if $1 \neq 0$ and $R \setminus \{0\}$ is a group under $\cdot$, so every nonzero $r \in R$ has a multiplicative inverse. A **field** is a commutative division ring.

Definition 1.4. A commutative ring R is a **domain**, sometimes called an **integral domain** if it has no zerodivisors: $ab = 0 \Rightarrow a = 0$ or $b = 0$.

For some familiar examples, $\mathrm{M}_n(R)$ (the set of $n \times n$ matrices) is a ring with the usual addition and multiplication of matrices, $\mathbb{Z}$ and $\mathbb{Z}/n$ are commutative rings, $\mathbb{C}$ and $\mathbb{Q}$ are fields, and the real Hamiltonian quaternion ring $\mathbb{H}$ is a division ring.

Definition 1.5. A **ring homomorphism** is a function $f\colon R \to S$ satisfying the following:

- $f(a+b) = f(a) + f(b)$ for all $a, b \in R$.
- $f(ab) = f(a)f(b)$ for all $a, b \in R$.
- $f(1_R) = 1_S$.

Under this definition, the map $f\colon \mathbb{R} \to \mathrm{M}_2(\mathbb{R})$ sending $a \mapsto \begin{bmatrix} a & 0 \\ 0 & 0 \end{bmatrix}$ preserves addition and multiplication but not the multiplicative identities, and thus it is not a ring homomorphism.

Exercise 1. For any ring R, there exists a unique homomorphism $\mathbb{Z} \to R$.

Definition 1.6. A subset S of a ring R is a **subring** of R if it is a ring under the same addition and multiplication operations and $1_R = 1_S$.

So under this definition, $2\mathbb{Z}$, the set of even integers, is not a subring of $\mathbb{Z}$; in fact, it is not even a ring, since it does not have a multiplicative identity!

Definition 1.7. Let R be a ring. A subset I of R is an **ideal** if:

- I is nonempty.
- $(I, +)$ is a subgroup of $(R, +)$.
- For every $a \in I$ and every $r \in R$, we have $ra \in I$ and $ar \in I$.

The final property is often called **absorption**. A **left ideal** satisfies only absorption on the left, meaning that we require only that $ra \in I$ for all $r \in R$ and $a \in I$. Similarly, a **right ideal** satisfies only absorption on the right, meaning that $ar \in I$ for all $r \in R$ and $a \in I$.

When R is a commutative ring, the left ideals, right ideals, and ideals over R are all the same. However, if R is not commutative, then these can be very different classes.

One key distinction between unital rings and nonunital rings is that if one requires every ring to have a 1, as we do, then the ideals and subrings of a ring R are very different creatures. In fact, the *only* subring of R that is also an ideal is R itself. The change lies in what constitutes a subring; notice that nothing has changed in the definition of ideal.

Remark 1.8. Every ring R has two **trivial ideals**: R itself and the zero ideal $(0) = \{0\}$.

A **nontrivial ideal** I of R is an ideal that $I \neq R$ and $I \neq (0)$. An ideal I of R is a **proper ideal** if $I \neq R$.

1.2 Modules: definition and examples

Definition 1.9. Let R be a ring with $1 \neq 0$. A **left R-module** is an abelian group $(M, +)$ together with an action $R \times M \to M$ 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 $M \times R \to M, (m, r) \mapsto mr$, 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 1.10. 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 1.11 (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 $(-1_R)m = -m$. □

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 1.12. 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 1.13. *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:

$$\begin{cases} [i]_n \cdot m = (\underbrace{[1]_n + \cdots + [1]_n}_{i}) \cdot m = \underbrace{[1]_n \cdot m + \cdots + [1]_n \cdot m}_{i} = \underbrace{m + \cdots + m}_{i} & \text{if } i > 0 \\ 0 \cdot m = 0_M & \\ [i]_n \cdot m = -(-[i]_n) \cdot m = -(\underbrace{m + \cdots + m}_{-i}) & \text{if } i < 0. \end{cases}$$

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. □

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 $\ell = 0, m \geqslant 1$ and $d_i \mid n$ for $1 \leqslant i \leqslant m$ 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 1.14. (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 $R[x_1, \ldots, x_n]$ is an R-module for any $n \geqslant 1$. This is a special case of (4).

(6) If R is a commutative ring and G is a group, then $R[G]$ is an R-module. This is a special case of (4).

(7) If R is a commutative ring, let $\mathrm{M}_n(R)$ denote set of $n \times n$ matrices with entries in R. Then $\mathrm{M}_n(R)$ is an R-module for $n \geqslant 1$, with the R-action given by multiplying all the entries of the given matrix by the given element of R.

(8) The **free module** over R of rank n is the set

$$R^n = \left\{ \begin{bmatrix} r_1 \\ \vdots \\ r_n \end{bmatrix} \mid r_i \in R, 1 \leqslant i \leqslant n \right\}$$

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 $(r_1, \ldots, r_n)$ instead. Notice that R is the free R-module of rank 1.

(9) 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 $r(a_i) := (ra_i)$.

1.3 Submodules and restriction of scalars

Definition 1.15. Let R be a ring and let M be a left R-module. An **R-submodule** of M is a subgroup N 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 2. 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 1.16. 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 1.17 (One-step test for submodules)**.** *Let R be a ring with $1 \neq 0$ and let M be a left R-module. A nonempty subset N of M is an R-submodule of M if and only if $rn + n' \in N$ for all $r \in R$ and $n, n' \in N$.*

Proof. The One-step Test for subgroups says that if for all $n, n' \in N$ we have $n' - n \in N$, then N is a subgroup of M. By Lemma 1.11, by taking $r = -1$ we get $rn + n' = n' - n$, and by assumption this is an element of N. Therefore, N is a subgroup of M. As a consequence, $0_M \in N$. By taking $n' = 0_M$, we see that for all $n \in N$ and all $r \in R$ we have $rn = rn + n' \in N$, and thus we can now conclude that N is a submodule of M. □

Example 1.18. 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 3. 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 4. 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

$$IM := \left\{ \sum_{k=1}^{n} j_k m_k \mid n \geqslant 0, j_k \in I, m_k \in M \text{ for } 1 \leqslant k \leqslant n \right\}$$

is a submodule of M.

Example 1.19. 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 1.13. The submodules of a $\mathbb{Z}$-module M coincide with the subgroups of the abelian group M.

Definition 1.20. Let R be a ring with $1 \neq 0$ and let M be an R-module. Given elements $m_1, \ldots, m_n \in M$, the **submodule generated by** $m_1, \ldots, m_n$ is the subset of M given by

$$Rm_1 + \cdots + Rm_n := \{r_1 m_1 + \cdots + r_n m_n \mid r_1, \ldots, r_n \in R\}.$$

Exercise 5. Let R be a ring with $1 \neq 0$ and M be an R-module. Given $m_1, \ldots, m_n \in M$, the submodule generated by $m_1, \ldots, m_n$ is indeed a submodule of M. Moreover, this is the smallest submodule of M that contains $m_1, \ldots, m_n$, meaning that every submodule of M containing $m_1, \ldots, m_n$ must also contain $Rm_1 + \cdots + Rm_n$.

Definition 1.21. 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 often 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 1.22 (Restriction of scalars)**.** *Let $\phi\colon R \to S$ 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 1.23. If I is an ideal of a ring R, applying restriction of scalars along the quotient homomorphism $q\colon R \to R/I$ 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 1.24. Given any ring R there exists a unique ring homomorphism $\mathbb{Z} \to R$, by Exercise 1. 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*[1] – actually coincide.

The next example explains why restriction of scalars is called a *restriction*.

Example 1.25. Let R be a subring of S, and let $i\colon R \to S$ 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.

[1] This is a concrete abstract nonsense construction that we will discuss in Homological Algebra next Fall.

1.4 Module homomorphisms and isomorphisms

Definition 1.26. Given R-modules M and N, an **R-module homomorphism** from M to N is a function $f\colon M \to N$ 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 1.27. 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 1.28. Let M and N be vector spaces over a field F. A **linear transformation** from M to N is an F-module homomorphism $M \to N$.

Example 1.29. Let R be a commutative ring and M be an R-module. For each $r \in R$, the multiplication map $\mu_r : M \to M$ 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

$$\mu_r(sm) = r(sm) = (rs)m = (sr)m = s(rm) = s\mu_r(m).$$

Definition 1.30. An R-module homomorphism $h\colon M \to N$ is an **R-module isomorphism** if there is an R-module homomorphism $g : N \to M$ such that $h \circ g = \mathrm{id}_N$ and $g \circ h = \mathrm{id}_M$. We say M and N are **isomorphic**, denoted $M \cong N$, if there exists an isomorphism $M \to N$.

To check that an R-module homomorphism $f\colon M \to N$ is an isomorphism, it is sufficient to check that it is bijective.

Exercise 6. Let $f\colon M \to N$ be a homomorphism of R-modules. Show that if f is bijective, then its set-theoretic inverse $f^{-1}\colon N \to M$ 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 1.31. Let $f\colon M \to N$ 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 $\mathrm{im}(f)$ or $f(M)$, is

$$\mathrm{im}(f) := \{f(m) \mid m \in M\}.$$

Exercise 7. 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 $i\colon N \to M$ is an R-module homomorphism.

Exercise 8. If $f\colon M \to N$ is an R-module homomorphism, then $\ker(h)$ is an R-submodule of M and $\mathrm{im}(f)$ is an R-submodule of N.

Definition 1.32. Let R be a ring and let M and N be R-modules. Then $\mathrm{Hom}_R(M,N)$ denotes the set of all R-module homomorphisms from M to N, and $\mathrm{End}_R(M)$ denotes the set $\mathrm{Hom}_R(M,M)$. We call $\mathrm{End}(M)$ the **endomorphism ring** of M, and elements of $\mathrm{End}(M)$ 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 $\mathrm{Hom}_R(M,N)$ is not a ring.

When R is commutative, $\mathrm{Hom}_R(M,N)$ is, however, an R-module; let us describe its R-module structure. Given $f, g \in \mathrm{Hom}_R(M,N)$, $f+g$ is the map defined by

$$(f+g)(m) := f(m) + g(m),$$

and given $r \in R$ and $f \in \mathrm{Hom}_R(M,N)$, $r \cdot f$ is the R-module homomorphism defined by

$$(r \cdot f)(m) := r \cdot f(m) = f(rm).$$

The zero element of $\mathrm{Hom}_R(M,N)$ is the **zero map**, the constant equal to 0_N.

Lemma 1.33. *Let M and N be R-modules over a commutative ring R. Then the addition and multiplication by scalars defined above make $\mathrm{Hom}_R(M,N)$ an R-module.*

Proof. There are many things to check, including:

- The addition and the R-action are both well-defined: given $f, g \in \mathrm{Hom}_R(M,N)$ and $r \in R$, we always have $f+g, rf \in \mathrm{Hom}_R(M,N)$.
- The axioms of an R-module are satisfied for $\mathrm{Hom}_R(M,N)$.

We leave the details as exercises. □

We will see later that for an n-dimensional vector space V over a field F, there is an isomorphism of vector spaces $\mathrm{End}_F(V) \cong M_n(F)$. This says that every linear transformation $T : V \to V$ corresponds to some $n \times n$ matrix. However, the story for general R-modules is a lot more complicated.

Lemma 1.34. *For any commutative ring R with $1 \neq 0$ and any R-module M there is an isomorphism of R-modules $\mathrm{Hom}_R(R,M) \cong M$.*

Before we write a formal proof, it helps to think about *why* this theorem is true. What does it mean to give an R-module homomorphism $f : R \to M$? More precisely, what information do we need to determine such an f? Do we need to be given the values of $f(r)$ for every $r \in R$? Since f is a homomorphism of R-modules, for any $r \in R$ we have

$$f(r) = f(r \cdot 1) = rf(1),$$

so the value of $f(1)$ *completely determines* which R-module homomorphism we are talking about. On the other hand, we can choose *any* $m \in M$ to be the image of 1, since thanks to the axioms for modules, the function

$$f(r) := rm$$

is a well-defined R-module homomorphism for any $m \in M$. In summary, to give an R-module homomorphism $R \to M$ is the same as choosing an element $m \in M$, and $\mathrm{Hom}_R(R, M) \cong M$.

Proof. Let $f\colon M \to \mathrm{Hom}_R(R, M)$ 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 $\mathrm{Hom}_R(R, M)$, 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 $\psi \in \mathrm{Hom}_R(R, M)$ 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. □

Definition 1.35. 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 $f\colon R \to A$ 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. We will discuss these further in Homological Algebra next Fall.

Example 1.36. Let R be a commutative ring with $1_R \neq 0_R$. The ring $R[x_1, \ldots, x_n]$ together with the inclusion map $R \hookrightarrow R[x_1, \ldots, x_n]$ is an R-algebra. More generally, any quotient of $R[x_1, \ldots, x_n]$ is an R-algebra.

The ring of matrices $M_n(R)$ with the homomorphism $r \mapsto rI_n$ is also an R-algebra, as is the group ring $R[G]$ for any group G with the inclusion of R into $R[G]$ given by $r \mapsto re_G$.

Lemma 1.37. *Let R be a commutative ring with $1 \neq 0$ and let M be an R-module. Then $\mathrm{End}_R(M)$ is an R-algebra, with addition and R-action defined as above, and multiplication defined by composition $(fg)(m) = f(g(m))$ for all $f, g \in \mathrm{End}_R(M)$ and all $m \in M$.*

Proof. There are many things to check here, including that:

- The axioms of a (unital) ring are satisfied for $\mathrm{End}_R(M)$.
- There is a ring homomorphism $f\colon R \to \mathrm{End}_R(M)$ such that $f(1_R) = 1_{\mathrm{End}_R(M)} = \mathrm{id}_M$ and $f(R) \subseteq Z(\mathrm{End}_R(M))$.

We will just check the last item and leave the others as exercises. Define $f : R \to \mathrm{End}_R(M)$ by $f(r) = r\,\mathrm{id}_M$. Notice that this is the map μ_r from Example 1.29. Then

$$f(r+s) = (r+s)\,\mathrm{id}_M = r\,\mathrm{id}_M + s\,\mathrm{id}_M = f(r) + f(s)$$

and

$$f(rs) = (rs)\,\mathrm{id}_M = (r\,\mathrm{id}_M) \circ (s\,\mathrm{id}_M) = f(r)f(s)$$

show that f is a ring homomorphism. Moreover, $\mathrm{id}_M \in Z(\mathrm{End}_R(M))$, and once can check easily that $\mu_r \in \mathrm{End}_R(M)$: given any other $g \in \mathrm{End}_R(M)$, 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 $f(R) \subseteq \mathrm{End}_R(M)$. □

Remark 1.38. Let R be a commutative ring with $1 \neq 0$ and let M be an R-module. Then M is also an $\mathrm{End}_R(M)$-module with the action $\phi m = \phi(m)$ for any $\phi \in \mathrm{End}_R(M)$, $m \in M$.

Definition 1.39. 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 1.40. *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 $q\colon M \to M/N$ 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 $rm + N = rm' + N$. □

Definition 1.41. Given an R-module M and a submodule N of M, the map $q\colon M \to M/N$ is the **canonical quotient map**, or simply the canonical map from M to N.

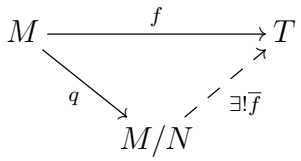
Example 1.42. If R is a field, quotient modules are the same thing as quotient vector spaces. When $R = \mathbb{Z}$, recall that $\mathbb{Z}$-modules are the same as abelian groups, by Lemma 1.13. Quotients of $\mathbb{Z}$-modules coincide with quotients of abelian groups.

Theorem 1.43. *Let N be a submodule of M, let T be an R-module, and let $f : M \to T$ be an R-module homomorphism. If $N \subseteq \ker f$, then the function*

$$\begin{aligned} M/N &\xrightarrow{\overline{f}} T \\ m + N &\longmapsto f(m) \end{aligned}$$

is a well-defined R-module homomorphism. In fact, $\overline{f} : M/N \to T$ is the unique R-module homomorphism such that $\overline{f} \circ q = f$, where $q\colon M \to M/N$ denotes the canonical map.

We can represent this in a more visual way by saying that $\overline{f}$ is the unique R-module homomorphism that makes the diagram



commute.

Proof. By 817, we already know that $\overline{f}$ is a well-defined homomorphism of groups under $+$ and that it is the unique one such that $\overline{f} \circ q = f$. It remains only to show $\overline{f}$ is an R-linear map:

$$\overline{f}(r(m+N)) = \overline{f}(rm+N) = f(rm) = rf(m) = r\overline{f}(m+N).$$

where the third equation uses that f preserves scaling. □

Theorem 1.44 (First Isomorphism Theorem)**.** *Let N be an R-module and let $h : M \to N$ be an R-module homomorphism. Then $\ker(h)$ is a submodule of M and there is an R-module isomorphism $M/\ker(h) \cong \operatorname{im}(h)$.*

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

$$\begin{aligned} \overline{h} : M/\ker(h) &\xrightarrow{\cong} \operatorname{im}(h) \\ m + \ker(f) &\longmapsto h(m). \end{aligned}$$

It remains only to show this map preserves multiplication by scalars. And indeed:

$$\begin{aligned} \overline{h}(r(m+\ker(h))) &= \overline{h}(rm + \ker(h)) && \text{by definition of the } R\text{-action on } M/\ker(h) \\ &= h(rm) && \text{by definition of } \overline{h} \\ &= rh(m) && \text{since } h \text{ is an } R\text{-module homomorphism} \\ &= r\overline{h}(m + \ker h) && \text{by definition of } h. \end{aligned}$$

Theorem 1.45 (Second 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 3, $A+B$ and $A\cap B$ are submodules of M. By the Second Isomorphism Theorem for Groups, there is an isomorphism of abelian groups

$$\begin{aligned} h\colon A/(A\cap B) &\xrightarrow{\cong} (A+B)/B \\ a+(A\cap B) &\longmapsto a+B \end{aligned}$$

It remains only to show h preserves multiplication by scalars:

$$h(r(a+(A\cap B))) = h(ra+A\cap B) = ra+B = r(a+B) = rh(a+(A\cap B)). \qquad \square$$

Theorem 1.46 (Third 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 Third 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:

$$\begin{aligned} h(r((m+A)+B/A)) =& h(r(m+A)+B/A) = h((rm+A)+B/A) \\ &= rm+B = r(m+B) \\ &= rh((m+A)+B/A). \qquad \square \end{aligned}$$

Theorem 1.47 (Lattice Isomorphism Theorem)**.** *Let R be a ring, let N be a R-submodule of an R-module M, and let $q\colon M\to M/N$ be the quotient map. Then the function*

$$\begin{aligned} \{R\text{-submodules of } M \text{ containing } N\} &\xrightarrow{\Psi} \{R\text{-submodules of } M/N\} \\ K &\longmapsto K/N \end{aligned}$$

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 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 Third 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. □

We come to a very important class of examples which will help us study linear transformations using module theory.

Lemma 1.48 ($F[x]$-modules)**.** *Let F be a field. There is a bijection*

$$\{V \text{ an } F[x]\text{-module}\} \longleftrightarrow \{V \text{ an } F\text{-vector space and } T \in \mathrm{End}_F(V)\}.$$

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 $T : V \to V$ be defined by $T(v) = xv$. To show that $T \in \mathrm{End}_F(V)$, note that for any $c \in F$ and $v, v_1, v_2 \in V$ the axioms of the $F[x]$-module give us

$$T(v_1 + v_2) = x(v_1 + v_2) = xv_1 + xv_2 = T(v_1) + T(v_2) \text{ and } T(cv) = x(cv) = c(xv).$$

Conversely, let V be an F-vector space and $T \in \mathrm{End}_F(V)$. We claim that the action of of $F[x]$ on V given by

$$f(x)v = (f(T))(v)$$

satisfies the axioms for a module (exercise!). Alternatively, we can explain this module structure in a more conceptual way, as follows. Consider the evaluation homomorphism $\varphi : F[x] \to \mathrm{End}_F(V), \quad \varphi(f(x)) = f(T)$. Since V is an $\mathrm{End}_F(V)$-module by Remark 1.38, then V is also an $F[x]$-module by restriction of scalars along ϕ; the $F[x]$ action is the one we described above:

$$f(x)v = \varphi(f)(v) = (f(T))(v)$$

Finally, one can check that the two constructions above are inverse to each other. □

Notation 1.49. We shall denote the $F[x]$-module structure on an F-vector space V induced by $T \in \mathrm{End}_F(V)$ by V_T.

Example 1.50. The proposition above says that if we fix an F-vector space V then any linear transformation T gives a different $F[x]$ module structure on V. For example,

- for $T = 0$ the $F[x]$ module V_0 carries an action given by scaling by the constant coefficient of f, that is if $f(x) = a^n x^n + \cdots + a_0$ then
$$f(x)v = (f(0))v = a_0 v \text{ for all } f \in F[x].$$
- for T the "shift operator" that takes $T(e_i) = e_{i-1}$, where e_i is the i-th standard basis vector, the $F[x]$ module V_T is has the action $x^m \begin{bmatrix} v_1 \\ \vdots \\ v_{n-m} \\ v_{n-m+1} \\ \vdots \\ v_n \end{bmatrix} = \begin{bmatrix} v_{m+1} \\ \vdots \\ v_n \\ 0 \\ \vdots \\ 0 \end{bmatrix}$.

1.5 Module generators, bases and free modules

Definition 1.51. Let M be an R-module. A **linear combination** of finitely many elements $a_1, ..., a_n$ of M is an element of M of the form $r_1m + 1 + \cdots + r_nm_n$ for some $r_1, \ldots, r_n \in R$.

Definition 1.52. 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

$$RA := \{r_1a_1 + \cdots + r_na_n \mid n \geq 0, r_i \in R, a_i \in A\}.$$

We M is **generated by** A if $M = RA$. If M is an F-vector space, we 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 9. Let M be an R-module and let $A \subseteq M$. Then RA is the smallest submodule of M containing A, that is

$$RA \quad = \bigcap_{A \subseteq N, N \text{ submodule of } M} N.$$

Exercise 10. Being finitely generated and being cyclic are R-module isomorphism invariants.

Example 1.53. 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 $\{1, x, x^2, \ldots, x^n, \ldots\}$ of monic monomials in the variable x.

(4) Let $M = \mathbb{Z}[x, y]$. M is generated by

- $\{1, x, y\}$ as a ring,
- $\{1, y, y^2, \ldots, y^n, \ldots\}$ as an $\mathbb{Z}[x]$-module, and
- $\{x^iy^j \mid i, j \in \mathbb{Z}_{\geqslant 0}\}$ as a group ($\mathbb{Z}$-module).

Lemma 1.54. *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 $\bar{A} = \{a + N \mid a \in A\}$. □

Definition 1.55. Let M be an R-module and let A be a subset of M. The set A is **linearly independent** if whenever $r_1, \ldots, r_n \in R$ and $a_1, \ldots, a_n$ are distinct elements of A satisfying $r_1a_1 + \cdots + r_na_n = 0$, then $r_1 = \cdots = r_n = 0$. Otherwise A is **linearly dependent**.

Definition 1.56. 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 1.57. 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 $\{1, x, x^2, \ldots, x^n, \ldots\}$ is a basis.

(4) Let $M = \mathbb{Z}[x,y]$. Then $\{1, y, y^2, \ldots, y^n, \ldots\}$ is a basis for the $\mathbb{Z}[x]$-module M, and $\{x^iy^j \mid i,j \in \mathbb{Z}_{\geqslant 0}\}$ is a basis for the $\mathbb{Z}$-module M.

Example 1.58. $\mathbb{Z}/2$ 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 1.59. *If A is a basis of M then every nonzero element $0 \neq m \in M$ can be written uniquely as $m = r_1a_1 + \cdots + r_na_n$ 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. □

Theorem 1.60. *Let R be a ring, M be a free R-module with basis B, N be any R-module, and let $j : B \to N$ be any function. Then there is a unique R-module homomorphism $h : M \to N$ such that $h(b) = j(b)$ for all $b \in B$.*

Proof. We have two things to prove: existence and uniqueness.
Existence: By Lemma 1.59, 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 $h \colon M \to N$ 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 $h : M \to N$ be an R-module homomorphism such that $h(b_i) = j(b_i)$. Then in particular $h \colon (M, +) \to (N, +)$ 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 $h(b_i) = j(b_i)$. □

Corollary 1.61. *If A and B are sets of the same cardinality, and fix a bijection $j : A \to B$. 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 $g : M \to N$ and $h : N \to M$ be the module homomorphisms induced by the bijection $j : A \to B$ and its inverse $j^{-1} : B \to A$, which exist by Theorem 1.60. We will show that h and g are inverse homomorphisms. First, note that $g \circ h : N \to N$ 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 $id_N(b) = b$ for every $b \in B$, by the uniqueness in Theorem 1.60 we have $g \circ h = \mathrm{id}_n$. Similarly, one shows that $h \circ g = \mathrm{id}_M$. □

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 1.62. 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

$$\begin{aligned} F_R(A) &= \{r_1 a_1 + \cdots + r_n a_n \mid n \geqslant 0, r_i \in R, a_i \in A\} \\ &= \left\{ \sum_{a \in A} r_a a \mid r_a \in R, r_a = 0 \text{ for all but finitely many } a \right\}, \end{aligned}$$

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

$$r \left(\sum_{a \in A} r_a a \right) = \sum_{a \in A} (r r_a) a.$$

Exercise 11. 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 1.63 (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 $A \to B$.*

Proof. You will show this in the next problem set (at least in the case where M has a finite basis). □

Definition 1.64. 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 1.65. Let R be a commutative ring with $1 \neq 0$. The rank of R^n is n. Note that by Corollary 1.61, 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 1.66. 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 12. The direct sum and the direct product of an arbitrary family of R-modules are R-modules.

Example 1.67. Suppose that $|A| = n < \infty$. Let $M_1, \ldots, M_n$ be R-modules. The direct product module $M_1 \times \cdots \times M_n$ is the abelian group $M_1 \times \cdots \times M_n$ with ring action given by $r(m_1, \ldots, m_n) = (rm_1, \ldots, rm_n)$ for all $r \in R$ and $m_i \in M_i$. Comparing the definitions we see that

$$M_1 \times \cdots \times M_n = M_1 \oplus \cdots \oplus M_n.$$

If $M_i = R$ for $1 \leqslant i \leqslant n$, then we denote $R^n = \underbrace{R \times \cdots \times R}_{n} = \underbrace{R \oplus \cdots \oplus R}_{n}$.

It is useful to talk about maps from the factors/summands to the direct product/ direct sum and conversely.

Definition 1.68. For $i \in J$ the *inclusion of the i-th factor* into a direct product or direct sum is the map

$$\iota_i \colon M_i \to \prod_{a \in J} M_a \text{ or } \iota_i \colon M_i \to \bigoplus_{a \in J} M_a, \iota_i(m) = (m_a)_{a \in J}, \text{ where } m_a = \begin{cases} m & \text{if } a = i \\ 0 & \text{if } a \neq i \end{cases}.$$

For $i \in J$ the i-th *projection map* from a direct product or a direct sum module is

$$\pi_i \colon \prod_{a \in J} M_a \to M_i \text{ or } \pi_i : \bigoplus_{a \in J} M_a \to M_i, \pi_i \left((m_a)_{a \in J}\right) = m_i.$$

Lemma 1.69. *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*

$$\pi_i \circ \iota_i = \mathrm{id}_{M_i}.$$

Also, $\iota_i(M_i)$ is an R-submodule of the direct product/sum which is isomorphic to M_i.

Note, however, that $\iota_i \circ \pi_i \neq \mathrm{id}$.

Chapter 2

Vector spaces and linear transformations

2.1 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:

$$\text{span}(A) := \{\sum_{i=1}^{n} c_i a_i \mid n \geqslant 0, c_i \in F, a_i \in A\}.$$

Lemma 2.1. *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 $w_1, \ldots, w_n$ 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

$$v = \sum_{i \geqslant 2} \frac{c_i}{c_1} w_i \in \text{span}(I),$$

contrary to assumption. This proves that $I \cup \{v\}$ is a linearly independent set. □

To prove that every vector space has a basis, we will need to use Zorn's Lemma. Before we recall what Zorn's Lemma says, let's recall some notation:

Definition 2.2. A **poset** is a set S with an order relation $\leqslant$ such that for all elements $x, y, z \in S$ we have

- $x \leqslant x$,
- if $x \leqslant y$ and $y \leqslant z$ then $x \leqslant z$, and
- if $x \leqslant y$ and $y \leqslant x$ then $x = y$.

A **totally ordered** set is a poset $(T, \leq)$ such that for all $x, y \in T$ either $x \leqslant y$ or $y \leqslant x$.

Example 2.3. Given a set X, the collection $\mathcal{P}(X)$ of all subsets of X forms a poset with $\leqslant$ defined to be set containment $\subseteq$. Unleess X is empty or a singleton, the poset $\mathcal{P}(X)$ is not totally ordered.

Definition 2.4. Let $(\mathcal{A}, \leqslant)$ be a **poset**, meaning that $\mathcal{A}$ is a set with a partial order $\leqslant$. A subset $\mathcal{B}$ of $\mathcal{A}$ is **totally ordered** if for all $b, b' \in \mathcal{B}$ either $b \leqslant b'$ or $b' \leqslant b$; a totally ordered subset of $\mathcal{A}$ is sometimes called a **chain**. We say a subset $\mathcal{B}$ of $\mathcal{A}$ has an **upper bound** in $\mathcal{A}$ if there exists an element $u_B \in \mathcal{A}$ such that $b \leqslant u_{\mathcal{B}}$ for all $b \in \mathcal{B}$. We say $\mathcal{A}$ has a **maximal element** if there exists $m \in \mathcal{A}$ such that whenever $x \in \mathcal{A}$ and $m \leqslant x$ then $m = x$.

Axiom 2.5 (Zorn's Lemma)**.** If $\mathcal{A}$ is a nonempty poset such that every totally ordered subset $\mathcal{B} \subseteq \mathcal{A}$ has an upper bound in $\mathcal{A}$, then there is a maximal element $m \in \mathcal{A}$.

Some mathematicians refuse to accept Zorn's Lemma into their axiom system. We will at least pretend to be mathematicians who do. Fun fact: Theorem 2.6 is actually equivalent to the Axiom of Choice, meaning that if one replaces the Axiom of Choice in the ZFC axioms for set theory by Theorem 2.6, that does not change set theory – and one would then be able to deduce the Axiom of Choice.

If we accept Zorn's Lemma, we can now show that every vector space has a basis.

Theorem 2.6 (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, note that a corollary of Theorem 2.6 is that every vector space has a basis; in particular, this says that every module over a field is free!

Corollary 2.7. *Every vector space V has a basis. 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. □

Example 2.8. $\mathbb{R}$ has a basis as a $\mathbb{Q}$-vector space; just don't ask me what it looks like.

We will not prove Theorem 2.6. But before we give a formal proof, let's first give a heuristic proof. To so that, start with I. If $\text{span}(I) = V$, then $B = I$ does the job. If not, then since $\text{span}(S) = V$, there must be a $v \in S \setminus \text{span}(I)$. Let $I' := I \cup \{v\}$. Then $I' \subseteq S$ and, by Lemma 2.1, I' is linearly independent. If $\text{span}(I') = V$, we have found our B, and if not we construct I'' from I' just as we constructed I' from I. At this point we would like to say that this process cannot go on for ever, and this is more-or-less true. But at least in an infinite dimensional setting, we need to use Zorn's Lemma to complete the proof rigorously.

Proof of Theorem 2.6. 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 $z_1, \ldots, z_m \in Z$, for each i we have $z_i \in Y_i$ for some $Y_i \in T$. Since T is totally ordered, one of $Y_1, \ldots, Y_m$ contains all the others and hence contains all the z_i's. Since Y_i is linearly independent, this shows $z_1, \ldots, z_m$ 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 $S \subseteq \operatorname{span}(B)$, then $\operatorname{span}(B)$ would have to be all of V. So, there is at least one $v \in S$ such that $v \notin \operatorname{span}(B)$, and set $X := B \cup \{v\}$. Clearly, $I \subset X \subseteq S$ and, by Lemma 2.1, 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. □

Corollary 2.9. *Suppose F is a field an W is 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 2.7 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. □

Remark 2.10. 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 2.11. A vector space is **finite dimensional** if there is spanned by a finite subset.

Thanks to Theorem 2.6, 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.

Lemma 2.12 (Exchange Property)**.** *Let B be a basis for the vector space V and consider any finite set of linearly independent vectors $C = \{c_1, \ldots, c_m\}$ in V. Then there are distinct vectors $b_1, \ldots, b_m$ in B such that $(B \setminus \{b_1, \ldots, b_m\}) \cup C$ is also a basis V.*

Proof. Using induction on k, we will show that for each k with $0 \leqslant k \leqslant m$ there are distinct vectors $b_1, \ldots, b_k$ in B such that $(B \setminus \{b_1, \ldots, b_k\}) \cup \{c_1, \ldots, c_k\})$ is also a basis of V. In the base case, $k = 0$, there is nothing to show. The terminal case, $k = m$, gives us the desired statement.

For the inductive step, assume $B' = (B \setminus \{b_1, \ldots, b_k\}) \cup \{c_1, \ldots, c_k\})$ is also a basis of V. Since $c_{k+1} \in V$, we can write

$$c_{k+1} = \sum_{i=1}^{n} \lambda_i b_i + \sum_{i=1}^{k} \mu_i c_i$$

for some scalars $\lambda_i, \mu_i \in F$ and some elements $b_i \in B \setminus \{b_1, \ldots, b_k\}$. Note that since C is linearly independent, at least one of the scalars λ_i is nonzero. Let i_0 be such that $\lambda_{i_0} \neq 0$, and notice that solving for b_{i_0} from the displayed equation gives that $b_{i_0} \in \operatorname{span}(B'')$ where $B'' = (B' \setminus \{b_{i_0}\}) \cup \{c_{k+1}\})$. Now we can "replace" b_{i_0} by c_k, since the previous statement implies $\operatorname{span}(B'') = \operatorname{span}(B') = V$ and moreover B'' is linearly independent since otherwise B' would be linearly dependent. □

Next, we will show that all bases of the same vector space have the same cardinality. We will only prove this under the assumption that V is finite dimensional, though it is true even if V has infinite dimension.

Theorem 2.13 (Dimension Theorem)**.** *Any two bases of the same vector space have the same cardinality.*

Proof of the finite dimensional case. Suppose V is finite dimensional. Then it has a finite basis B. Let B' be any other basis, and note that we cannot yet assume B' is necessarily finite. Let $\{c_1, \ldots, c_m\}$ be any m-element subset of B' for any m. An immediate consequence of Lemma 2.12 is that $m \leqslant |B|$, since otherwise we could not find m distinct elements of B to replace the c_i's by. Since every finite subset of B' has cardinality no larger than $|B|$, this proves that B' is finite and $|B'| \leqslant |B|$. By symmetry, we obtain $|B| \leqslant |B|'$ too, hence equality follows. □

Definition 2.14. The **dimension** of a vector space V, denoted $\dim_F(V)$ or $\dim(V)$, is the cardinality of any of its bases.

Example 2.15. $\dim_F(F^n) = |\{e_1, e_2, \ldots, e_n\}| = n$.

Theorem 2.16 (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 $m, n \in \mathbb{Z}_{\geqslant 0}$, $F^m \cong F^n$ if and only if $m = n$.

Proof. To show (1), let V be a finite dimensional F-vector space. Then F has a finite spanning set S and by Theorem 2.6 there is a basis $B \subseteq S$ for V. Notice that B is necessarily finite and $V = FB$. Set $|B| = n$ and $B = \{b_1, \ldots, b_n\}$. By Theorem 1.60, there is a linear transformation $f\colon F^n \to V$ such that $f(e_i) = b_i$ as well as a linear transformation $g\colon V \to F^n$ such that $g(b_i) = e_i$. Then both $f \circ g : V \to V$ and $g \circ f : F^n \to F^n$ are linear transformation which agree with the identity map on a basis. Hence by the uniqueness part of Theorem 1.60 we have $f \circ g = \mathrm{id}_V$ and $g \circ f = \mathrm{id}_{F^n}$. 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 $1 \leqslant i \leqslant m$. 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

$$\dim_F(F^n) = n = |\varphi(B)| = |B| = m.$$ □

Remark 2.17.

(1) The same proof as in part (1) of Theorem 2.16 above shows that every finitely generated free R-module is isomorphic to R^n for some $n \geqslant 0$.

(2) Part (2) of the Classification Theorem can be extended to modules over commutative rings as stated in Theorem 1.63; this is a problem in Problem Set 3.

(3) The Classification Theorem yields that dimension is an isomorphism invariant.

Corollary 2.18. *Two finite dimensional vector spaces V and V' over the same field F are isomorphic if and only if* $\dim_F(V) = \dim_F(V')$.

Proof. By Theorem 2.16, V and V' are both of the form $V \cong F^m$ and $V' \cong F^n$, while $F^m \cong F^n$ if and only if $m = n$. □

A word on infinite-dimensional vector spaces.

Example 2.19. Consider the vector space $F[x]$. This cannot be a finite dimensional vector space. For instance, if $\{f_1, \ldots, f_n\}$ were a basis, then setting

$$M = \max_{1 \leqslant j \leqslant n} \{\deg(f_j)\}$$

we see that the element x^{M+1} is not be in the span of $\{f_1, \ldots, f_n\}$. We can find a basis for this space though. Consider the collection $B = \{1, x, x^2, \ldots\}$. This set is linearly independent and spans $F[x]$, thus it forms a basis for $F[x]$. This basis is *countable*, so $\dim_F(F[x]) = \aleph_0 = |\mathbb{N}|$.

Example 2.20. Consider the real vector space

$$V := \mathbb{R}^{\mathbb{N}} = \mathbb{R} \times \mathbb{R} \times \mathbb{R} \times \cdots .$$

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 $E = \{e_1, e_2, \ldots\}$. 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 $v = (1, 1, \ldots)$ can not be represented as a finite linear combination of these elements. Since v is not in $\text{span}(E)$, then by Lemma 2.1 we know that $E \cup \{v\}$ is a linearly independent set. However, this new set $E \cup \{v\}$ does not span BV either, as $(1, 2, 3, 4, \ldots)$ 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 $\dim_{\mathbb{R}}(\mathbb{R}^{\mathbb{N}}) = |\mathbb{R}|$.

We now deduce some formulas that relate the dimensions of various vector spaces.

Theorem 2.21. *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 $n \in \mathbb{Z}_{\geqslant 0}$. We leave the proof for Problem Set 4.

Example 2.22. 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

$$V/W \to \text{span}\{e_2\} \quad (x, y) + W \mapsto (0, y).$$

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 2.23. Let $V = F[x]$, which we saw in Example 2.19 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, then $\dim(F[x]/E) = \infty$.

Definition 2.24. Let $T \colon V \to W$ be a linear transformation. The **nullspace** of T is $\ker(T)$. The **rank** of T is $\dim(\text{im}(T))$.

Corollary 2.25 (Rank-Nullity Theorem)**.** *Let $f \colon V \to W$ 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

$$\dim\left(V/\ker(f)\right) = \dim(\text{im}(V)).$$

By Theorem 2.21, we have

$$\dim(V) = \dim(\ker(V)) + \dim\left(V/\ker(f)\right).$$

Thus

$$\dim(V) = \dim(\ker(V)) + \dim\left(V/\ker(f)\right) = \dim(\ker(V)) + \dim(\text{im}(V)).$$

□

2.2 Linear transformations and homomorphisms between free modules

Exercise 13. If W is a free R-module with basis $C = \{c_1, \ldots, c_m\}$ and $w \in W$, then w can be written *uniquely* as $w = \sum_{j=1}^{m} a_j c_j$ with $a_1, \ldots, a_m \in R$.

Definition 2.26 (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-modules of rank n, and let W be a finitely generated free R-module of rank m. Let $B = \{b_1, \ldots, b_n\}$ and $C = \{c_1, \ldots, c_m\}$ be *ordered* bases of V, W. Given an R-module homomorphism $f : V \to W$, we define elements $a_{ij} \in R$ for $1 \leqslant i \leqslant m$ and $1 \leqslant j \leqslant n$ by the formulas

$$f(b_i) = \sum_{j=1}^{m} a_{j,i} c_j. \tag{2.2.1}$$

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 2.27. By Exercise 13, the coefficients $a_{j,i}$ in equation 2.2.1 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 ith column of $[f]_B^C$. Note that $[f]_B^C$ is an $m \times n$ matrix with entries in R.

Definition 2.28. Let V and W be finite F-vector spaces of dimension n and m with ordered bases B and C respectively and let $f : V \to W$ 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 2.29. If $\mathrm{id}_V : V \to V$ is the identity automorphism of an n-dimensional free R-module V, then for any basis B of V we have $\mathrm{id}_V(b_i) = b_i$ for all i and hence

$$[\mathrm{id}_V]_B^B = I_n.$$

Example 2.30. 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 $d : P_3 \to P_3$ given by taking the derivative $d(f) = f'$. Let $B = \{1, x, x^2, x^3\}$. Then

$$[f]_B^B = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 3 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Example 2.31. Let F be a field and consider a linear transformation $f : V \to W$, 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

$$f\left(\sum l_i b_i\right) = \sum_i l_i f(b_i).$$

Each $f(b_i)$ can be written uniquely as a linear combination of the c_j's as in (2.2.1):

$$f(b_i) = \sum_j a_{j,i} c_j.$$

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

$$f(\sum l_i b_i) = \sum_i l_i f(b_i).$$

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 $f : F^n \to F^m$ 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 2.32.

Theorem 2.32. *Let R be a commutative ring with $1 \neq 0$. 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*

$$\mathrm{Hom}_R(V, W) \cong \mathrm{M}_{m,n}(R) \qquad f \mapsto [f]_B^C.$$

If $V = W$, so that in particular $m = n$, and $B = C$, then the above map is an R-algebra isomorphism $\mathrm{End}_R(V) \cong \mathrm{M}_n(R)$.

Proof. Let $\varphi : \mathrm{Hom}_R(V, W) \to \mathrm{M}_{m,n}(R)$ 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 $f, g \in \mathrm{Hom}_R(V, W)$ and $\lambda \in R$. Let $A = [f]_B^C$ and $A' = [g]_B^C$. Then

$$(f + g)(b_i) = f(b_i) + g(b_i) = \sum_j a_{j,i} c_j + \sum_j a'_{j,i} c_j = \sum_j (a_{j,i} + a'_{i,j}) c_j$$

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 $f, g \in \mathrm{End}_R(V)$ we have $[f \circ g]_B^B = [f]_B^B [g]_B^B$ as an exercise.

Finally, the argument described in Example 2.31 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. □

Corollary 2.33. *For any field F and finite F-vector spaces V and W of dimension n and m respectively,* $\dim(\mathrm{Hom}_F(V, W)) = mn$.

Proof. The isomorphism $\mathrm{Hom}_F(V, W) \cong \mathrm{M}_{m,n}(F)$ gives

$$\dim\left(\mathrm{Hom}_F(V, W)\right) = \dim\left(\mathrm{M}_{m,n}(F)\right) = mn.$$

□

2.3 Change of basis

Definition 2.34. 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 $[\mathrm{id}_V]_B^C$ is a matrix called the **change of basis matrix** from B to C.

In Theorem 2.39 we will show that $[\mathrm{id}_V]_B^C$ is invertible with inverse $\left([\mathrm{id}_V]_B^C\right)^{-1} = [\mathrm{id}_V]_C^B$.

Example 2.35. 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

$$\begin{aligned}\mathrm{id}_V(1) &= 1,\\ \mathrm{id}_V(x) &= 2\cdot 1 + 1\cdot(x-2),\\ \mathrm{id}_V(x^2) &= 4\cdot 1 + 4\cdot(x-2) + 1\cdot(x-2)^2.\end{aligned}$$

Thus, the change of basis matrix is given by $[\mathrm{id}_V]_B^C = \begin{bmatrix}1 & 2 & 4\\ 0 & 1 & 4\\ 0 & 0 & 1\end{bmatrix}$.

Lemma 2.36. *If V, W, U are finitely generated free R-modules spaces with ordered bases B, C, and D, and if $f\colon V \to W$ and $g\colon W \to U$ are R-module homomorphisms, then*

$$[g \circ f]_D^B = [g]_D^C \cdot [f]_C^B.$$

Proof. Given $v \in V$, we have

$$(f\circ g)(v) = f(g(v)) = f([g]_B^C v) = [f]_C^D([g]_B^C v) = ([f]_C^D[g]_B^C)v,$$

so $[f\circ g]_B^B = [f]_B^B[g]_B^B$. □

Definition 2.37. Let V be a finitely generated free module over a commutative ring R. Two R-module homomorphisms $f, g : V \to V$ are **similar** if there is a bijective linear transformation $h : V \to V$ 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 2.38. For elements $A, B \in \mathrm{GL}_n(R)$, the notions of similar and conjugate are the same.

Theorem 2.39. *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 $f : V \to W$ be a homomorphism. Then*

$$[f]_{B'}^{C'} = [\mathrm{id}_W]_C^{C'}[f]_B^C[\mathrm{id}_V]_{B'}^B \tag{2.3.1}$$

In particular, if $g\colon V \to V$ 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 2.36 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 (2.3.1) 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 (2.3.1) to obtain

$$[g]_{B'}^{B'} = [\mathrm{id}_V]_B^{B'}[g]_B^B[\mathrm{id}_V]_{B'}^B = P[g]_B^B P^{-1}. \qquad \square$$

We now come to certain special changes of basis and their matrices:

Definition 2.40. Let R be a commutative ring with $1 \neq 0$, let M be a free R-module of finite rank n, and let $B = \{b_1, \ldots, b_n\}$ be an ordered basis for M. An **elementary basis change operation** on the basis B is one of the following three types of operations:

1. Replacing b_i by $b_i + rb_j$ for some $i \neq j$ and some $r \in R$,
2. Replacing b_i by ub_i for some i and some unit u of R,
3. Swapping the indices of b_i and b_j for some $i \neq j$.

Definition 2.41. Let R be a commutative ring with $1 \neq 0$. An **elementary row operation** on a matrix $A \in \mathrm{M}_{m,n}(R)$ is one of the following three types of operations:

1. Adding an element of R times a row of A to a different row of A.
2. Multiplying a row of A by a unit of R.
3. Interchanging two rows of A.

Definition 2.42. Let R be a commutative ring with $1 \neq 0$. An **elementary matrix** over R is an $n \times n$ matrix obtained from I_n by applying a single elementary row operation:

1. For $r \in R$ and $1 \leqslant i, j \leqslant n$ 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 $1 \leqslant i \leqslant n$ 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 $1 \leqslant i, j \leqslant n$ 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 2.43. Let E be an $n \times n$ elementary matrix.

- E is the change of basis matrix $[\mathrm{id}_V]_{B'}^B$, where B is any basis of V and B' is the basis obtained from B by the corresponding elementary basis change operation.
- If $A \in \mathrm{M}_{n,q}(R)$, then the product matrix EA is the result of performing the corresponding elementary row operation on A.
- If $B \in \mathrm{M}_{m,n}(R)$, then the product matrix BE is the result of performing the corresponding elementary column operation on B.

2.4 A warning on the differences between vector spaces and general free modules

Many of the nice theorems we showed about vector spaces, basis, and dimension do not extend well to general free modules over a commutative ring. Most importantly, Theorem 2.6, which says that for a vector space V every linearly independent set can be extended to a basis and every set that spans V contains a basis, does not hold in general, even in simple cases.

Example 2.44. Let $R = \mathbb{Z}$ and consider the free R-module R. The set $\{2\}$ is linearly independent but it is not a basis for R; given any other element $n \in R$, $\{2, n\}$ is necessarily linearly dependent, since $n \cdot 2 - 2 \cdot n = 0$. Thus we cannot extend $\{2\}$ to a basis of this free module.

Conversely the set $\{2, 3\}$ spans the free module $R = \mathbb{Z}$, but it is not linearly independent, and the subsets $\{2\}$ and $\{3\}$ do not generate the entire free module R.

The failure of Theorem 2.6 leads to the failure of other properties we might expect. For example, one can show that Theorem 2.6 implies that if W is a subspace of V, then $\dim(W) \leqslant \dim(V)$. But when R is a general commutative ring, submodules of free modules do not have to be free, so we can't even talk about dimension; and if we count the number of generators needed, even cyclic modules might have submodules that are not cyclic.

Example 2.45. Let R be a ring an I be an ideal that is not principal. For example, we can take $R = k[x, y]$ with k a field and $I = (x, y)$. Then the R-module R is cyclic, while the submodule I needs at least 2 generators.

Moreover, submodules of free modules are not necessarily free:

Example 2.46. Let k be a field and $R = k[x]/(x^2)$. The submodule $I = (x)$ of the free module R is not free: it is cyclic, generated by x, but $\operatorname{ann}(I) = (x)$ is nontrivial, and thus I is not free.

Chapter 3

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.

3.1 Every module is a quotient of a free module

Lemma 3.1. *Given any ring R with $1 \neq 0$, any direct sum of copies of R is always a free R-module.*

Proof. Suppose that $F = \bigoplus_{i \in \Lambda} R$ is a direct sum of copies of R indexed by some set Λ. The tuples

$$e_i = (a_j)_{j \in \Lambda} \quad \text{with } a_j = 0 \text{ for all } j \neq i \text{ and } a_i = 1$$

generate F, since we can write any element as

$$(c_i)_{i \in \Lambda} = \sum_{i \in \Lambda} c_i e_i.$$

Notice that by definition $c_i \neq 0$ for only finitely many i, so the sum on the right has finitely many nonzero terms. Moreover, the e_i are linearly independent, and thus they form a basis for F. □

We will show in the next chapter that every when R is a field, every R-module is free. In contrast, we will also see that if R is a commutative ring that is not a field, there always exists an R-module that is not free – in fact, given a ring R that is not a field, one can give a very concrete recipe for building nonfree modules.

However, even though not all modules are free, what is true is that every R-module can be written as a quotient of a free module, as follows. Given a module M, first take a generating set for M, say $\Gamma = \{m_i\}_{i \in \Lambda}$. Notice that a generating set always exists: for example, we can take $\Gamma = M$, though of course that is a bit of an overkill, since it's quite likely that some elements can be obtained from linear combinations of others.

Next, we construct a free module on the set Λ; more precisely, we take a free module on as many generators as generators for M that we picked. Now the map

$$\bigoplus_{i\in\Lambda} R \xrightarrow{\pi} M$$

$$(r_i) \longmapsto \sum_{i\in\Lambda} r_i m_i.$$

Notice this map actually makes sense: the tuples (r_i) have only finitely many nonzero entries, and thus $\sum_{i\in\Lambda} r_i m_i$ is a (finite) linear combination of our chosen generators. Moreover, since we chose the m_i to be generators for M, this map π is surjective. It is also easy to check that it is an R-module homomorphism: in fact, this is the R-module homomorphism we would get from Theorem 1.60 by setting $e_i \mapsto m_i$.

By the First Isomorphism Theorem,

$$M \cong \bigoplus_{i\in\Lambda} R/\ker\pi.$$

This shows the following:

Theorem 3.2. *Every R-module is a quotient of a free R-module.*

Notice that the map π we constructed above depends on a choice of generating set for M. Given the map π corresponding to the set of generators $\Gamma = \{m_i\}$, each element in $\ker(\pi)$ is a **relation** among the generators for M: the tuple (r_i) is a relation for the generators $\{m_i\}$ if

$$\sum_{i\in\Lambda} r_i m_i = 0.$$

A nonzero relation among the m_i tells us that the set $\{m_i\}$ is linearly dependent. Thus we see that

$$\pi \text{ is injective} \iff \{m_i\} \text{ is linearly independent} \iff \{m_i\} \text{ is a basis for } M.$$

In particular, the existence of such a map π that is injective is equivalent to M being free. Since π is always surjective (as long as $\{m_i\}$ forms a generating set for M, we can now rephrase this as

$$\pi \text{ is an isomorphism} \iff \{m_i\} \text{ is a basis for } M.$$

The module M is free if and only if we can find a basis for M, thus M if M is free then M is isomorphic to a direct sum of copies of R. Since we have already shown that a direct sum of copies of R is free, we conclude the following:

Theorem 3.3. *An R-module is free if and only if it is isomorphic to a direct sum of copies of R.*

3.2 Presentations for finitely generated modules over noetherian rings

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 3.4. Let R be a commutative ring with $1 \neq 0$, let $A \in \mathrm{M}_{m,n}(R)$, and let $t_A : R^n \to R^m$ be the R-module homomorphism represented by A with respect to the standard bases. Notice that this homomorphism is given by the rule $t_A(v) = Av$. The **R-module presented by** A is the R-module $R^m/\operatorname{im}(t_A)$.

The R-module M presented by $A \in \mathrm{M}_{m,n}(R)$ 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 $\operatorname{im}(t_A)$, which are the images of the standard basis of R^n by t_A.

Example 3.5. The $\mathbb{Z}$-module $M = \mathbb{Z}/6$ is presented by

$$\mathbb{Z} \xrightarrow{6} \mathbb{Z},$$

since $M \cong \mathbb{Z}/\operatorname{im}(t_6) = \mathbb{Z}/(6)$. Notice here we abused notation and wrote 6 instead of the 1×1 matrix $[6]$.

Example 3.6. 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 3.7. 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 $M = \mathbb{Z}^4/\operatorname{im}(t_A)$, where $t_! : \mathbb{Z}^3 \to \mathbb{Z}^4$ 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 1.54 that $M = \mathbb{Z}^4/\operatorname{im}(t_A)$ is generated by the cosets of the e_i. To keep the notation short, we set $m_i = e_i + \operatorname{im}(t_A)$.

Let $N = \operatorname{im}(t_A)$ 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 $q\colon \mathbb{Z}^4 \to M = \mathbb{Z}^4/N$, 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 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

$$M = R\{m_1, m_2, m_3, m_4\} = \{m_1, m_2, m_4\}.$$

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:

$$A = \begin{bmatrix} 2 & 1 & 0 \\ 3 & 9 & 5 \\ 1 & -2 & 7 \\ 0 & 1 & 2 \end{bmatrix} \to \begin{bmatrix} 0 & 5 & -14 \\ 0 & 15 & -16 \\ 1 & -2 & 7 \\ 0 & 1 & 2 \end{bmatrix} \to \begin{bmatrix} 0 & 0 & -24 \\ 0 & 0 & -46 \\ 1 & 0 & 13 \\ 0 & 1 & 0 \end{bmatrix}$$

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

$$C = E_{1,2}(-2)B = \begin{bmatrix} -24 \\ 2 \end{bmatrix}.$$

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 subtracting 12 row 2 from row 1 we also form

$$D = E_{1,2}(12)C = \begin{bmatrix} 0 \\ 2 \end{bmatrix},$$

The last matrix D presents the module $M' = \mathbb{Z}^2/\operatorname{im}(t_D)$ with generators a, b, where

$$a = e_1 + \operatorname{im}(t_D), \quad b = e_2 + \operatorname{im}(t_D)$$

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 $\mathbb{Z}^2 \to \mathbb{Z} \oplus \mathbb{Z}/2$ 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 $(2e_2)\mathbb{Z} = \operatorname{im}(t_D)$. Then the first isomorphism theorem gives $M' = \mathbb{Z}^2/\operatorname{im}(t_D) \cong \mathbb{Z} \oplus \mathbb{Z}/2$.

Lemma 3.8. *Let R be a commutative ring with $1 \neq 0$, $A \in \mathrm{M}_{m,n}(R)$ and $B \in M_{m',n'}(R)$ for some $m, n, m', n' \geqslant 1$. 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:*

(a) an elementary row operation,

(b) an elementary column operation,

(c) deletion of the jth column and ith row of A if $Ae_j = e_i$, that is, if the jth column of A is the vector e_i,

(d) the reverse of 3: insertion of a row and column satisfying $Ae_j = e_i$,

(e) deletion of a column of all 0's,

(f) 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 $M = R^m/\operatorname{im}(t_A)$ and $M' = R^{m'}/\operatorname{im}(t_{A'})$. 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 $E : R^n \to R^n$ maps $\operatorname{im}(A)$ bijectively onto $\operatorname{im}(A')$. Thus Q induces an isomorphism

$$M = R^m/\operatorname{im}(t_A) \xrightarrow{\cong} R^m/\operatorname{im}(t_{A'}) = M'.$$

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

$$\operatorname{im}(t_{A'}) = \operatorname{im}(t_{AE}) = \operatorname{im}(t_A \circ t_E) = \operatorname{im}(t_A)$$

and so $m = m'$ and $M = R^m/\operatorname{im}(t_A) = R^{m'}/\operatorname{im}(t_{A'}) = M'$. 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 $p\colon R^m \to R^{m-1}$ sending

$$\begin{gathered} e_1 \mapsto e'_1, \dots, e_{i-1} \mapsto e'_{i-1} \\ e_i \mapsto 0 \\ e_{i+1} \mapsto e'_i, \dots, e_m \mapsto e'_{m-1} \end{gathered}$$

Similarly, there is a unique R-module homomorphism $q\colon R^n \to R^{n-1}$ sending

$$\begin{gathered} e_1 \mapsto e'_1, \dots, e_{j-1} \mapsto e'_{j-1}, \\ e_j \mapsto 0, \\ e_{j+1} \mapsto e'_j, \dots, e_n \mapsto e'_{n-1}. \end{gathered}$$

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

$$\begin{array}{ccc} R^n & \xrightarrow{A} & R^m \\ {\scriptstyle q}\big\downarrow & & \big\downarrow{\scriptstyle p} \\ R^{n-1} & \xrightarrow[A']{} & R^{m-1} \end{array}$$

commutes by the definition of A'. In particular, $p(\operatorname{im}(t_A)) \subseteq \operatorname{im}(t_{A'})$ and so p induces an R-module homomorphism

$$\overline{p}\colon M \to M',$$

and we claim $\overline{p}$ is bijective.

Since p is onto, so is $\overline{p}$. Suppose $m \in \ker(\overline{p})$. Then $m = v + \operatorname{im}(t_A)$ for some $v \in R^m$ and $p(v) \in \operatorname{im}(t_{A'})$. 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 $v = t_A(u + re_j) \in \mathrm{im}(t_A)$ 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. □

Which modules have presentations? The discussion in Section 3.1 shows that the answer is every module. But if we want to make the presentation be finite (that is, so that the matrix describing the module has finitely many rows and columns) then we need to restrict ourselves to finitely generated modules. 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 3.9. Let k be a field and $R = k[x_1, x_2, \ldots]$ 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 $(x_1, x_2, \ldots)$ generated by all the variables.

Definition 3.10 (Noetherian ring)**.** A ring R is **noetherian** if every ascending chain of ideals

$$I_1 \subseteq I_2 \subseteq I_3 \subseteq \cdots$$

eventually stabilizes: there is some N for which $I_n = I_{n+1}$ for all $n \geqslant N$.

The following characterization of noetherian rings is the key to guaranteeing that submodules of finitely generated modules are also finitely generated.

Theorem 3.11. *A commutative ring R is noetherian if and only if every ideal of R is finitely generated.*

Many rings are noetherian.

Example 3.12.

a) Every field k is noetherian, since (0) and k are the only ideals.

b) If R is a principal ideal domain (PID), then by definition every ideal is generated by a single element, and hence R is noetherian.

c) If R is noetherian, then you will show in Problem Set 5 that every quotient of R is also noetherian.

For more examples, the following famous theorem is useful.

Theorem 3.13 (Hilbert Basis Theorem)**.** *If R is a noetherian ring, then so is $R[x_1, \ldots, x_n]$ for all integers $n \geqslant 1$.*

In the interest of time, and since we really won't need it in this class, I will not give a proof of the Hilbert Basis Theorem. Combining the facts above together gives the following very nice fact:

Corollary 3.14. *Let k be a field and let I be an ideal in $S = k[x_1, \ldots, x_n]$ for some $n \geqslant 1$. Then the ring S/I is noetherian.*

This includes a large collection of the rings that are of most interest in the fields of commutative algebra and algebraic geometry. In contrast, above we saw an example of a ring that is not noetherian:

Example 3.15. Let k be any field and $R = k[x_1, x_2, \ldots]$ be the polynomial ring in arbitrarily many variables. Then R is not noetherian: the ideal $(x_1, x_2, \ldots)$ generated by all the variables is not finitely generated. Alternatively, we can see that the ascending chain of ideals

$$(x_1) \subseteq (x_1, x_2) \subseteq (x_1, x_2, x_3) \subseteq \cdots$$

does not stop.

Theorem 3.16. *Let R be a commutative ring. If R is a noetherian ring, then every submodule of a finitely generated module is also finitely generated.*

Proof. We will first prove that for each $n \geqslant 1$, every submodule of R^n is finitely generated. The base case $n = 1$ holds by Theorem 3.11, since a submodule of R^1 is the same thing as an ideal of R. Assume $n > 1$ and that every submodule of R^{n-1} is finitely generated. Let M be any submodule of R^n. Define

$$\pi \colon R^n \twoheadrightarrow R^1$$

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 finitely generated by Theorem 3.11. Furthermore, by the First Isomorphism Theorem $M/\ker(\pi) \cong \pi(M)$. By Lemma 1.54, 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 $q \colon R^n \twoheadrightarrow T$ 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 $v_1, \ldots, v_m \in q^{-1}(N)$. We claim that $q(v_1), \ldots, q(v_m)$ 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 $v_1, \ldots, v_m$ generated $q^{-1}(N)$, we can find $c_1, \ldots, c_m \in R$ such that

$$b = c_1 v_1 + \cdots + c_m v_m \implies c_1 q(v_1) + \cdots + c_m q(v_m) = q(c_1 v_1 + \cdots + c_m v_m) = q(b) = a. \quad \square$$

In fact, the converse of Theorem 3.16 is also true. More precisely, a commutative ring R is noetherian if and only if every submodule of a finitely generated module is also finitely generated.

Remark 3.17. Let R be a commutative ring. Note that R is a module over itself and a submodule of R is exactly the same thing as an ideal. This module R is always finitely generated as an R-module: 1 generates R, for example. If R is not noetherian, then by Theorem 3.11 R has an ideal I that is not finitely generated. Then I is a submodule of a finitely generated module that fails to be finitely generated.

Theorem 3.18. *Any finitely generated module M over a noetherian ring R has a finite presentation given by an $m \times n$ matrix A, that is, there is an isomorphism*

$$M \cong R^m / \operatorname{im}(t_A),$$

where $t_A \colon R^n \to R^m$ is the map on free modules $t_A(v) = Av$ induced by A.

Proof. Let M be a finitely generated module over a noetherian ring. We start by following the general argument we described in Section 3.1: we choose a finite generating set $y_1, \dots y_m$ of M and obtain an R-module map $\pi \colon R^m \to M$ 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 noetherian, by Theorem 3.16 the submodule $\ker(\pi)$ is also finitely generated, say by $z_1, \dots, z_n$. This too leads to a surjective R-module map $g \colon R^n \to \ker(\pi)$ that sends $e_i \mapsto z_i$. The composition of $g \colon R^n \twoheadrightarrow \ker(\pi)$ followed by the inclusion of $\iota \colon \ker(\pi) \hookrightarrow R^m$ is an R-module homomorphism $t = \iota \circ g : R^n \to R^m$ and hence by Theorem 2.32 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 $M \cong R^m / \operatorname{im}(t_A)$. First note that since $t_A = \iota \circ g$ and g is surjective we have

$$\operatorname{im}(t_A) = \operatorname{im}(\iota \circ g) = \iota(\operatorname{im}(g)) = \iota(\ker(\pi)) = \ker(\pi).$$

By the First Isomorphism Theorem we now have

$$M = \operatorname{im}(\pi) \cong R^m / \ker(\pi) = R^m / \operatorname{im}(t_A). \qquad \square$$

3.3 Classification of finitely generated modules over PIDs

Since any PID is a noetherian ring, 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 3.19 (Smith Normal Form (SNF))**.** *Let R be a PID and let $A \in \mathrm{M}_{m,n}(R)$. 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*

$$a_{11} \mid a_{22} \mid a_{33} \mid \cdots .$$

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 multiplication by units.

Remark 3.20. 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 3.19, 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 3.21. *Let R be a commutative ring with $1 \neq 0$, let $m \geqslant n$, let $A = [a_{ij}] \in \mathrm{M}_{m,n}(R)$ 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 3.22 (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 $r \geqslant 0$, $k \geqslant 0$, and nonzero nonunit elements $d_1, \ldots, d_k$ of R satisfying $d_1 \mid d_2 \mid \cdots \mid d_k$ 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 3.18, M has a presentation matrix A. By Theorem 3.19, A can be put into Smith Normal Form B, where the diagonal entries of B are $b_1, \ldots, b_\ell$ and satisfy $b_1 \mid b_2 \mid \cdots \mid b_k$. 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 3.18, M is isomorphic to the module presented by B. By Lemma 3.21, this is isomorphic to

$$M \cong R^r \oplus R/(b_1) \oplus \cdots \oplus R/(b_{\ll}).$$

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. □

Definition 3.23. Let R be a PID, let $r \geqslant 0, k \geqslant 0$, and let $d_1, \ldots, d_k$ be nonzero nonunit elements of R satisfying $d_1 \mid d_2 \mid \cdots \mid d_k$. 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** $d_1, \ldots, d_k$.

Notice that the invariant factors of M are only defined up to multiplication by units.

Remark 3.24. The classification theorem can be interpreted as saying that M decomposes into a free submodule R^r and a torsion submodule $\mathrm{Tor}(M) = R/(d_1) \oplus \cdots \oplus R/(d_k)$.

Corollary 3.25 (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 $r \geqslant 0$, $k \geqslant 0$, and $n_i \geqslant 2$ for all i, satisfying $n_{i+1} \mid n_i$ for all i. Moreover, the integers r, k, and $n_1, \ldots n_k$ are uniquely determined by G.

Example 3.26. 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:

$$A = \begin{bmatrix} 1 & 6 & 5 & 2 \\ 2 & 1 & -1 & 0 \\ 3 & 0 & 3 & 0 \end{bmatrix} \xrightarrow[R3 \to R3-3R1]{R2 \to R2-2R1} \begin{bmatrix} 1 & 6 & 5 & 2 \\ 0 & -11 & -11 & -4 \\ 0 & -18 & -12 & -6 \end{bmatrix} \to \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -11 & -11 & -4 \\ 0 & -18 & -12 & -6 \end{bmatrix}$$

$$\xrightarrow{C2 \leftrightarrow C4} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -4 & -11 & -11 \\ 0 & -6 & -12 & -18 \end{bmatrix} \xrightarrow[C4 \to C4+3C1]{C3 \to C3+2C2} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -4 & -3 & 1 \\ 0 & -6 & 0 & 0 \end{bmatrix} \xrightarrow{C2 \leftrightarrow C4} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & -3 & -4 \\ 0 & 0 & 0 & -6 \end{bmatrix}$$

$$\to \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -6 \end{bmatrix} \xrightarrow{C3 \leftrightarrow C4} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -6 & 0 \end{bmatrix} \xrightarrow{C3 \to -C3} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 6 & 0 \end{bmatrix}.$$

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 3.27 (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 $r \geqslant 0$, $s \geqslant 0$, prime elements $p_1, \ldots, p_s$ of R (not necessarily distinct), and $e_1, \ldots, e_s \geqslant 1$ 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 $p_1^{e_1}, \ldots, p_s^{e_s}$ 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. □

Definition 3.28. Let R be a PID, let $r \geqslant 0$, $s \geqslant 0$, $p_1, \ldots, p_s$ be prime elements of R, and let $e_1, \ldots, e_s \geqslant 1$. 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 $p_1^{e_1}, \ldots, p_s^{e_s}$ 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 3.29. 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 3.30. *Let G be a finitely generated abelian group. Then there exist $r, s \geqslant 0$, prime integers $p_1, \ldots, p_s$, and positive integers $e_i \geqslant 1$ 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.

We have yet to show Theorem 3.19: every matrix over a PID has a Smith Normal Form. We will need a few auxiliary lemmas.

Definition 3.31. Let R be a PID and let $a_1, \ldots, a_n \in R$. The **greatest common divisor** or **gcd** of $a_1, \ldots, a_n$, denoted $\gcd(a_1, \ldots, a_n)$, is a generator for the principal ideal $(a_1, \ldots, a_n)$. Given a matrix $A \in \mathrm{M}_{m,n}(R)$, $\gcd(A)$ is the gcd of the entries of A. We adopt the convention that $\gcd(0, 0) = 0$ and thus if A is the matrix of all zeroes, then $\gcd(A) = 0$.

Notice that the greatest common divisor is only defined up to multiplication by a unit.

Lemma 3.32. *Let R is a PID. Let $A \in \mathrm{M}_{m,n}(R)$ be any matrix and let $P \in \mathrm{M}_m(R)$ and $Q \in \mathrm{M}_n(R)$ be invertible matrices. Then $\gcd(A) = \gcd(PA) = \gcd(QA)$. In particular, if $B \in \mathrm{M}_{m,n}(R)$ and B is obtained from A by a finite sequence of elementary row and column operations, then $\gcd(A) = \gcd(B)$.*

Proof. First, suppose that $n = 1$, meaning that A is a column, say

$$A = \begin{bmatrix} a_1 \\ \vdots \\ a_m \end{bmatrix} \text{ and let } PA = \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix}.$$

We need to show that $(a_1, \ldots, a_m) = (b_1, \ldots, b_m)$. On the one hand, each b_i is a linear combination of the a_j, so $(b_1, \ldots, b_m) \subseteq (a_1, \ldots, a_m)$. On the other hand,

$$\begin{bmatrix} a_1 \\ \vdots \\ a_m \end{bmatrix} = P^{-1}(PA) = P^{-1} \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix}$$

so each a_j is a combination of the b_i, and $a_j \in (b_1, \ldots, b_m)$. We conclude that we have an equality of ideals $(a_1, \ldots, a_m) = (b_1, \ldots, b_m)$, and thus multiplying a column vector by an invertible matrix does not change the greatest common divisor of the entries.

Now given $A \in \mathrm{M}_{m,n}(R)$, if we denote the ith column of A by A_i, we have

$$PA = \begin{bmatrix} PA_1 & \cdots & PA_m \end{bmatrix}.$$

Since the gcd of each column remains the same, the gcd of all the entries does not change.

To show that $\gcd(AQ) = \gcd(A)$, note that transposing a matrix does not change its gcd nor the fact that its invertible, so we can apply what we have already shown:

$$\gcd(AQ) = \gcd((AQ)^T) = \gcd(Q^T A^T) = \gcd(A^T) = \gcd(A).$$

Finally, applying elementary row and column operations corresponds to multiplying by an elementary matrix on the left or right, and elementary matrices are invertible. □

Lemma 3.33. *Let R be a PID and $x, y \in R$. There exists an invertible 2×2 matrix $P \in \mathrm{M}_2(R)$ such that*

$$P \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \gcd(x, y) \\ 0 \end{bmatrix}.$$

Proof. 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

$$P^{-1} = \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} \begin{bmatrix} 1 & 0 \\ e & 1 \end{bmatrix}. \qquad \square$$

By transposing the matrices in Lemma 3.33, we can show that there exists an invertible 2×2 matrix Q such that

$$\begin{bmatrix} x & y \end{bmatrix} Q = \begin{bmatrix} \gcd(x, y) & 0 \end{bmatrix}.$$

Exercise 14. Show that for any $i \geqslant 1$ and any commutative ring R, the ideal generated by the $i \times i$ minors of a matrix with entries in R is unchanged by row and column operations.

We are now finally ready to show that every matrix over a PID can be put into Smith Normal Form.

Theorem 3.19. (Smith Normal Form) Let R be a PID and let $A \in \mathrm{M}_{m,n}(R)$. 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

$$a_{11} \mid a_{22} \mid a_{33} \mid \cdots.$$

Moreover, the number ℓ of nonzero entries of M is uniquely determined by A, and the nonzero diagonal entries $a_{11}, \ldots, a_{\ell\ell}$ are unique up to multiplication by units.

Proof. Before we begin, note that we will apply a sequence of steps that correspond to multiplication by an invertible matrix on the left or right, and by Lemma 3.32, none of these steps will change the gcd.

To prove existence of such a matrix M, we claim we can multiply A on the right and the left by invertible matrices to transform it into a matrix of the form

$$\begin{bmatrix} g & 0 \\ 0 & B \end{bmatrix}$$

for some $(n-1) \times (m-1)$ matrix B, where $g = \gcd(A)$. If our claim holds, then we are done: notice that g divides every entry of B, since $\gcd(A) = \gcd(g, B)$ by Lemma 3.32, and so by applying this fact to B, and then over and over again, we arrive at a matrix of the desired form M. Notice moreover that if C is an invertible matrix, then so is

$$\begin{bmatrix} 1 & 0 \\ 0 & C \end{bmatrix}.$$

To construct a matrix in the form above, let a be the upper left $(1,1)$ entry of A. First, we are going to show that we can turn A into a matrix of the form

$$\begin{bmatrix} * & 0 \\ 0 & B \end{bmatrix}$$

with the same gcd as A. If a happens to divide all the entries on the first row and column, then we can simply apply elementary row and column operations to get to the desired form. Suppose there exists b on the first column such that $a \nmid b$. Then we may apply an elementary row operation switching rows so that $b = a_{2,1}$ is on the first column, second row. By Lemma 3.33, we can now find an invertible matrix $C \in \mathrm{M}_2(R)$ such that

$$C \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \gcd(a,b) \\ 0 \end{bmatrix}.$$

Consider the $m \times m$ matrix

$$P := \begin{bmatrix} C & 0 \\ 0 & I_{m-2} \end{bmatrix}.$$

Note that PA has $(PA)_{1,1} = \gcd(a,b)$ and $(PA)_{2,1} = 0$, so this step replaces a by $\gcd(a,b)$ and b by 0. By Lemma 3.32, $\gcd(PA) = \gcd(A)$. We can keep repeating this until the top left corner entry divides every entry on the first column, and this process must stop after at most $m-1$ steps, since there are only m elements on the first column.

Similarly, if there exists b on the first row that a does not divide, we can repeat this process by instead multiplying A on the left by an invertible matrix, until we zero out all the remaining entries on the first row and column. Finally, we arrive at a matrix of the form

$$\begin{bmatrix} a & 0 \\ 0 & B \end{bmatrix}.$$

If $a = \gcd(A)$, we are done. If not, then we can find some entry $b = a_{i,j}$ such that $a \nmid b$. We can then add the jth column to the first column, which puts b into the first column without affecting a, since the remainder of the top row is zero. But this brings us back to the previous situation, and we have already shown that we can replace the top left corner by $\gcd(a,b)$.

At each step, we replace a by some c with is both a divisor of a and a multiple of $\gcd(A)$. Our ring R is a UFD, so there are finitely many factors of $a/\gcd(A)$, and this process must stop. This shows that we can eventually replace A by

$$\begin{bmatrix} \gcd(A) & 0 \\ 0 & B \end{bmatrix}.$$

Now it remains to show the uniqueness portion of the theorem. For any i and any matrix B, let $\gcd_i(B)$ denote the gcd of all the $i \times i$ minors of B. By Exercise 14, $\gcd_i$ is unchanged by row and column operations, so $\gcd_i(A) = \gcd_i(M)$.

For a matrix of the form M, the only minors that are nonzero are those where the choices of columns and rows are the same, and hence the only nonzero $i \times i$ minors of M are $g_{s_1} \cdots g_{s_i}$ for some $s_1 < \cdots < s_i$. Since $g_{s_1} \cdots g_{s_i}$ divide each other, it follows that

$$\gcd_i(A) = \gcd_i(M) = g_1 \cdots g_i.$$

In particular, the largest value of i such that some $i \times i$ minor of A is nonzero is ℓ. Also, we have

$$g_i = \frac{\gcd_i(A)}{\gcd_{i-1}(A)}.$$

This proves uniqueness, for it shows that $\ell, g_1, \ldots, g_\ell$ are all defined from A directly, without any choices. □

Example 3.34. 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:

$$\begin{bmatrix} 1 & x-2 \\ x-1 & 0 \end{bmatrix} \xrightarrow{R2 \to R2-(x-1)R1} \begin{bmatrix} 1 & x-2 \\ 0 & -(x-1)(x-2) \end{bmatrix} \xrightarrow{C2 \to C2-(x-2)C1} \begin{bmatrix} 1 & 0 \\ 0 & -(x-1)(x-2) \end{bmatrix}.$$

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

$$\begin{bmatrix} 1 & x-2 \\ x-1 & 0 \end{bmatrix} \xrightarrow{R2 \to R2-(x-1)R1} \begin{bmatrix} 1 & x-2 \\ 0 & -(x-1)(x-2) \end{bmatrix} \xrightarrow{C2 \to C2-(x-2)C1} \begin{bmatrix} 1 & 0 \\ 0 & (x-1)(x-2) \end{bmatrix}.$$

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 4

Canonical forms for endomorphisms

4.1 Rational canonical form

Recall that given an F-vector space V with $\dim_F(V) = n$ and an ordered basis B for V we have proven in Proposition 2.32 that $\mathrm{End}_F(V) \cong \mathrm{M}_n(F)$ via the maps $t \mapsto [t]_B^B$ and $A \mapsto t_A$. Recall also from Lemma 1.48 that to give a finitely generated module over $F[x]$ is the same data as a finite dimensional vector space V and a linear transformation $V \to V$:

Definition 4.1. Let F be a field, let V be a finite dimensional vector space over F, and let $t : V \to V$ be a linear transformation. The $F[x]$-module V_t is defined to be the vector space V with the unique $F[x]$-action satisfying $xv = t(v)$ for all $v \in V$. That is,

$$(r_n x^n + \cdots + r_0)v = r^n t^n(v) + \cdots + r_0 v \quad \text{for all } r_n x^n + \cdots + r_0 \in F[x].$$

Theorem 4.2. *Let F be a field, let V be an F-vector space of dimension n, let $t\colon V \to V$ be a linear transformation, let B be an ordered basis for V, and let $A = [t]_B^B$. Then the matrix $xI_n - A \in \mathrm{M}_n(F[x])$ presents the $F[x]$-module V_t.*

Proof. Let $B = \{b_1, \ldots, b_n\}$ be any basis for V, and note that B is a generating set for V_t as a module over $F[x]$. As we described in Section 3.1, V_t can then be written as a quotient of $F[x]^n$. More precisely, let $e_1, \ldots, e_n$ denote the standard $F[x]$-basis for the free $F[x]$-module $F[x]^n$, and let $\pi\colon F[x]^n \to V_t$ be the surjective $F[x]$-module homomorphism sending e_i to b_i. That is,

$$\pi((g_1(x), \ldots, g_n(x)) = \pi\left(\sum_{i=1}^n g_i(x)e_i\right) = \sum_{i=1}^n g_i(x)b_i = \sum_{i=1}^n g_i(t)b_i.$$

By the First Isomorphism Theorem, we have $V_t \cong F[x]^n/\ker(\pi)$. On the other hand, the matrix $xI_n - A$ determines a map

$$t_{xI_n - A} : F[x]^n \to F[x]^n,$$

and to show that $V_t \cong F[x]^n/\operatorname{im}(t_{xI_n-A})$ it suffices to show that $\operatorname{im}(t_{xI_n-A}) = \ker(\pi)$. Now

$$(\pi \circ t_{xI_n-A})(e_i) = \pi((xI_n - A)e_i) = (xI_n - A)\pi(e_i) = (xI_n - A)b_i = xb_i - Ab_i = t(b_i) - t(b_i) = 0.$$

This proves $\operatorname{im}(xI_n - a) \subseteq \ker(\pi)$. It follows by Theorem 1.43 that there is a surjection of $F[x]$-modules

$$W := F[x]^n / \operatorname{im}(xI_n - A) \twoheadrightarrow V_t.$$

We may also regard this as a surjection of F-vector spaces. Since $\dim_F(V_t) = n$ and the map above is surjective, we have $\dim_F(W) \geqslant n$, which follows from the Rank Nulity Theorem. To establish that the map above is an isomorphism, it suffices to show that $\dim_F(W) \leqslant n$.

Denote by $c_i = e_i + \operatorname{im}(xI_n - A)$ the image of the standard basis of $F[x]^n$ in W. The ith 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 $c_1, \ldots, c_n$ in W; that is; $c_1, \ldots, c_n$ span W as an F-vector space. Therefore, we have the desired inequality $\dim_F(W) \leqslant n$, which completes our proof. □

Corollary 4.3. *Suppose F is a field, V is an F-vector space, and $t : V \to V$ 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 $g_1, \ldots, g_k$ are both the invariant factors of the $F[x]$-module V_t and the entries on the diagonal of the Smith normal form of $xI_n - [t]_B^B$ for any basis B of V.

Proof. Theorem 4.2 says that $xI_n - [t]_B^B$ presents the $F[x]$-module V_t, 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_t is a finite dimensional vector space. If V_t had a free summand, then it would contain an infinite linearly independent set over F, and thus it could not be finite-dimensional. □

Definition 4.4. The polynomials $g_1, \ldots, g_k$ in Corollary 4.3 are called the **invariant factors** of the linear transformation t.

Example 4.5. Let

$$A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \in \mathrm{M}_2(\mathbb{Q}).$$

Then

$$xI_2 - A = \begin{bmatrix} x-1 & -1 \\ 0 & x-1 \end{bmatrix}.$$

We could compute the invariant factors of $t : \mathbb{Q}^2 \to \mathbb{Q}^2$ 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 3.19 that d_1 is the gcd of the entries of $xI_2 - A$ and $d_1 d_2 = \det(xI_2 - A)$. Thus $d_2 = \det(xI_2 - A) = (x-1)^2$ and $d_1 = 1$. Therefore the only invariant factor of t_A is $(x-1)^2$.

You will show the following lemma in Problem Set 6:

Lemma 4.6. *For a monic polynomial $f(x) = x^n + a_{n-1}x^{n-1} + \cdots a_1 x + a_0$ with $n \geqslant 1$, the classes of $1, x, \ldots, x^{n-1}$ form a basis for $F[x]/(f(x))$ regarded as an F-vector space. Relative to this basis, the F-linear operator $l_x : F[x]/(f(x)) \to F[x]/(f(x))$ defined by $l_x(v) = xv$ is given by the following matrix:*

$$C(f) := \begin{bmatrix} 0 & 0 & \cdots & 0 & -a_0 \\ 1 & 0 & \cdots & 0 & -a_1 \\ 0 & 1 & \ddots & 0 & -a_2 \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & 1 & -a_{n-1} \end{bmatrix} = \begin{bmatrix} 0 & \cdots & 0 & -a_0 \\ & & & -a_1 \\ & I_{n-1} & & \vdots \\ & & & -a_{n-1} \end{bmatrix}.$$

Definition 4.7. In the setup of Lemma 4.6, the matrix $C(f)$ is called the **companion matrix** of the monic polynomial f.

Definition 4.8. Given square matrices $A_1, \ldots, A_m$ 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 4.9. If $f : V_1 \to W_1$ and $g : V_2 \to W_2$ are linear transformations, then the map $f \oplus g : V_1 \oplus V_2 \to W_1 \oplus W_2$ 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

$$[f \oplus g]^{\mathcal{C}}_{\mathcal{B}} = \begin{bmatrix} [f]^{C_1}_{B_1} & 0 \\ 0 & [g]^{C_2}_{B_2} \end{bmatrix}.$$

Theorem 4.10 (Rational Canonical Form)**.** *Let F be a field, V a finite dimensional F-vector space, and $t \colon V \to V$ an F-linear transformation. There is a basis B of V such that*

$$[t]^B_B = C(g_1) \oplus \cdots \oplus C(g_l) = \begin{bmatrix} C(g_1) & 0 & 0 & \cdots & 0 \\ 0 & C(g_2) & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & & \vdots \\ 0 & 0 & \cdots & 0 & C(g_k) \end{bmatrix}$$

where $g_1, \ldots, g_l$ are the invariant factors of t, meaning they are monic polynomials of positive degree such that $g_1 \mid g_2 \mid \cdots \mid g_k$. Moreover, the polynomials $g_1, \ldots, g_k$ are unique.

Proof. By Corollary 4.3, $V_t \cong \bigoplus_{i=1}^k F[x]/(g_i(x))$ for unique g_i as in the statement. Set $V_i = F[x]/(g_i(x))$ and note that $V_t = V_1 \oplus \cdots \oplus V_k$. The map $l_x : V_t \to V_t$ given by multiplication by x preserves each summand in this decomposition: $l_x(V_i) \subseteq V_i$. 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 Remark 4.9, B is a basis of V_t and $[t]^B_B = [t|_{V_1}]^{B_1}_{B_1} \oplus \cdots \oplus [t|_{V_k}]^{B_k}_{B_k}$. The result now follows from Lemma 4.6. $\square$

Definition 4.11. In the setup of Theorem 4.10, the matrix $C(g_1) \oplus \cdots \oplus C(g_l)$ is called the **rational canonical form** (RCF) of the linear transformation t. The rational canonical form of a matrix $A \in \mathrm{M}_n(F)$ 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 4.12. Let $A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \in \mathrm{M}_2(\mathbb{Q})$ as in Example 4.5. Because the only invariant factor of $xI_2 - A$ is $(x-1)^2$, the Rational Canonical Form of t_A is

$$RCF(A) = C((x-1)^2) = C(x^2 - 2x + 1) = \begin{bmatrix} 0 & -1 \\ 1 & 2 \end{bmatrix}.$$

We will later show that two matrices have the same Rational Canonical Form if and only if they are similar.

4.2 The Cayley-Hamilton Theorem

Definition 4.13. Let F be a field and let $A \in \mathrm{M}_n(F)$. The **characteristic polynomial** of A is the polynomial $c_A = \det(xI_n - A)$.

Definition 4.14. Let V be an F-vector space of dimension n, and let $t : V \to V$ 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 4.15. 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 2.39, so $C = PAP^{-1}$ for some invertible matrix P. Moreover, diagonal matrices are in the center of $\mathrm{M}_n(R)$, 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 4.16. For any matrices A and B, $c_{A \oplus B} = c_A c_B$.

Definition 4.17. Let F be a field and let $A \in \mathrm{M}_n(F)$. 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 4.18. Let V be an F-vector space of dimension n, and let $t : V \to V$ be a linear transformation. The **minimal polynomial** of t, denoted m_t, is the unique monic polynomial generating the ideal $\mathrm{ann}_{F[x]}(V_t)$ in the PID $F[x]$.

Lemma 4.19. *Let F be a field. Let V be an F-vector space of dimension n with basis B and let $t : V \to V$ 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 $\mathrm{ann}_{F[x]}(V_t) = (m_A)$. Indeed,

$$\begin{aligned}
f \in \mathrm{ann}_{F[x]}(V_t) &\iff f(x)v = 0 \text{ for all } v \in V_t \\
&\iff f(A)v = 0 \text{ for all } v \in V_t \\
&\iff \ker(f(A)) = V_t \\
&\iff \mathrm{rank}(f(A)) = 0 && \text{by the Rank-Nulity Theorem} \\
&\iff f(A) = 0 \\
&\iff f \in (m_A) && \text{by definition of } m_A. \qquad \square
\end{aligned}$$

Remark 4.20. 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 $f(x) \in \mathrm{ann}(V_t) = (m(x))$, 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 Lemma 4.19, 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 $m(x) \mid f(x)$.

Lemma 4.21. *Let F be a field, let V be a finite dimensional F-vector space, and $t : V \to V$ 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 $g_1, \ldots, g_k$ 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, $\mathrm{ann}_{F[x]}(V_t) = (g_k)$, and since g_k is monic we deduce that $m_t = g_k$. $\square$

We can now prove the famous Caley-Hamilton theorem.

Theorem 4.22 (Cayley-Hamilton)**.** *Let F be a field, and let V be a finite dimensional F-vector space. If $t : V \to V$ is a linear transformation, then $m_t \mid c_t$, and hence $c_t(t) = 0$. Similarly, for any matrix $A \in \mathrm{M}_n(F)$ 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 Lemma 4.21, $m = g_k$ and $c = g_1 \cdots g_k$, so $m \mid c$. By definition, we $m(A) = 0$. Since $m | c$, we conclude that $c(A) = 0$. $\square$

Remark 4.23. As a corollary of the Cayley-Hamilton Theorem, we obtain that the minimal polynomial of $t : V \to V$ has degree at most $n = \dim(V)$, since m_t divides c_t, which is a polynomial of degree n.

Lemma 4.24. *Let F be a field and let V be a finite dimensional F-vector space. If $t : V \to V$ is a linear transformation, then $c_t \mid m_t^k$.*

Proof. Since $g_i \mid g_k$ for $1 \leqslant i \leqslant k$, we have $c_t = g_1 \cdots g_k \mid g_k^k = m_t^k$. □

It follows that c_t and m_t have the same roots, not counting multiplicities.

Definition 4.25. Let V be $t\colon V \to V$ 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 $A \in \mathrm{M}_n(F)$, a nonzero $v \in F^n$ satisfying $Av = \lambda v$ for some $\lambda \in F$ is an **eigenvector** of A with **eigenvalue** λ.

Theorem 4.26. *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 Lemma 4.24 we know that $c_t \mid m_t^k$, 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 $c_A(\lambda) = 0$. □

Theorem 4.27. *Let F be a field and let $A, A' \in \mathrm{M}_n(F)$. The following are equivalent:*

(1) A and A' are similar matrices.

(2) A and A' have the same Rational Canonical Form.

(3) A and A' have the same invariant factors.

Proof. To show (1) $\Rightarrow$ (2), suppose A is similar to A'. Then there exists an invertible matrix P such that $A' = PAP^{-1}$, and thus

$$xI_n - A' = xI_nPAP^{-1} = P(xI_n - A)P^{-1}.$$

Thus the matrices $xI_n - A$ and $xI_n - A'$ are also similar. Moreover, by definition we see that similar matrices have the same Smith normal form, and thus A and A' 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 A' have the same invariant factors then there is an isomorphism of $F[x]$-modules $F^n_{t_A} \cong F^n_{t_{A'}}$, which implies by a homework problem in Problem Set 6 that A and A' must be similar. □

Example 4.28. Let us find the minimal and characteristic polynomials of $T : \mathbb{R}^2 \to \mathbb{R}^2$ 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

$$C(x^2 - 1 + 1) = \begin{bmatrix} 0 & -1 \\ 1 & 1 \end{bmatrix}$$

is the rational canonical form of T.

Example 4.29. 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 $j \leqslant 4$. 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 4.30. 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 $j \leqslant 4$. This time we notice that $(A - I_4)^2 = 0$ and so, since $(A - I_4) \neq 0$, we have $m_A(x) = c_A(x) = (x - 1)^2$.

4.3 Jordan canonical form

We now turn to the Jordan canonical form. To motivate it, let us do an example.

Example 4.31. Let us consider

$$A = \begin{bmatrix} 0 & 0 & 8 \\ 1 & 0 & -12 \\ 0 & 0 & 6 \end{bmatrix} = C((x-2)^3) \in \mathrm{M}_3(\mathbb{Q}).$$

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

$$B = \{\overline{1}, \overline{x}, \overline{x^2}\}$$

But notice that

$$B' = \{\overline{(x-2)^2}, \overline{x-2}, \overline{1}\}$$

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 $T' = T - 2\,\mathrm{id}_V$ 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 4.32. 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 4.33 (Jordan Canonical Form Theorem)**.** *Let F be a field, let V be a finite dimensional vector space, and let $t : V \to V$ 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 $e_i \geqslant 1$. Moreover, the polynomials $(x - r_1)^{e_1}, \ldots, (x - r_s)^{e^s}$ 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 Example 4.31. 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 $B_i' = \{\overline{(x - r_i)^{e_i - 1}}, \ldots, \overline{x - r_i}, \overline{1}\}$ 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 $T \colon F[x]/((x - r)^e) \xrightarrow{x} F[x]/((x - r)^e)$ in the basis $B = \{\overline{(x - r)^{e-1}}, \ldots, \overline{x - r}, \overline{1}\}$. Let $T' := T - r \cdot \mathrm{id}$, 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 $T = T' + r\,\mathrm{id}$, we conclude that $[T]_B^B$ is indeed given by the Jordan block $J_e(r)$, as desired. □

Definition 4.34. Let F be a field, V be a finite dimensional vector space, and let $t : V \to V$ be a linear transformation satisfying the property that the characteristic polynomial c_t of t factors completely into linear factors and has elementary divisors $(x - r_1)^{e_1}, \ldots, (x - r_s)^{e_s}$. The matrix $J_{e_1}(r_1) \oplus \cdots \oplus J_{e_s}(r_s)$ is a **Jordan canonical form** (JCF) of t.

A **Jordan canonical form** for $A \in \mathrm{M}_n(F)$ is a Jordan canonical form for the linear transformation $t_A \colon F^n \to F^n$ 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 4.35. We revisit the example of the rotation by 60° but extend scalars to $\mathbb{C}$. That is, start with a matrix A with $c_A(x) = x^2 - x + 1 = (x - w)(x - \overline{w})$ where $w = \frac{1+\sqrt{3}i}{2}$. Since the minimal polynomial of A is $m_A = x^2 - x + 1$, we deduced in Example 4.28 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 $x^2 - x + 1 = (x - w)(x - \overline{w})$, and thus by the CRT

$$\mathbb{C}[x]/(x^2 - x + 1) \cong \mathbb{C}[x]/(x - w) \oplus \mathbb{C}[x]/(x - \overline{w}).$$

Therefore,

$$A \sim C(x - w) \oplus C(x - \overline{w}) = \begin{bmatrix} w & 0 \\ 0 & \overline{w} \end{bmatrix}.$$

The latter matrix is the JCF of A, and in this case the JCF is a diagonal matrix. Notice that if we consider $A \in \mathrm{M}_2(\mathbb{R})$ then the characteristic polynomials fails to factor into linear factors. Hence $A \in \mathrm{M}_2(\mathbb{R})$ does not have a JCF.

Definition 4.36. Let F be a field, let V be a finite dimensional vector space, and let $t : V \to V$ 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 $A \in \mathrm{M}_n(F)$. Then A is **diagonalizable** if A is similar to a diagonal matrix.

Theorem 4.37. *Let F be a field, let V be a finite dimensional vector space, and consider a linear transformation $t : V \to V$. 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). □

Chapter 5

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 of 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 some results about irreducible polynomials from 817.

Theorem 5.1. *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 5.2 (Eisenstein's Criterion)**.** *Suppose R is a domain and let $n \geqslant 1$, and consider the monic polynomial*

$$f(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 \in R[x].$$

If there exists a prime ideal P of R such that $a_0, \ldots, a_{n-1} \in P$ and $a_0 \notin P^2$, then f is irreducible in $R[x]$.

Theorem 5.3 (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 5.4. *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 $R[x]$.*

5.1 Definition and first properties

Definition 5.5. 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 5.6. 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 $\phi\colon R \to S$ is a ring homomorphism, then S is a left R-module via $r \cdot s := \phi(r)s$ by restriction of scalars.

Definition 5.7. 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 5.8. Since $\mathbb{C} = \mathbb{R}(i) = \{a + bi \mid a, b \in \mathbb{R}\}$, we have $[\mathbb{C} : \mathbb{R}] = 2$.

Example 5.9. 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 5.10. *Let F be a field, $p \in F[x]$ with $\deg(p) \geqslant 1$, 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 $\overline{x} := x + (p) \in L$ 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 $F \to L$ 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 $1, x, \ldots, x^{n-1}$ 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 $\overline{x} = x + (p) \in L$, the element $\overline{x}$ is a root of $p(x) \in L[x]$ since

$$p(\bar{x}) = p(x) + (p(x)) = 0_L. \qquad \square$$

Example 5.11. The polynomial $f(x) = x^2 + 1$ is irreducible over $\mathbb{R}$. Theorem 5.10 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 5.12. 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 5.13. Since the intersection of any two subfields of L is again a subfield, $F(a)$ exists and is given by

$$F(\alpha) \quad = \bigcap_{\substack{E \text{ field} \\ F \cup \{\alpha\} \subseteq E \subseteq L}} E.$$

Definition 5.14. Let $F \subseteq L$ be a field extension and $\alpha \in L$. We write

$$F[\alpha] := \{f(\alpha) \mid f \in F[x]\}.$$

Remark 5.15. 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 5.16. *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 5.17. 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 5.18. 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 5.19. 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 $A = \{a_1, \ldots, a_n\}$ is a finite set, we write $F(a_1, \ldots, a_n)$ for $F(A)$.

Remark 5.20. 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 5.21. 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 5.22. *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 $\phi|_F = \mathrm{id}_F$ and*

$$\begin{array}{c} \frac{F[x]}{(p(x))} \longrightarrow F(\alpha) \\ x + (p(x)) \longmapsto \alpha \\ f(x) + (p(x)) \longmapsto f(\alpha). \end{array}$$

Proof. Let $\tilde{\phi} : F[x] \to F(\alpha)$ 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 1.43 we get an injective ring homomorphism

$$\phi\colon \frac{F[x]}{(p(x))} \to F(\alpha)$$

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 $\mathrm{im}(\phi) = F[\alpha] = F(\alpha)$. 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 $\mathrm{im}(\tilde{\phi}) = F(\alpha)$. □

Let's formalize the extra information we have obtained in the course of proving the theorem. First we used the following useful fact:

Remark 5.23. If $\phi : F \to L$ is an injective ring homomorphism and F and L are fields then the image of ϕ is a subfield of L.

Corollary 5.24. *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 5.25 (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 5.26. 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 Corollary 5.25 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 5.27. Another example illustrating Corollary 5.25 is that $\mathbb{Q}(\sqrt{2})$ and $\mathbb{Q}(-\sqrt{2})$ are isomorphic fields. In fact, the are equal: $\mathbb{Q}(\sqrt{2}) = \mathbb{Q}(-\sqrt{2})$. But again Corollary 5.25 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 $\phi(a + b\sqrt{2}) = a - \sqrt{2}$ for $a, b \in \mathbb{Q}$.

The two examples above preview the central idea of Galois theory.

Example 5.28. In Example 5.21, 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 5.29. 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 Example 5.28, 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.

5.2 Algebraic and transcendental extensions

Definition 5.30. For a field extension $F \subseteq L$ and $\alpha \in L$, we say α is **algebraic** over F if $f(x) = 0$ for some nonconstant polynomial $f(x)$. Otherwise, α is **transcendental** over F.

Example 5.31. 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 5.32. *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 (p) 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 Theorem 5.22.

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 $1, \alpha, \alpha^2, \ldots$ 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 $a_0, \ldots, a_n \in F$ 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 $\tilde{\phi} : F(x) \to L$. The image of this field map is

$$\left\{ \frac{g(\alpha)}{f(\alpha)} \mid g, f \in \mathbb{F}[x], f(x) \neq 0 \right\},$$

which is precisely $F(\alpha)$ by Lemma 5.16. The map is injective since it is a field homomorphism that is not identically zero. □

Definition 5.33. 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 5.34. 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 5.35. Note that the minimal polynomial of i over $\mathbb{R}$ is $m_{i,\mathbb{R}}(x) = x^2 + 1$.

Theorem 5.36 (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 $a_1, \ldots, a_m \in A$ and $l_1, \ldots, l_m \in L$. 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 $a_1, \ldots, a_m$ and $b_1, \ldots, b_n$ 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

$\{a_i b_j \mid i = 1, \ldots, m, j = 1, \ldots, n\}$ 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

$$[K : F] = [K : L][L : F]. \qquad \square$$

Example 5.37. In Example 5.21 we showed that $\mathbb{Q}(\sqrt{2}, \sqrt{3}) = \mathbb{Q}(\beta)$ with $\beta = \sqrt{2} + \sqrt{3}$. We claim that $m_{\beta,\mathbb{Q}}(x) = x^4 - x^2 + 1$. 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 Theorem 5.32 (3).

Definition 5.38. A field extension $F \subseteq L$ is **algebraic** if every element $a \in L$ is algebraic over F.

Definition 5.39. 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 5.40. The extension $\mathbb{R} \subseteq \mathbb{C}$ is finite, since $[\mathbb{C} : \mathbb{R}] = 2$.

Lemma 5.41. *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

$$\{a^n \mid n \geqslant 0\}$$

is linearly dependent. Does there exists n such that

$$\{1, a, a^2, \ldots, a^n\}$$

is linearly dependent. Writing an equation of linear dependence, say

$$b_n a^n + \cdots + b_1 a + b_0 = 0$$

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 Theorem 5.32 (5), a must be algebraic over K. $\square$

The converse if false, as shown by the following example:

Example 5.42. 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 Theorem 5.32 (5) we conclude that the degree of this extension is $[\mathbb{Q}(a) : \mathbb{Q}] = n$. Thus $\mathbb{Q}$ 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 $\mathbb{Q}$ 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 5.43. *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 $a_n x^n + \cdots + a_1 x + a_0 \in L[x]$. Note that this polynomial belongs to $F(a_0, \ldots, a_n)[x]$ too, and so α is algebraic over $F(a_0, \ldots, a_n)$.

Consider the chain of field extensions

$$F \subseteq F(a_0) \subseteq F(a_0, a_1) \subseteq \cdots \subseteq F(a_0, a_1, \ldots, a_n) \subseteq F(a_0, \ldots, a_n, \alpha).$$

Each $a_i \in L$ is algebraic over F for all i, and α is algebraic over $F(a_0, a_1, \ldots, a_n)$, so each step in our tower of extensions consists of adding an algebraic element to the previous field. By Theorem 5.32, each step in this chain has finite dimension. By the Degree Formula,

$$[F(a_0, \ldots, a_n, \alpha) : F] = [F(a_0, \ldots, a_n, \alpha) : F(a_0, \ldots, a_n)] \cdots [F(a_0) : F]$$

is finite. Moreover, if we reorder the tower above to start from $F \subseteq F(\alpha)$, by the Degree Formula we have

$$[F(\alpha) : F][F(a_0, \ldots, a_n, \alpha) : F(\alpha)] = [F(a_0, \ldots, a_n, \alpha) : F] < \infty.$$

Therefore, $[F(\alpha) : F]$ is finite. By Theorem 5.32 (5) again, α is algebraic over F. □

In the proof of Theorem 5.43, we also showed the following corollary of the Degree Formula:

Corollary 5.44. *If $\alpha_1, \ldots, \alpha_n$ are algebraic over F, then $F \subseteq F(\alpha_1, \ldots, \alpha_n)$ is a finite algebraic extension.*

5.3 Algebraically closed fields and algebraic closure

Definition 5.45. 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 5.46. *For any field extension $F \subseteq L$, the set $\overline{F}_L$ is a 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 Lemma 5.41. 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 $F(\alpha, \beta) \subseteq \overline{F}_L$. □

The notion of algebraic closure is closely related (pun intended) to being algebraically closed.

Definition 5.47. 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 5.48. 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 5.49. *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 Theorem 5.43, $F \subseteq \overline{F}_L(\alpha)$ 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. □

Remark 5.50. 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 5.51. Lemma 5.49 shows that the field $\overline{\mathbb{Q}}$ defined in Example 5.42 is algebraically closed.

Definition 5.52. 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 5.53. Let L be an algebraic closure of F. Since L is algebraically closed by definition, by Lemma 5.49 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 5.54.

1) Since $[\mathbb{C} : \mathbb{R}] = 2$, the extension $\mathbb{R} \subseteq \mathbb{C}$ is finite, and thus by Lemma 5.41 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 Lemma 5.49, 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 5.55. *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. □

We are now ready to show that every field has an algebraic closure, and that algebraic closures are unique up to isomorphism.

Theorem 5.56 (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 $\phi|_F = \mathrm{id}_F$.*

Proof of existence of algebraic closures. First, we reduce the proof of existence to the following claim:

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's assume the claim holds. By using this fact repeatedly, we may form a tower of field extensions

$$F = F_0 \subseteq F_1 \subseteq F_2 \subseteq \cdots$$

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 Lemma 5.55, 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

$$1 = g_1 f_1(y_{f_1}) + \cdots + g_m f_m(y_{f_m})$$

in R. There exists finite extension E of F in which each f_i has a root α_i: by Theorem 5.10, f_i has a root α_i in some extension of F, and $F(\alpha_1, \ldots, \alpha_n)$ is a finite extension of F by Corollary 5.44. 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 $F \to K$ 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 $\overline{y_f} \in K$ 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 a 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. □

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 $i|_F = \mathrm{id}_F$. 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 Lemma 5.55, there is an isomorphism $E(\alpha) \to E'(\alpha')$ extending the isomorphism i. Since $E'(\alpha) \subseteq L$, 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 $i(L) = L'$. □

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 5.57. Given a field F, we will write $\overline{F}$ for its algebraic closure inside an algebraically closed field extension of F.

By Theorem 5.56, $\overline{F}$ is defined only up to isomorphism.

Example 5.58. The field $\mathbb{C}$ is the algebraic closure of $\mathbb{R}$, so we write $\overline{R} = \mathbb{C}$.

Example 5.59. In Example 5.42, 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 Example 5.42. This field $\overline{\mathbb{Q}}$ is sometimes called the **field of algebraic numbers**.

5.4 Splitting fields

Definition 5.60. 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 5.61. *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 $c, \alpha_1, \ldots, \alpha_n \in E$, then $F(\alpha_1, \ldots, \alpha_n)$ 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 $F(\alpha_1, \ldots, \alpha_n)[x]$. Hence, given some splitting field L of f over F, by the minimality condition in the definition, we must have $L \subseteq F(\alpha_1, \ldots, \alpha_n)$. 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 $F(\alpha_1, \ldots, \alpha_n) \subseteq L$. □

Remark 5.62. Note that there may be repetitions in the list $\alpha_1, \ldots, \alpha_n$, but that does not affect the validity of anything here.

Theorem 5.63 (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 Theorem 5.56. Let $\alpha_1, \ldots, \alpha_m$ be the roots of f in $\overline{F}$. By construction, $F(\alpha_1, \ldots, \alpha_m)$ is a splitting field of f. □

Example 5.64.

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 5.65. 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 $g_1, \ldots, g_n$: to adjoin all the roots of $g_1, \ldots, g_n$ is the same as forming the splitting field of $g_1 \cdots g_n$.

Example 5.66. 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 Example 5.21.

Lemma 5.67. *For every field F and every nonconstant polynomial $f \in F[x]$ of degree $n \geqslant 1$, there exists a splitting field L for f over F with $[L : F] \leqslant n!$.*

Proof. Intuitively, we just need to adjoin all the roots of f, which is possible since we already know we can adjoint 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 Theorem 5.10 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 $[E : K] \leqslant (n - 1)!$ in which g splits completely. Then f also splits completely in E and by the Degree Formula

$$[E : F] = [E : K][K : F] \leqslant (n - 1)!n = n!.$$

Finally, let

$$f(x) = \prod_i (x - \alpha_i)$$

be the factorization of f in $E[x]$, and set $L = F(\alpha_1, \ldots, \alpha_n)$. By Lemma 5.61, L is a splitting field of f over F. By the Degree Formula,

$$[E : F(\alpha_1, \ldots, \alpha_n)][F(\alpha_1, \ldots, \alpha_n) : F] = [E : F] \leqslant n! \implies [F(\alpha_1, \ldots, \alpha_n) : F] \leqslant n!. \quad \square$$

The degree of the splitting field of f can be $n!$, but it can also be much smaller.

Example 5.68. 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 Lemma 5.67 above that $[L : \mathbb{Q}] \leqslant 3! = 6$. 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

$$[L : \mathbb{Q}] = [L : \mathbb{Q}(\sqrt[3]{2})][\mathbb{Q}(\sqrt[3]{2}) : \mathbb{Q}] \geqslant 3 \cdot 2 = 6.$$

We conclude that $[L : \mathbb{Q}] = 6$.

Example 5.69. 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 nth roots of 1. One of these is $\zeta_n := e^{2\pi i/n}$. Notice that every other nth 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 Lemma 5.67. In fact, it is at most $n-1$, since f factors as $(x-1)(x^{n-1}+\cdots+x+1)$, and hence the minimum polynomial of ζ_n is a divisor of $x^{n-1} + \cdots + x + 1$.

When $n = p$ is prime, then one can show that $x^{p-1} + \cdots + x + 1$ 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 that $n - 1$ 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 $m_{i,\mathbb{Q}}(x) = x^2 + 1$.

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 5.70. *Let F be a field, f be a nonconstant polynomial, and consider a field isomorphism $\theta : F \to F'$. Consider the isomorphism $\tilde{\theta} : F[x] \to F'[x]$ induced by θ, and let $f' = \tilde{\theta}(f) \in F'[x]$ be the polynomial corresponding to f.*

(a) Suppose f is irreducible. 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

$$\widehat{\theta} : F(\alpha) \to F'(\alpha')$$

that extends the map θ and sends α to α'.

(b) Suppose L is a splitting field of f over F and L' is a slitting field of g over F'. Then there is a field isomorphism $\widehat{\theta} : L \to L'$ extending θ.

Proof.

(a) 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 Corollary 5.25. 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':

$$\alpha \mapsto x + (f) \mapsto x + (f') \mapsto \alpha'.$$

(b) 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 Corollary 5.25.

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 part (a), there is an isomorphism

$$\phi\colon F(\alpha) \to F'(\alpha')$$

extending θ and sending α to α'.

In $F(\alpha)$, p factors as $p = (x-\alpha)g$, and in $F'(\alpha')$, $p' = (x-\alpha')g'$. Moreover, since ϕ extends θ and $\phi(\alpha) = \alpha'$, it follows that $\phi(p) = p'$ and $\phi(x-\alpha) = x - \alpha'$. Thus we have

$$(x-\alpha')\phi(g) = \phi(x-\alpha)\phi(g) = \phi((x-\alpha)g) = \phi(p) = p' = (x-\alpha')g'.$$

Since F' is a domain, we conclude that $\phi(g) = g'$.

Note that L is a splitting field of 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 $\widehat{\theta} : L \to L'$ that extends ϕ and hence extends θ. □

Corollary 5.71 (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 $\phi : L \to L'$ that fixes F, i.e. $\phi|_F = \mathrm{id}_F$.*

Proof. Apply part (2) of Theorem 5.70 to $\theta = \mathrm{id}_F$. □

We will now be referring to *the* splitting field of F, rather than *a* splitting field, thanks to the uniqueness result above.

Corollary 5.72. *If L is the splitting field over F of an irreducible polynomial $f(x) \in \mathbb{F}[x]$ and if $\alpha, \beta \in L$ are any two roots of f, then there is a field automorphism $s\colon L \to L$ such that $s|_F = \mathrm{id}_F$ 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 Corollary 5.25 there is an isomorphism $\tau : F(\alpha) \to F(\beta)$ such that $\tau|_F = \mathrm{id}_F$ 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 Corollary 5.71, an isomorphism such as s exists. □

Example 5.73. Let L be the splitting field of x^3-2 over $\mathbb{Q}$; so $L = \mathbb{Q}(\sqrt[3]{2}, e^{2\pi i/3}\sqrt[3]{2}, e^{4\pi i/3}\sqrt[3]{2})$. Corollary 5.72 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.

Corollary 5.72 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.

Corollary 5.74. *Let F be a field and $f \in F[x]$ a polynomial of degree $n \geqslant 1$. For any splitting field S of f, $[S : F] \leqslant n!$.*

Proof. By Lemma 5.67, there exists a splitting field E of f with $[E : F] \leqslant n!$. By Corollary 5.71, splitting fields are unique up to isomorphism, so we conclude that $[S : F] \leqslant n!$ for any splitting field S of f. □

5.5 Separability

Definition 5.75. Let R be a commutative ring. The **characteristic** $\operatorname{char}(F)$ of F is defined to be the smallest positive integer n such that

$$n \cdot 1_F = \underbrace{1_F + \ldots + 1_F}_{n} = 0_F$$

if such and integer exists, and 0 otherwise.

Example 5.76. Here are some familiar examples: $\operatorname{char}(\mathbb{Z}) = 0$ and $\operatorname{char}(\mathbb{Z}/n) = n$.

Definition 5.77. For a field F its **prime field** is the subfield of F generated by 1_F.

You proved the following lemma in Problem Set 8:

Lemma 5.78. *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 $\operatorname{char}(F)$ is either 0 or a prime number p.

The most important tool we have at our disposal if $\operatorname{char}(R) = p$ is a prime is the Frobenius endomorphism. This is a simple but very powerful tool, given by taking pth powers of each element. The fact that the pth power map is a ring homomorphism is known as the **Freshman's Dream**.

Lemma 5.79 (Freshman's Dream)**.** *If R is a commutative ring of prime characteristic p, then the function*

$$F \colon R \to R, \quad \phi(c) = c^p$$

is a ring homomorphism

Proof. Since

$$(a+b)^p = \sum_{k=0}^{p} \binom{p}{k} a^k b^{n-k}$$

and the binomial coefficients $\binom{p}{k}$ are divisible by p for any $1 \leqslant k \leqslant p-1$, 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. □

Remark 5.80. Let R be a commutative ring of prime characteristic p. Since $\mathrm{End}(R)$ is closed under composition, the eth iterate of the Frobenius endomorphism

$$F^e = \underbrace{\phi \circ \cdots \circ \phi}_{n} \colon R \to R, F^e(x) = x^{p^n}$$

is also a ring homomorphisms.

Definition 5.81. Let F be a field, $f \in F[x]$, and let α be a root of f in some field extension L of F. The **multiplicity** of α in f to be the number of times $x-\alpha$ appears in the factorization

$$f = \prod_i (x - \beta_i)$$

of f in some (any) splitting field. If the multiplicity of every root of f is 1, we say f is **separable**.

Example 5.82. 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 5.83. For any field F and $f = a_n x^n + \cdots + a_1 x + a_0 \in F[x]$, define its **derivative** to be

$$f' = n a_n x^{n-1} + (n-1) a_{n-1} x^{n-2} + \cdots + 2a_2 x + a_1.$$

Remark 5.84. The derivative is F-linear: For $f, g \in F[x]$, $(f+g)' = f' + g'$ and $(af)' = af'$ for all $a \in F$.

Lemma 5.85 (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 10, $\gcd(f, f') = 1$ if and only if f and f' have no common roots in $\overline{F}$.

c) Assume f is irreducible. Since the degree of f' is strictly less than the degree of f, we have that $\gcd(f, f') \neq 1$ if and only if $f' = 0$. □

Definition 5.86. An algebraic field extension L/F is called **separable** if for every $\alpha \in L$, the minimal polynomial $m_{\alpha,F}(x)$ of α over F is separable.

Corollary 5.87. *If* $\mathrm{char}(F) = 0$*, then 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 $\mathrm{char}(F) = 0$, $m'_{\alpha,F} \neq 0$. Since $m_{\alpha,F}$ is irreducible in $F[x]$, Lemma 5.85 implies $m_{\alpha,F}(x)$ is separable. □

Lemma 5.88. *Let* F *be a field with* $\mathrm{char}(F) = p$ *for some prime number* p*, and let* K/F *be an algebraic extension.*

a) If b *is an element of* F *that is not a* p*th power of an element of* F*, and* K/F *is an algebraic extension of* F *that contains a root of* $x^p - b$*, then* K/F *is not separable.*

b) If every element of F *is the* p*th power of another element of* F*, then every algebraic extension* K/F *is separable.*

Proof.

a) In general, for such an 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 $q := m_{\alpha,F}$ is not separable. Since α is a root of $x^p - b$, we have $q \mid x^p - b$. In $L[x]$, using the Freshman's Dream, we have

$$(x - \alpha)^p = x^p - \alpha^p = x^p - b.$$

If follows that q must divide $(x - \alpha)^p$ in $L[x]$ and hence must have the form $(x - \alpha)^i$ for some $1 \leqslant i \leqslant p$. But $i \neq 1$ since $\alpha \notin F$. Thus α is a multiple root of q and q is irreducible in $F[x]$.

b) Assume $\mathrm{char}(F) = p$ and every element of F is the pth power of another element. If $q' = 0$, then we must have that q is a sum of terms of the form bx^{mp}, for some $m \geqslant 0$ 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.

□

Corollary 5.89. *Every algebraic field extension of a finite field is separable.*

Proof. Problem Set 10. □

Fields which have the property that every one of their algebraic extensions is separable are called **perfect fields**. To summarize the section on separability, we have shown that fields of characteristic 0 and fields K of characteristic p such that $K = K^p$, in particular finite fields, are separable.

Chapter 6

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

$$G \times S \to S,$$

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 $s \in S$.

Definition 6.2. 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 6.3. 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 $\mathrm{Orb}_G(p) = S$ for any $p \in S$.

Definition 6.4. 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

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

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

6.1 Group actions on field extensions

Definition 6.5. Let K be a field. The **automorphism group** of K, denoted $\mathrm{Aut}(K)$, is the collection of field automorphisms of K, with the binary operation of composition.

Let K/F be a field extension. The **automorphism group** of K/F, denoted $\mathrm{Aut}(K/F)$, is the collection of field automorphisms of K that restrict to the identity on F, with the binary operation of composition.

Exercise 15. Let K/F be a field extension. Then $\mathrm{Aut}(K)$ is a group and $\mathrm{Aut}(K/F)$ is a subgroup of $\mathrm{Aut}(K)$.

Example 6.6. The automorphism group $\mathrm{Aut}(\mathbb{C}/\mathbb{R})$ has two elements, the identity and the element given by complex conjugation. It is easy to see each of these is an element of $\mathrm{Aut}(\mathbb{C}/\mathbb{R})$: this amounts to the fact that complex conjugation commutes with addition and multiplication of complex numbers. To see these are all the automorphisms, suppose $\tau \in \mathrm{Aut}(\mathbb{C}/\mathbb{R})$. Since $\tau|_{\mathbb{R}} = \mathrm{id}_{\mathbb{R}}$, 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 $\tau(i) = \pm 1$.

Example 6.7. For any squarefree integer d, $\mathrm{Aut}(\mathbb{Q}(\sqrt{d})/\mathbb{Q})$ also has two elements, the identity and the map sending $a + b\sqrt{d}$ to $a - b\sqrt{d}$. The details are similar to the previous example, so we leave them as an exercise.

Remark 6.8. Let L be a field and let $\sigma \in \mathrm{Aut}(L)$. Then the UMP of polynomial rings gives that there is an induced ring homomorphism $(-)^\sigma : L[x] \to L[x]$ such that for each $q = a_n x^n + \cdots + a_0 \in K[x]$, we have

$$q^\sigma(x) = \sigma(a_n)x^n + \cdots + \sigma(a_0).$$

If $\sigma \in \mathrm{Aut}(L/K)$ and $q \in K[x]$, then $q^\sigma = q$.

Lemma 6.9. *Let K/F be a field extension, let $\sigma \in \mathrm{Aut}(K/F)$, and let $q \in F[x]$.*

a) For all $c \in K, \sigma(q(c)) = q(\sigma(c))$.

b) If $b \in K$ is a root of q, then $\sigma(b)$ also is a root of q.

Proof.

a) σ is a homomorphism and it restricts to the identity on F.

b) If $\sigma \in \mathrm{Aut}(L/F)$ and $q \in F[x]$, then we have $q^\sigma = q$. Since $\sigma(q(\alpha)) = q^\sigma(\sigma(\alpha))$ for all $\alpha \in L$, it follows that if α is a root of $f(x)$, then

$$0 = \sigma(q(\alpha)) = q^\sigma(\sigma(\alpha)) = q(\sigma(\alpha))$$

showing that $\sigma(\alpha)$ is also a root of q. □

We now come to the main idea connecting filed 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 6.10. *Let K/F be the splitting field of a polynomial $q \in F[x]$. Let S be the set of distinct roots of q in K, and let $n = |S|$.*

a) $Aut(K/F)$ acts faithfully on Y, via $\sigma \cdot b = \sigma(b)$ for all $\sigma \in \mathrm{Aut}(K/F)$ and $b \in Y$, and hence $\mathrm{Aut}(K/F)$ is isomorphic to a subgroup of S_n.

b) If f is an irreducible polynomial in $F[x]$, then $\mathrm{Aut}(K/F)$ acts transitively on S.

c) The orbits of the action of $\mathrm{Aut}(K/F)$ on S are the subsets of S that are the roots of the same irreducible factor of q.

Proof.

a) Let $G = Aut(K/F)$. To see that the the action claimed above is well-defined, notice that if $b \in S$ then $\sigma(b) \in S$ by Lemma 6.9. Now we have

$$\sigma \cdot \sigma' \cdot b = \sigma(\sigma'(b)) = (\sigma \circ \sigma')(b), \quad \forall \sigma, \sigma' \in G, b \in S$$

$$1_G \cdot b = \mathrm{id}_K(b) = b, \quad \forall \sigma \in G, b \in S$$

so the given formula indeed defines an action of G on S.

The action is faithful since if σ fixes all the roots $\alpha_1, \ldots, \alpha_n$ of f, then it fixes every element of $F(\alpha_1, \ldots, \alpha_n) = L$.

b) Now assume $f(x)$ is an irreducible polynomial. Let α, β be any two roots of $f(x)$. Theorem 5.70 (1) shows there is an isomorphism $\theta : F(\alpha) \to F(\beta)$ that fixes F.

We have $f(x) = (x - \alpha)g(x)$ and $f(x) = (x - \beta)h(x)$.Since $f^\theta = f$ and $(x - \alpha)^\theta = x - \beta$, we must have $g^\theta(x) = h(x)$. Theorem 5.70 (2) applies, showing there is an automorphism $\sigma : L \to L$ that extends θ. It is easy to see that σ fixes F, so $\sigma \in \mathrm{Aut}(L/F)$, and that $\sigma(\alpha) = \beta$. This proves the action is transitive on the set of roots of any irreducible polynomial.

c) For each $b \in S$ the orbit of b is $\{\sigma(b) \mid \sigma \in \mathrm{Aut}(K/F)\}$. Since b is a root of $f(x)$ there exists an irreducible factor $p(x) \in F[x]$ of $f(x)$ such that b is a root of $p(x)$. Then, since $p(x) \in F[x]$, Lemma 6.9 shows that $\sigma(b)$ will be a root of $p(x)$ for any $\sigma \in \mathrm{Aut}(K/F)$. Thus the orbit of b is contained in the set of roots of $p(x)$ in K.

Conversely, since by part (2) $\mathrm{Aut}(K/F)$ acts transitively on the set of roots of $p(x)$, we have that every root of $p(x)$ is in the orbit of b under the action of $\mathrm{Aut}(K/F)$, hence the desired conclusion follows. □

Index