

Chapter 3

Introduction to modules

3.1 Modules, submodules and homomorphisms

The problem of classifying all rings is much too general to ever hope for an answer. But one of the most important tools available – for general non-commutative rings – is really to focus not on the ring itself, but on the structure of its *module category*.

Let R be a ring. A *left R -module* means an Abelian group M together with a multiplication $R \times M \rightarrow M$ denoted $(r, m) \mapsto rm$ such that

$$(M1) \quad r(m_1 + m_2) = rm_1 + rm_2;$$

$$(M2) \quad (r_1 + r_2)m = r_1m + r_2m;$$

$$(M3) \quad (r_1r_2)m = r_1(r_2m);$$

$$(M4) \quad 1_R m = m$$

for all $r, r_1, r_2 \in R, m, m_1, m_2 \in M$. Strictly speaking, what I am calling a left R -module should be called a *unital* left R -module because I always include axiom (M4). By the way, the axioms imply that $(-r)m = -(rm)$, $0_R m = 0_M$ and $r0_M = 0_M$ for all $r \in R, m \in M$.

There is another notion called a right R -module. As you can probably guess, this is exactly the same idea, but the operation of $r \in R$ on $m \in M$ is written on the *right*, i.e. the operation is a map $M \times R \rightarrow M$ denoted $(m, r) \mapsto mr$. The axioms become

$$(M1') \quad (m_1 + m_2)r = m_1r + m_2r;$$

$$(M2') \quad m(r_1 + r_2) = mr_1 + mr_2;$$

$$(M3') \quad m(r_1r_2) = (mr_1)r_2;$$

$$(M4') \quad m1_R = m$$

for all $r, r_1, r_2 \in R, m, m_1, m_2 \in M$.

You need to be somewhat ambidextrous when working with modules. I will try usually to work with left modules – and all the results we prove for left modules have right module analogues. You really cannot avoid the need for right modules from time to time, however. If necessary, we will write ${}_R M$ to emphasize that M is a left R -module, or M_R to emphasize that M is a right R -module.

Now, given any ring R , let R^{op} denote the same Abelian group but with new multiplication \cdot defined by $r \cdot s := sr$, the right hand side being the old multiplication in R . If M is a left R -module, then we can view M as a right R^{op} -module, by defining the right action of R^{op} on M by

$mr := rm$, where the right hand side of this equation is the old left action of R on M . Similarly, any right R -module can be viewed as a left R^{op} -module. This “op” trick will occasionally be useful for technical reasons.

In the special case that R is commutative, $R^{op} = R$. So we obtain from the previous paragraph *in the commutative case* the standard way to view any left R -module as a right R -module (or vice versa): if M is a left R -module, define a right action of R on M by $mr := rm$. In view of this, when working with commutative rings, I allow myself to be especially careless and usually just talk about R -modules without making the “left” or “right” clear.

Before giving the many examples you already know, let me define R -submodules. Given a left R -module M , an R -submodule $N \leq M$ means a sub-Abelian group of M such that $rn \in N$ for all $r \in R, n \in N$. I leave you to formulate the definition for right modules!

As with rings, I use the convention that XY denotes the sub-Abelian group of M generated by $\{xy \mid x \in X, y \in Y\}$, for any subsets $X \subseteq R, Y \subseteq M$. Then, saying that N is an R -submodule of M means simply that $N = RN$. In that case, N is itself a left R -module with the operation being the restriction of the operation on M .

Given any subset $X \subseteq M$ (a left R -module), RX is the submodule of M generated by X . So, elements of RX look like $r_1x_1 + \cdots + r_nx_n$ for $n \geq 0, r_i \in R, x_i \in X$. In particular, we say that X generates M if $M = RX$. Then, M is a *finitely generated* R -module if M is generated by some finite subset X of M , and is a *cyclic* R -module if M is generated by a single element $x \in M$. In this last case, we have that $M = Rx$ so every element of M looks like rx for some $r \in R$.

Now for first examples:

3.1.1. Any ring R itself is a left R -module, denoted ${}_R R$, the left action just being the multiplication; similarly, R is itself a right R -module, denoted R_R , the right action being just the multiplication. These are called the *left regular* and *right regular* modules, respectively.

Observe that ${}_R R$ is actually a *cyclic* left R -module, because $R = R1_R$.

The left R -submodules of ${}_R R$ are precisely the sub-Abelian groups I of R such that $RI = I$. These were called *left ideals* of R in section 2.1. Similarly, the right R -submodules of R_R are the *right ideals* of R . If R is commutative, left ideals, right ideals and (two-sided) ideals coincide, i.e. the left submodules of ${}_R R$ are the right submodules of R_R . So, in the commutative case we just talk of submodules of R , a.k.a. ideals. For instance, if R is a PID, then all submodules of the regular module R are cyclic.

3.1.2. Now take $R = \mathbb{Z}$. Any Abelian group M is a left (but we henceforth omit “left” since \mathbb{Z} is commutative) \mathbb{Z} -module, defining $nm = m + m + \cdots + m$ (n times) for $n \in \mathbb{N}, m \in M$. Conversely, given a \mathbb{Z} -module, it is in the first place an Abelian group and the \mathbb{Z} -module structure is determined uniquely by the Abelian group structure. So: *Abelian groups = \mathbb{Z} -modules*. Thus, you can think of the notion of R -module for general R as a generalization of Abelian groups!. The case of Abelian groups is always the *most important case* in the general module theory we will be developing.

3.1.3. Let R be a field F . Then, an F -module (left but we omit it) is exactly the same as a *vector space* over F ; F -submodules are the same as vector subspaces. So the notion of R -module also captures the notion of vector spaces over a field.

3.1.4. Let R be any ring. Let $M_n(R)$ be the set of all $n \times n$ matrices over R , itself a ring under matrix addition and multiplication. Let A be any left R -module and consider the space $C_n(A)$ of column vectors of height n with entries in A , viewed as an Abelian group under vector addition. Then, $C_n(A)$ is a left $M_n(R)$ -module via multiplication of a matrix by a vector. Similarly, $R_n(A)$, the space of row vectors of width n with entries in A is a right $M_n(R)$ -module.

As you might expect, the next job is to introduce homomorphisms of R -modules and discuss the isomorphism theorems. Let M, N be left (or right) R -modules. A *homomorphism* $f : M \rightarrow N$ means a morphism of Abelian groups such that $f(rm) = rf(m)$ for all $r \in R, m \in M$.

Remark. Ring theorists tend to adopt the convention of writing homomorphisms between left R -modules on the right. So they would write mf instead of $f(m)$, for instance. I'm not going to do this – which occasionally later on we will need to use the “*op*” trick mentioned earlier as a result. You need to be flexible on this issue.

We then obtain categories $R\text{-mod}$ and $\text{mod-}R$ of all *left* R -modules and all *right* R -modules respectively, morphisms being the R -module homomorphisms as just defined. If $f : M \rightarrow N$ is an R -module homomorphism, its kernel and image, defined in the same way as for Abelian groups, are automatically R -submodules of M and N respectively.

Given any R -submodule K of M , the quotient Abelian group M/K becomes an R -module if we define $r(m + K) = rm + K$ for all $r \in R, m \in M$, and this gives us the *quotient R -module* of M by the R -submodule K . The map $\pi : M \rightarrow M/K, m \mapsto m + K$ is an R -module homomorphism, the *canonical quotient map*. We have the all important:

Universal property of quotients. *Let $N \leq M$ be an R -submodule of a (left or right) R -module M , $\pi : M \rightarrow M/N$ be the canonical quotient map. Given any R -module homomorphism $f : M \rightarrow M'$ with $N \subseteq \ker f$, there exists a unique homomorphism $\bar{f} : M/N \rightarrow M'$ such that $f = \bar{f} \circ \pi$.*

Now the following results all follow as in the case of Abelian groups:

First isomorphism theorem. *Let $f : M \rightarrow M'$ be an R -module homomorphism and $N = \ker f$. Then, f factors through the quotient M/N to induce an isomorphism $\bar{f} : M/N \rightarrow \text{im } f$.*

Second isomorphism theorem. *Let $K, L \leq M$ be R -submodules of an R -module M . Then, $K/(K \cap L) \cong (K + L)/L$.*

Third isomorphism theorem. *Let $K \leq L \leq M$ be R -submodules of an R -module M . Then, L/K is an R -submodule of M/K and $M/L \cong (M/K)/(L/K)$.*

We have the lattice isomorphism theorem for submodules. We should first observe that the set of all R -submodules of a fixed left R -module M form a complete lattice: meet is given by taking intersections and join is given by taking sums. Then:

Lattice isomorphism theorem. *Let $f : M \rightarrow M'$ be an epimorphism with kernel K . Then, the map $N \mapsto f(N)$ gives an isomorphism between the lattice of R -submodules of M containing K and the lattice of R -submodules of M' .*

I conclude this basic introductory section on modules with some discussion on how you can try to understand their structure.

In general, the structure of R -modules can be very varied. The best possible case is when R is a field (or more generally when R is a *simple ring* discussed later on), in which case the fundamental theorem of vector spaces classifies R -modules up to isomorphism by their dimension.

One useful way of trying to understand R -modules for more general rings R than fields is by considering *composition factors*. By definition, a *simple* (or *irreducible*) R -module means a non-zero R -module M having no R -submodules other than M itself and (0) . We say that an R -module M has a *composition series* if there is a chain of R -submodules of M

$$M = M_0 > M_1 > \cdots > M_n = (0)$$

such that each consecutive factor M_i/M_{i-1} for $i = 1, \dots, n$ is a simple R -module. You should compare the definitions just made with the analogous definitions we made when studying groups.

Of course, a general R -module M may or may not have a composition series (see later when we discuss *Artinian modules*). But if it does, we have the analogue of the *Jordan-Hölder theorem* for modules (the proof is exactly the same as the proof we gave for groups). This asserts that two different composition series of a given R -module M have the same length and that the composition factors appearing in the two series are isomorphic (after reordering). Thus the length of a

composition series of M , and the set of isomorphism types of the composition factors appearing in the composition series, give important invariants of the module M up to isomorphism.

But even if you are lucky and you can prove that M has a composition series and can in some sense determine the simple composition factors appearing in any such composition series, the precise way the composition factors fit together to form the module M – for instance the precise structure of the lattice of submodules of M – can be very difficult to understand.

3.2 Direct products and direct sums

Let R be a ring. Throughout the section, we will only discuss *left R -modules*, though of course all the definitions and results have right module analogues. We first want to explain that the category $R\text{-mod}$ is an *additive category* in the sense of section 0.5.

Recall this means first of all that given two left R -modules M, N , the set $\text{Hom}_R(M, N)$ of all R -module homomorphisms from M to N actually has the additional structure of an Abelian group. Indeed, given homomorphisms $f, g : M \rightarrow N$, we define their sum $f + g : M \rightarrow N$ by $(f + g)(m) = f(m) + g(m)$ for all $m \in M$. This gives the operation on $\text{Hom}_R(M, N)$ making it into an Abelian group. For instance, the zero element of $\text{Hom}_R(M, N)$ is the homomorphism 0 with $0(m) = 0_N$ for all $m \in M$. Composition of homomorphisms distributes over addition.

The category $R\text{-mod}$ clearly has a *zero object*, namely, the zero module. All that is left for $R\text{-mod}$ to be an additive category is that every pair of R -modules has both a product and a coproduct. These are defined in exactly the same way as products and coproducts of Abelian groups:

- (P) The *product* of two R -modules M_1, M_2 is the Cartesian product $M_1 \times M_2$ as an Abelian group with action of $r \in R$ defined by $r(m_1, m_2) = (rm_1, rm_2)$. The R -module homomorphisms $\pi_i : M_1 \times M_2 \rightarrow M_i$ satisfying the universal property of products are just the projections, $\pi_i(m_1, m_2) = m_i$. I always refer to products of R -modules as *direct products*.
- (C) The *coproduct* of two R -modules M_1, M_2 is the same R -module $M_1 \times M_2$ as the product (bear in mind the corollary in section 0.5). But for some perverse reason, we denote it by $M_1 \oplus M_2$ in this case and write the element (m_1, m_2) instead as the sum $m_1 + m_2$. The maps $\iota_i : M_i \rightarrow M_1 \oplus M_2$ making $M_1 \oplus M_2$ into the coproduct are then just the natural inclusions. Of course, as the notation suggests, we always call $M_1 \oplus M_2$ the *direct sum* instead of the coproduct.

Actually, the category $R\text{-mod}$ is much richer than being just an additive category, as we shall see. For instance, it actually possesses *arbitrary* products and coproducts (i.e. not just of finite families of objects). So now let M_i ($i \in I$) be a possibly infinite family of left R -modules. Then their product $\prod_{i \in I} M_i$ is just their Cartesian product, with the action of R being “coordinatewise”, together with the natural projections $\pi_i : \prod_{i \in I} M_i \rightarrow M_i$. Note I always try to visualize an element of $\prod_{i \in I} M_i$ as an “infinite tuple” $m = (m_i)_{i \in I}$.

Turn now to coproducts for our family M_i ($i \in I$), which turns out to be the more useful notion in module theory. Then, their coproduct $\coprod_{i \in I} M_i$ is defined to be the R -submodule of $\prod_{i \in I} M_i$ consisting of all tuples $m = (m_i)_{i \in I}$ such that $m_i = 0$ for all but finitely many $i \in I$ (the same as for coproducts of Abelian groups in (0.3.5)). We will write an element of $\coprod_{i \in I} M_i$ not as an infinite tuple, but as a sum $\sum_{i \in I} m_i$, since all but finitely many m_i are zero. In this notation, the obvious inclusions $M_i \hookrightarrow \coprod_{i \in I} M_i$ are precisely the maps appearing in the abstract definition of coproduct. You should of course see that for non-zero modules M_i , $\coprod_{i \in I} M_i = \prod_{i \in I} M_i$ if and only if the index set I is finite. Henceforth, I denote $\coprod_{i \in I} M_i$ as $\bigoplus_{i \in I} M_i$ and call it *direct sum*.

We have now defined *direct sum* of a family of modules M_i ($i \in I$). It gives a way of building a new module out of a collection of old modules. You can call the direct sum $\bigoplus_{i \in I} M_i$ the *external* direct sum. We now want to know how to recognize when a given module M is actually $\bigoplus_{i \in I} M_i$ for some submodules M_i of M , i.e. when is M an *internal* direct sum? (The distinction is similar to the difference between external and internal semidirect products made in section 1.7.)

So now suppose we are given some R -module M and a collection of R -submodules M_i ($i \in I$) of M . The R -submodule of M generated by the M_i is just their sum $\sum_{i \in I} M_i$, meaning the set of all elements of M which can be written as $\sum_{i \in I} m_i$ for $m_i \in M_i$ with all but finitely many being zero (so that the possibly infinite sum has meaning). Now we have the notions of *span* and *linear independence*. If

$$M = \sum_{i \in I} M_i$$

then we say that the M_i *span* M . If the property

$$\sum_{i \in I} m_i = 0 \quad \Rightarrow \quad m_i = 0 \quad \forall i \in I$$

holds whenever we are given elements $m_i \in M_i$ with all but finitely many being zero, we say that $\sum_{i \in I} M_i$ is *direct*, and the submodules M_i are called *linearly independent*.

Exercise. The following properties are equivalent:

- (1) $\sum_{i \in I} M_i$ is direct;
- (2) $M_i \cap \left(\sum_{j \in I - \{i\}} M_j \right) = (0)$ for all $i \in I$;
- (3) any $m \in \sum_{i \in I} M_i$ can be written as $\sum_{i \in I} m_i$ for *unique* elements $m_i \in M_i$, all but finitely many being zero.

If the M_i span M , so $\sum_{i \in I} M_i = M$, and they are linearly independent, so $\sum_{i \in I} M_i$ is direct, then we write

$$M = \bigoplus_{i \in I} M_i$$

as say that M is the *internal direct sum* of the submodules M_i . You can check that if M is the internal direct sum of the submodules M_i , then the unique map $\bigoplus_{i \in I} M_i \rightarrow M$ induced by the inclusions $M_i \hookrightarrow M$ according to the universal property of coproducts is in fact an isomorphism, so that M is *isomorphic* to the external direct sum of the M_i . This should explain the language.

Let me end this section with two basic definitions used when discussing direct sums. An R -module M is called *decomposable* if M is the internal direct sum of two non-zero proper submodules M_1, M_2 of M . Otherwise, M is *indecomposable*. An R -submodule N of an R -module M is called a *summand* of M if there exists another R -submodule C of M such that $M = N \oplus C$. This submodule C is then called a *complement* to N in M .

Example. Consider the \mathbb{Z} -module $M = \mathbb{Z}_2 \oplus \mathbb{Z}_2$ (otherwise known as the Klein 4 group!) It has *three* \mathbb{Z} -submodules (otherwise known as subgroups!) of order 2, namely $A = \{(0, 0), (1, 0)\}$, $B = \{(0, 0), (0, 1)\}$ and $C = \{(0, 0), (1, 1)\}$. We therefore have that

$$M = A \oplus B = A \oplus C = B \oplus C.$$

Thus the module M is *decomposable*, but there are *many* ways of decomposing it as a direct sum of two non-zero proper submodules. The submodule A of M is a *summand* of M , while both B and C are *complements* to A in M . Thus a summand can in general have *many different* complements. Corresponding to the three subgroups of M of order 2, we obtain three different *composition series* of M , namely

$$M \supset A \supset (0), \quad M \supset B \supset (0), \quad M \supset C \supset (0).$$

These are composition series because all factors are isomorphic to \mathbb{Z}_2 which is simple. Thus M in this case has exactly *three different composition series*.

Another example. This time consider the \mathbb{Z} -module $M = \mathbb{Z}_4 = \{0, 1, 2, 3\}$ (a.k.a. the cyclic group of order 4). It contains a *unique* \mathbb{Z} -submodule $A = \{0, 2\}$ of order 2 (hence $\cong \mathbb{Z}_2$), and the quotient M/A is also of order 2 (hence $\cong \mathbb{Z}_2$). In this case, A is *not* a summand of M , for it cannot possibly possess a complement. The chain

$$M \supset A \supset (0)$$

is a *composition series* of M in this case, and this is the *unique* composition series of M .

3.3 Free modules

Let R be a ring, and continue to work only with *left* R -modules. We next introduce the notion of *free R -module*. You should compare the definitions with that of *free groups* (section 1.4), but also keep in mind that *free modules are the next best thing to vector spaces!*

Let F be a (left) R -module and $X \subseteq F$ be a subset. We say that F is *free on X* if the following universal property holds:

- (F) given any R -module M and a set map $f : X \rightarrow M$, there exists a unique R -module homomorphism $\bar{f} : F \rightarrow M$ extending f (i.e. such that $\bar{f}(x) = f(x)$ for all $x \in X$).

If there exists an R -module F that is free on the set X , then certainly by the usual argument (e.g. Lemma 1.4.1) F is unique up to a canonical isomorphism. So we just talk about *the* free module on X . Such a module always exists because:

Existence of free modules. *For any set X , there is a left R -module that is free on X .*

Proof. For $x \in X$, let R_x be a copy of the left regular R -module ${}_R R$, denoting the element in R_x corresponding to the 1_R by x . Consider

$$F = \bigoplus_{x \in X} R_x$$

and view X as a subset of F in the obvious way. We claim that F is free on X . Take a set map $f : X \rightarrow M$ to an R -module M . Now, every element of $R_x \cong R$ can be written as rx for a unique $r \in R$. Hence, every element of F looks like $\sum_{x \in X} r_x x$ for uniquely determined coefficients $r_x \in R$, all but finitely many r_x 's being zero. So if we are trying to extend f to an R -module homomorphism, there is no option but to define

$$\bar{f} \left(\sum_{x \in X} r_x x \right) = \sum_{x \in X} r_x f(x).$$

Moreover, as you easily check, this equation really does define an R -module homomorphism. \square

By definition, a subset X of an R -module M is called a *basis* of M if X is *linearly independent*, meaning

$$\sum_{x \in X} r_x x = 0 \quad \Rightarrow \quad r_x = 0 \quad \forall x \in X$$

whenever $r_x \in R$ are coefficients with all but finitely many being zero, and moreover X *spans* or *generates* M , meaning that $M = RX$. Now let F be the free R -module on X . In the course of the above existence proof, we saw that every element of F can be written as $\sum_{x \in X} r_x x$ for unique coefficients $r_x \in R$, all but finitely many being zero. In other words, if F is free on X , then X is a *basis* for F .

You should be able to show conversely (copying the proof of the existence theorem) that if F is any R -module such that $X \subseteq F$ is a basis, then F is free on X . In other words, an R -module M is the *free R -module on the subset $X \subseteq M$* if and only if X is a *basis* for M . Thus, the free

R -modules are exactly the R -modules possessing a basis. In that case, the module is isomorphic to a direct sum of copies of the regular R -module ${}_R R$, the number of such copies that suffices being the cardinality of a basis of M .

Free R -modules are extremely important, mainly because:

3.3.1. Theorem. *Every (finitely generated) R -module M is the quotient of a (finitely generated) free module.*

Proof. Let F be the free module on the set M . The set map $M \rightarrow M$ extends to a unique R -module homomorphism $F \rightarrow M$ which is clearly surjective. Thus, M is a quotient of F . In case M is finitely generated, the argument is the same, but one takes F to be the free module on some finite generating set of M instead. \square

It is worth pointing out a special case of this argument. Recall an R -module M is *cyclic* if $M = Rx$ for some $x \in M$. In that case, observing that ${}_R R$ is free on 1_R , the set map $1_R \mapsto x$ extends to a unique R -module homomorphism $f : R \rightarrow M, r \mapsto rx$. It is surjective since $M = Rx$. Hence, $M \cong R/\ker f$. This shows: *any cyclic R -module is a quotient of the regular module ${}_R R$ by a left ideal.*

To motivate the next result, suppose that R is a *field*, when R -modules are just the same as vector spaces. The fundamental theorem of linear algebra asserts that every vector space possesses a basis and moreover that any two bases have the same cardinality. This shows in particular that *every module over a field is free*, having a basis. But now you need to be very careful! The property (of vector spaces) that any two bases have the same cardinality is a special property that can fail for modules over sufficiently bad rings. We do at least have:

Rank for free modules over commutative rings. *Let R be a commutative ring and M be a free R -module. Then, any two bases of M have the same cardinality.*

Proof. Let I be a maximal ideal of R , so that $F = R/I$ is a field. Note IM is an R -submodule of M , so M/IM is a quotient module. Moreover, for any $a \in I$, $aM \subseteq IM$ so that a acts as zero on the quotient module M/IM . Hence, we obtain a well-defined action of F on M/IM by setting $(r+I)(m+IM) = rm+IM$ for any $r \in R, m \in M$. In other words, the quotient module M/IM can actually be viewed as a vector space over the field $F = R/I$.

Now let X be a basis for M . Consider the quotient map $\pi : M \rightarrow M/IM$. Suppose that $r_1\pi(x_1) + \cdots + r_n\pi(x_n) = 0$ in M/IM for some $r_i \in R$ and distinct $x_i \in X$. Then, $r_1x_1 + \cdots + r_nx_n \in IM$ so can be written (since X spans M) as $\sum_j a_j x_j$ for some $a_j \in I, x_j \in X$. So,

$$r_1x_1 + \cdots + r_nx_n = \sum_j a_j x_j$$

which implies using the linear independence of X that in fact r_1, \dots, r_n are all elements of the ideal I , i.e. their images in the field F are zero. We have shown: the elements of the set $\pi(X)$ are distinct and form an F -basis for the vector space M/IM .

Hence, the cardinality of the set X is the same as the dimension of the vector space M/IM . Now the fundamental theorem of linear algebra gives that any two bases of M have the same cardinality, since any two bases of the vector space M/IM do. \square

In view of the theorem, if R is a commutative ring, we can define the notion of *rank* of a free R -module, namely, the cardinality of any basis. Thus, if R is a field, rank is exactly what is more usually called dimension. For more general rings than commutative rings, there may or may not be a well-defined notion of rank of free modules, depending on which ring you are talking about. About the only thing we can say about the cardinality of a basis of a free module in full generality is:

3.3.2. Lemma. *Let F be a free R -module. The following are equivalent:*

- (1) F is a finitely generated R -module;
- (2) F has a finite basis;
- (3) every basis of F is finite.

Proof. Clearly if F has a finite basis, it is finitely generated. Conversely, suppose F is generated by finitely many elements f_1, \dots, f_n and let X be any basis for F . We can write $f_i = \sum_{x \in X} a_{i,x}x$ for each $i = 1, \dots, n$, where all but finitely many of the coefficients $a_{i,x} \in R$ are zero for each fixed i . But then $X' = \{x \in X \mid a_{i,x} \neq 0 \text{ for some } i = 1, \dots, n\}$ is finite and every generator f_1, \dots, f_n lies in RX' . Hence, $M = RX'$, so since X is linearly independent we must actually have that $X = X'$. This shows that X is finite. \square

3.4 Elementary divisors

We introduced free modules in the previous section, and also the rank of a free module over a commutative ring. Basing your intuition on vector spaces, it is natural to ask questions like

- Are all submodules of a free module F free?
- For commutative rings, is the rank of a free submodule of F less than or equal to the rank of F ?

Unfortunately, the answer to both questions is in general *no* (so the analogy with vector spaces is not that good).

For the remainder of the section, R will denote a PID. In this case only, we can develop a reasonable theory of submodules of free modules.

3.4.1. Theorem. *Let F be a finitely generated free module over a PID R , and $N \leq F$ be an R -submodule. Then, F is also free and $\text{rank } N \leq \text{rank } F$.*

(Actually, this theorem is true even if F is not finitely generated, see Hungerford IV.6.)

Proof. Let x_1, \dots, x_n be a basis for F and $N \leq F$. Set $F_0 = (0)$, $F_i = Rx_1 + \dots + Rx_i$ and $N_i = N \cap F_i$. We prove by induction on $i = 0, 1, \dots, n$ that $N_i \leq F_i$ is free of rank $\leq i$, the base case $i = 0$ being trivial.

Now let $i \geq 1$ and consider the induction step. If $N_i = N_{i-1}$, there is nothing to prove. Otherwise, let

$$A = \{a \in R \mid \text{there exists } x \in F_{i-1} \text{ such that } x + ax_i \in N_i\}.$$

Then, A is an ideal of R , so $A = (b)$ for some $b \in R$ as R is a PID. Moreover, $b \neq 0$ as $N_i \neq N_{i-1}$. As $b \in A$, there is some $y \in F_{i-1}$ such that $y + bx_i \in N_i$. Let $z = y + bx_i$. We claim $N_i = N_{i-1} \oplus Rz$, so that N_i is free of rank one more than N_{i-1} .

Well, take any $f \in N_i$. Then, $f = x + cx_i$ for some $x \in F_{i-1}, c \in R$. So $c \in A = (b)$, so $c = bd$ for some $d \in R$ and $f = x + dbx_i$. But then, $f - dz = x - dy \in F_{i-1} \cap N_i = N_{i-1}$. This shows that $N_i = N_{i-1} + Rz$. Finally, to show that $N_{i-1} \cap Rz = (0)$, suppose $rz \in N_{i-1}$. Then, $rz = ry + rbx_i \in F_{i-1}$, hence $rbx_i \in F_{i-1} \cap Rx_i = (0)$, so $rb = 0$ whence $r = 0$ as R is an integral domain. \square

In order to obtain more precise information about submodules of free modules over PIDs, we first need to discuss a rather different topic. Let $M_{s,t}(R)$ denote the set of all $s \times t$ matrices with entries in the PID R . Call two matrices $A, B \in M_{s,t}(R)$ *equivalent* if there exist invertible square matrices P and Q such that $B = PAQ$. Note that “equivalence” is indeed an equivalence relation on $M_{s,t}(R)$. I will assume you are familiar with basic notions of matrices (over an arbitrary commutative ring), referring you to section 6.1 for some background. For instance, a square matrix with entries in R is invertible if and only if its determinant (defined by “Laplace expansion” along some row or column) is a unit in R .

The main job now is to prove:

Canonical form for matrices over PIDs. *If R is a PID then any matrix $A \in M_{s,t}(R)$ is equivalent to a matrix of the form $\text{diag}(d_1, \dots, d_u)$ (where $u = \min(s, t)$) with $d_1 | d_2 | \dots | d_u$ in R . Moreover, the diagonal entries d_1, \dots, d_u are unique up to associates.*

Proof. For the proof, we need the *elementary row and column operations*. Let me remind you of these, for a matrix $A \in M_{s,t}(R)$:

- (O1) Swap two rows (or columns) of A ;
- (O2) Scale any row (or column) of A by a unit in R ;
- (O3) Add a multiple of one row (or column) to another row (or column).

The point is that all of the elementary row and column operations can be performed on the matrix A by pre- or post-multiplying by an invertible matrix. Thus, they do not change the equivalence class of the matrix we are considering.

Now let me explain the *algorithm* to reduce an arbitrary matrix $A \in M_{r,s}(R)$ to the desired canonical form. To guarantee that the algorithm eventually terminates, we introduce the notion of *length* of the matrix A . This is defined to be the the number of primes appearing in the prime factorization of the leading entry $a_{1,1}$ of a , or 0 in case $a_{1,1}$ is zero or a unit.

Now, if $A = 0$, there is nothing to do. Else, some entry of the matrix A is non-zero, and swapping rows and columns, we can assume that $a_{1,1}$ is a non-zero entry of A of minimal length. Now there are three cases:

Case one. $a_{1,1} \nmid a_{1,j}$ for some $j > 1$. Without loss of generality, assume $a_{1,1} \nmid a_{1,2}$. Let d be a GCD of $a_{1,1}$ and $a_{1,2}$, so $a_{1,1} = dy_1, a_{1,2} = dy_2$ for y_1, y_2 coprime. Since $a_{1,1} \nmid a_{1,2}$, y_1 is not a unit, so $\lambda(y_1) \geq 1$, so $\lambda(d) < \lambda(a_{1,1})$. Write $1 = x_1 y_1 + x_2 y_2$ for $x_1, x_2 \in R$. Then

$$Q = \left[\begin{array}{cc|ccc} x_1 & -y_2 & 0 & \dots & 0 \\ x_2 & y_1 & 0 & \dots & 0 \\ \hline 0 & 0 & & & \\ \vdots & \vdots & & I & \\ 0 & 0 & & & \end{array} \right]$$

is invertible and AQ has leading term d . Since $\lambda(d) < \lambda(a_{1,1})$, the length of A has gone down, and we can now repeat the algorithm from the beginning.

Case two. $a_{1,1} \neq 0$ and $a_{1,1} \nmid a_{i,1}$ for some $i > 1$. This is proceeds in the same way as case one, working with columns not rows.

Case three. $a_{1,1}$ divides every entry in both the first row and the first column of A . Then, we can use elementary row and column operations to reduce the matrix A to the form

$$\left[\begin{array}{c|ccc} a_{1,1} & 0 & \dots & 0 \\ \hline 0 & & & \\ \vdots & & B & \\ 0 & & & \end{array} \right]$$

Now apply the algorithm recursively to put the matrix B into the canonical diagonal form. Then, if $a_{1,1}$ divides every entry of the (now diagonal) matrix B , we are done. Otherwise, $a_{1,1} \nmid b_{i,i}$ for some i . In this case, add the i th column of B to the first column of A and apply step one to reduce the length of $a_{1,1}$ and continue.

This algorithm gives existence: every $A \in M_{s,t}(R)$ is equivalent to a diagonal matrix in canonical form. It just remains to prove uniqueness. So now suppose $\text{diag}(d_1, \dots, d_u)$ and $\text{diag}(d'_1, \dots, d'_u)$ are equivalent matrices in $M_{s,t}(R)$, where $d_1 | d_2 | \dots | d_u$ and $d'_1 | \dots | d'_u$. We need to prove d_i ass d'_i for each i . To do this, define $J_i(A)$ to be the ideal of R generated by the determinants of all $i \times i$

minors (“sub-matrices”) of a matrix A . The point is that $J_i(A)$ depends only on the *equivalence class* of the matrix A . Clearly, $J_i(\text{diag}(d_1, \dots, d_u)) = (d_1 d_2 \dots d_i)$ if $d_1 | d_2 | \dots | d_u$. Hence,

$$(d_1) = (d'_1), (d_1 d_2) = (d'_1 d'_2), \quad \dots, \quad (d_1 \dots d_u) = (d'_1 \dots d'_u).$$

This implies d_i ass d'_i for each i . \square

The result just proved shows that to a matrix $A \in M_{s,t}(R)$ you can associate a sequence $d_1 | d_2 | \dots | d_u$ of R , unique up to associates, called the *invariant factor sequence* of the matrix A . Notice in the special case R is a field, each d_i is either 0 or 1 and the invariant factor sequence takes the form $1 | \dots | 1 | 0 | \dots | 0$; the number of 1’s simply records the *rank* of the original matrix, which should be familiar to you from linear algebra. Thus the invariant factor sequence is a generalization to matrices over an arbitrary PID of the notion of rank of a matrix over a field.

Now we go back to studying submodules of free modules:

Structure of submodules of free modules over PIDs. *Let R be a PID, F be a free R -module of rank s and N be an R -submodule of F of rank $t \leq s$. Then, there exists a basis f_1, \dots, f_s for F and elements $d_1, \dots, d_t \in R$ such that $d_1 | d_2 | \dots | d_t$ and $d_1 f_1, d_2 f_2, \dots, d_t f_t$ is a basis for N . Moreover, the elements d_1, \dots, d_t are unique in the sense that if we have another basis f'_1, \dots, f'_s for F and elements $d'_1, \dots, d'_t \in R$ such that $d'_1 | d'_2 | \dots | d'_t$ and $d'_1 f'_1, d'_2 f'_2, \dots, d'_t f'_t$ is a basis for N , then d_i ass d'_i for each i .*

Proof. We may as well assume that $N \neq (0)$ since that case is trivial. Let $\{f_1, \dots, f_s\}$ be a basis for F and $\{n_1, \dots, n_t\}$ ($t \leq s$) be a basis for N , applying Theorem 3.4.1. Write

$$n_j = \sum_{i=1}^s a_{i,j} f_i$$

for $j = 1, \dots, t$. So, $A = (a_{i,j})$ is an $s \times t$ matrix. Therefore, applying the canonical form for matrices over PIDs, we can find an invertible $s \times s$ matrix P and an invertible $t \times t$ matrix Q such that $P^{-1} A Q = D = \text{diag}(d_1, \dots, d_t)$ for elements $d_1 | d_2 | \dots | d_t$ in R .

Now let $n'_j = \sum_{i=1}^t q_{i,j} n_i$, $f'_j = \sum_{i=1}^s p_{i,j} f_i$. Since the matrices P and Q are invertible, these also give bases for N and F respectively. Moreover,

$$n'_j = \sum_{i=1}^t q_{i,j} n_i = \sum_{i=1}^t \sum_{k=1}^s a_{k,i} q_{i,j} f_k = \sum_{i=1}^s d_j p_{i,j} f_i = d_j f'_j$$

since $AP = QD$. This gives us the required bases for F and N .

It remains to prove uniqueness. So take the two bases f_1, \dots, f_s and f'_1, \dots, f'_s as in the statement of the second part of the theorem. Let P be the change of basis matrix from f_1, \dots, f_s to f'_1, \dots, f'_s and Q be the change of basis matrix from $d_1 f_1, \dots, d_t f_t$ to $d'_1 f'_1, \dots, d'_t f'_t$. Then $P \text{diag}(d_1, \dots, d_t) = \text{diag}(d'_1, \dots, d'_t) Q$, so the two $s \times t$ diagonal matrices are equivalent, so d_i ass d'_i by the uniqueness of the invariant factors of a matrix proved above. \square

3.5 Structure theorems for modules over PIDs

We now apply the results of the previous section to prove the structure theorems for finitely generated modules over PIDs.

Recall that a cyclic R -module is an R -module M generated by a single element m ; in that case, the R -module homomorphism $\pi : R \rightarrow M, r \mapsto rm$ is surjective so $M \cong R / \ker \pi$. We call $\ker \pi$ the *order ideal* of the cyclic module M , since it determines M uniquely up to isomorphism. Note $\ker \pi$ is the *annihilator in R of M* , that is, $\{r \in R \mid rM = (0)\}$. If R is even a PID, then $\ker \pi = (f)$ for some $f \in R$, unique up to associates, and then we call f the *order* of M .

For example, if $R = \mathbb{Z}$ then the cyclic \mathbb{Z} -modules are exactly the Abelian groups \mathbb{Z} , of order 0, or \mathbb{Z}_n of order $\pm n$ (note the order of a cyclic module is only defined up to associates). In other words, the order of a cyclic \mathbb{Z} -module is simply its order as an Abelian group, or 0 if it is infinite. We will see another important example, when $R = F[X]$ for F a field, in section 6.2 when we study normal forms for linear transformations.

We will need the following elementary but technical lemma.

Technical lemma. *Let R be a PID and F be a free R -module of rank s . Suppose $\pi : F \rightarrow M$ be a surjective homomorphism such that $M = M_1 \oplus \cdots \oplus M_t$ for some $t \geq s$ and non-zero cyclic modules M_i of order d_i , where $d_1 | d_2 | \cdots | d_t$. Then, there exists a basis f_1, \dots, f_s for F such that $\pi(f_i)$ generates M_{i+t-s} for each $i = 1, \dots, s$.*

Proof. Proceed by induction on t , the case $t = 0$ being immediate (since then $M = (0)$). Since π is onto we can find $0 \neq f \in F$ with $\pi(f)$ a generator of the non-zero module M_t . Applying the submodule theorem from the previous section to the submodule $N = Rf$ of F , we get a basis f_1, \dots, f_s of F and $a \in R^*$ such that af_s is a basis of N . Then, $a\pi(f_s)$ generates M_t . We claim that in fact, $\pi(f_s)$ generates M_t .

To prove this, let h be the greatest common divisor of a and d_t . Write $a = hb, d_t = hd$ with b, d coprime. Then, $da\pi(f_s) = dhb\pi(f_s) = bd_t\pi(f_s) = 0$, the last equality being true since d_t even annihilates all of M . So, d annihilates $a\pi(f_s)$ hence annihilates M_t . Since d_t generates the annihilator of M_t , we get that $d_t | d$, whence h is a unit so a and d_t are in fact coprime. Write $1 = au + d_tv$. Then, $\pi(f_s) = (au + d_tv)\pi(f_s) = au\pi(f_s)$, so $R\pi(f_s) = Rua\pi(f_s) \subseteq Ra\pi(f_s) \subseteq R\pi(f_s)$. This shows that $R\pi(f_s) = Ra\pi(f_s) = M_t$, and $\pi(f_s)$ generates M_t as claimed.

Let $M' = M_1 \oplus \cdots \oplus M_{t-1}$. Now, for $i = 1, \dots, s-1$, we have that $\pi(f_i) = m_i + c_i\pi(f_s)$ for some $m_i \in M'$ and $c_i \in R$. So, setting $f'_i = f_i - c_if_s$ for each $i = 1, \dots, s-1$, we now have a basis $f'_1, \dots, f'_{s-1}, f_s$ for F and f_s is the same as before, so $\pi(f_s)$ generates M_t ; but now, we have ensured that $\pi(f'_i) \in M'$ for each $i = 1, \dots, s-1$. Set $F' = Rf'_1 \oplus \cdots \oplus Rf'_{s-1}$. Then, F' is free of rank $s-1$ and the restriction of π maps F' surjectively onto M' . By the induction hypothesis, we can find a basis f''_1, \dots, f''_{s-1} for F' so that $\pi(f''_i)$ generates M_{i+t-s} for each $i = 1, \dots, s-1$. Now we are done: the basis $f''_1, \dots, f''_{s-1}, f_s$ for F does the job. \square

The main result is the following:

Structure theorem. *Let M be a finitely generated R -module, where R is a PID. Then, M can be decomposed as an internal direct sum as*

$$M = M_1 \oplus \cdots \oplus M_s$$

where M_i is a non-zero cyclic submodule of order d_i and $d_1 | \cdots | d_s$ in R . Moreover, s and the orders d_1, \dots, d_s are uniquely determined up to associates. In other words, if

$$M = M'_1 \oplus \cdots \oplus M'_t$$

for non-zero cyclic submodules M'_i of order d'_i such that $d'_1 | \cdots | d'_t$, then $t = s$ and $d_i \text{ ass } d'_i$.

Proof. Existence. Since M is finitely generated, we can find generators m_1, \dots, m_s for M , where s is taken to be *minimal*. Let F be the free R -module on $\{x_1, \dots, x_s\}$. Then, M is a quotient of F under the unique R -module homomorphism π sending $x_i \mapsto m_i$. Let $K = \ker \pi$. Applying the structure theorem for submodules of free modules over PIDs, we can find a basis f_1, \dots, f_s for F and elements $d_1 | d_2 | \cdots | d_s$ in R (where some of the d_i are possibly zero) so that the non-zero elements from d_1f_1, \dots, d_sf_s form a basis for K . If any d_i is a unit, then $f_i \in \ker \pi$ so that just $\pi(f_1), \dots, \pi(f_{i-1}), \pi(f_{i+1}), \dots, \pi(f_s)$ generate M , contradicting the minimality of s . So no d_i is a unit.

Now, let $M_i = \pi(Rf_i)$. Observe

$$M_i \cong Rf_i / (K \cap Rf_i) \cong Rf_i / Rd_if_i \cong R/(d_i).$$

Hence, M_i is cyclic of order d_i , so non-zero since d_i is not a unit. Moreover, since f_1, \dots, f_s generate F , $\pi(f_1), \dots, \pi(f_s)$ generate M hence $M = M_1 + \dots + M_s$. The sum is direct because if

$$m_1 + \dots + m_s = 0$$

for $m_i = \pi(r_i f_i) \in M_i$, then $r_1 f_1 + \dots + r_s f_s \in K$. Since $d_1 f_1, \dots, d_s f_s$ is a basis for K , we deduce that $d_i | r_i$ for each i so that each $m_i = 0$.

Uniqueness. Now take another such decomposition $M = M'_1 \oplus \dots \oplus M'_t$ with M'_i non-zero and cyclic of order d'_i . Observe that the generators of all the M'_i generate M , so $t \geq s$ by the minimality of the choice of s in the existence proof. Now we have the surjection $\pi : F \rightarrow M$ defined in the existence proof. Apply the technical lemma to obtain a basis f'_1, \dots, f'_s for F such that $\pi(f'_i)$ generates M'_{i+t-s} for each $i = 1, \dots, s$. But the f'_i generate F , so the $\pi(f'_i)$ must generate M since π is surjective. This proves that in fact, $t = s$ and $\pi(f'_i)$ generates M'_i for each $i = 1, \dots, s$. Now, M'_i is cyclic of order d'_i , which implies that $d'_1 f'_1, \dots, d'_s f'_s$ give a basis for the kernel K of π . But now you get that d'_i ass d_i from the uniqueness in the submodule theorem from the previous section. \square

The sequence $d_1 | d_2 | \dots | d_s$ appearing in the theorem is called the *invariant factor sequence* of the module M . It determines M uniquely up to isomorphism. Thus, the structure theorem is a *classification* of the finitely generated modules over a PID by their invariant factor sequences.

Now let R be any commutative ring and M be an R -module. We call an element $m \in M$ a *torsion element* if $rm = 0$ for some $r \in R^*$. Then, M is called *torsion* if all its elements are torsion elements, and M is *torsion-free* if it has no non-zero torsion elements. For example a torsion \mathbb{Z} -module means an Abelian group all of whose elements have finite order; all vector spaces over a field are torsion free. In general, for an arbitrary R -module M , we let

$$M_t = \{m \in M \mid m \text{ is a torsion element}\}.$$

Then:

3.5.1. Lemma. *Let R be an integral domain, M an R -module. Then, M_t is an R -submodule of M , and M/M_t is torsion-free.*

Proof. Take $m_1, m_2 \in M_t$. Then, there are $r_1, r_2 \in R^*$ such that $r_i m_i = 0$. Set $r = r_1 r_2$, non-zero since R is an integral domain. Then, $r(m_1 + m_2) = 0$ so $m_1 + m_2 \in M_t$. The rest of the proof is similar... \square

It is obvious that if R is an integral domain, then all free R -modules are torsion-free. In the case of finitely generated modules over PIDs, the converse is true, so that the local property “torsion-free” is equivalent to the global property “free”.

Torsion free \Rightarrow free for f.g. modules over PIDs. *Let R be a PID and M be a finitely generated R -module. Then, there is a (not necessarily unique!) free R -submodule M' of M such that $M = M_t \oplus M'$. In particular, M is free if and only if M is torsion-free.*

Proof. Apply the structure theorem to write $M = M_1 \oplus \dots \oplus M_s$ where M_i is cyclic of order d_i and $d_1 | \dots | d_s$. Let

$$M' = \bigoplus_{i \text{ s.t. } d_i=0} M_i.$$

Then, M' is free since it is a direct sum of copies of R . In particular, $M_t \cap M' = \{0\}$. On the other hand, each M_i with $d_i \neq 0$ is contained in M_t . This shows that $M = M_t + M'$ as required. The second statement in the theorem follows immediately since M is torsion-free if and only if $M_t = (0)$ which is if and only if $M = M'$. \square

Note: for more general integral domains, the torsion submodule M_t of M will *not* in general have a complement in M . For an example of a torsion-free \mathbb{Z} -module that is *not* free, take the Abelian group \mathbb{Q} (which is *not* finitely generated!).

The number of summands s in the decomposition of M in the statement of the structure theorem is the *smallest possible* number such that M can be written as a direct sum of s cyclic submodules. We turn to discussing the other extreme, when M is written as a direct sum of as many submodules as possible: the *primary decomposition* of M .

3.5.2. Lemma. *Let R be a PID and M be a cyclic R -module of order p^r where p is prime. Then, the only R -submodules of M are the following:*

$$(0) = p^r M \subset p^{r-1} M \subset \cdots \subset p M \subset M$$

and $p^i M / p^{i+1} M \cong R/(p)$. In particular, M is indecomposable.

Proof. Since $M \cong R/(p^r)$, the lattice of submodules of M is isomorphic to the lattice of ideals of R containing p^r , which in turn (since R is a PID) is isomorphic to the lattice of divisors of p^r . The result follows since p is prime. \square

Primary decomposition theorem. *Let R be a PID and M be a finitely generated torsion R -module. Then, M can be written as*

$$M = M_1 \oplus \cdots \oplus M_t$$

for cyclic R -modules M_i of order $p_i^{n_i}$, where the p_i are (not necessarily distinct) primes. Moreover, given another such decomposition

$$M = M'_1 \oplus \cdots \oplus M'_{t'}$$

with M'_i cyclic of prime power order $q_i^{m_i}$, we have that $t' = t$ and (after reordering) $q_i^{m_i} = p_i^{n_i}$ for each i .

Proof. Existence. In view of the structure theorem, we just need to be able to decompose a single cyclic module of non-zero order d into a direct sum of cyclic modules of prime power order. Well, write $d = p_1^{r_1} \cdots p_a^{r_a}$ as a product of pairwise-coprime prime powers. Then apply the Chinese remainder theorem (section 2.2) to obtain

$$R/(d) \cong R/(p_1^{r_1}) \oplus \cdots \oplus R/(p_a^{r_a}).$$

Uniqueness. You need to convince yourself (using the Chinese remainder theorem) that knowing the prime power orders $p_i^{n_i}$ with their multiplicities in a primary decomposition is exactly equivalent to knowing the invariant factor sequence $d_1 | d_2 | \cdots | d_s$ of M . Then the conclusion follows from the uniqueness of the invariant factor sequence from the structure theorem. \square

The primary decomposition theorem and Lemma 3.5.2 essentially give complete information about the submodule structure of a finitely generated module over a PID, which is remarkable. In particular, you should now be able to determine the *composition factors* of such an M . Note also that the modules M_i appearing in the primary decomposition are *indecomposable*.