Chapter 4

Module Fundamentals

4.1 Modules and Algebras

4.1.1 Definitions and Comments

A vector space M over a field R is a set of objects called vectors, which can be added, subtracted and multiplied by scalars (members of the underlying field). Thus M is an abelian group under addition, and for each $r \in R$ and $x \in M$ we have an element $rx \in M$. Scalar multiplication is distributive and associative, and the multiplicative identity of the field acts as an identity on vectors. Formally,

$$r(x+y) = rx + ry; \quad (r+s)x = rx + sx; \quad r(sx) = (rs)x; \quad 1x = x$$

for all $x, y \in M$ and $r, s \in R$. A module is just a vector space over a ring. The formal definition is exactly as above, but we relax the requirement that R be a field, and instead allow an arbitrary ring. We have written the product rx with the scalar r on the left, and technically we get a *left R-module* over the ring R. The axioms of a *right R-module* are

$$(x+y)r = xr + yr; \quad x(r+s) = xr + xs; \quad (xs)r = x(sr), \quad x1 = x.$$

"Module" will always mean left module unless stated otherwise. Most of the time, there is no reason to switch the scalars from one side to the other (especially if the underlying ring is commutative). But there are cases where we must be very careful to distinguish between left and right modules (see Example 6 of (4.1.3)).

4.1.2 Some Basic Properties of Modules

Let M be an R-module. The technique given for rings in (2.1.1) can be applied to establish the following results, which hold for any $x \in M$ and $r \in R$. We distinguish the zero vector 0_M from the zero scalar 0_R.

(1) $r0_M = 0_M$ $[r0_M = r(0_M + 0_M) = r0_M + r0_M]$

(2) $0_R x = 0_M$ $[0_R x = (0_R + 0_R)x = 0_R x + 0_R x]$

(3) $(-r)x = r(-x) = -(rx)$ [as in (2) of (2.1.1) with a replaced by r and b by x]

(4) If R is a field, or more generally a division ring, then $rx = 0_M$ implies that either $r = 0_R$ or $x = 0_M$. [If $r \neq 0$, multiply the equation $rx = 0_M$ by r^{-1}.]

4.1.3 Examples

1. If M is a vector space over the field R, then M is an R-module.

2. Any ring R is a module over itself. Rather than check all the formal requirements, think intuitively: Elements of a ring can be added and subtracted, and we can certainly multiply $r \in R$ by $x \in R$, and the usual rules of arithmetic apply.

3. If R is any ring, then R^n, the set of all n-tuples with components in R, is an R-module, with the usual definitions of addition and scalar multiplication (as in Euclidean space, e.g., $r(x_1, \dots, x_n) = (rx_1, \dots, rx_n)$, etc).

4. Let $M = M_{mn}(R)$ be the set of all $m \times n$ matrices with entries in R. Then M is an R-module, where addition is ordinary matrix addition, and multiplication of the scalar c by the matrix A means multiplication of each entry of A by c.

5. Every abelian group A is a $\mathbb{Z}$-module. Addition and subtraction is carried out according to the group structure of A; the key point is that we can multiply $x \in A$ by the integer n. If $n > 0$, then $nx = x + x + \cdots + x$ (n times); if $n < 0$, then $nx = -x - x - \cdots - x$ ($|n|$ times).

In all of these examples, we can switch from left to right modules by a simple notational change. This is definitely not the case in the next example.

6. Let I be a left ideal of the ring R; then I is a left R-module. (If $x \in I$ and $r \in R$ then rx (but not necessarily xr) belongs to I.) Similarly, a right ideal is a right R-module, and a two-sided ideal is both a left and a right R-module.

An R-module M permits addition of vectors and scalar multiplication. If multiplication of vectors is allowed, we have an R-algebra.

4.1.4 Definitions and Comments

Let R be a commutative ring. We say that M is an *algebra over* R, or that M is an R-*algebra*, if M is an R-module that is also a ring (not necessarily commutative), and the ring and module operations are compatible, i.e.,

$$r(xy) = (rx)y = x(ry) \text{ for all } x, y \in M \text{ and } r \in R.$$

4.1.5 Examples

1. Every commutative ring R is an algebra over itself (see Example 2 of (4.1.3)).

2. An arbitrary ring R is always a $\mathbb{Z}$-algebra (see Example 5 of (4.1.3)).

3. If R is a commutative ring, then $M_n(R)$, the set of all $n \times n$ matrices with entries in R, is an R-algebra (see Example 4 of (4.1.3)).

4. If R is a commutative ring, then the polynomial ring $R[X]$ is an R-algebra, as is the ring $R[[X]]$ of formal power series; see Examples 5 and 6 of (2.1.3). The compatibility condition is satisfied because an element of R can be regarded as a polynomial of degree 0.

5. If E/F is a field extension, then E is an algebra over F. This continues to hold if E is a division ring, and in this case we say that E is a *division algebra* over F.

To check that a subset S of a vector space is a subspace, we verify that S is closed under addition of vectors and multiplication of a vector by a scalar. Exactly the same idea applies to modules and algebras.

4.1.6 Definitions and Comments

If N is a nonempty subset of the R-module M, we say that N is a *submodule* of M (notation $N \le M$) if for every $x, y \in N$ and $r, s \in R$, we have $rx + sy \in N$. If M is an R-algebra, we say that N is a *subalgebra* if N is a submodule that is also a subring.

For example, if A is an abelian group (= $\mathbb{Z}$-module), the submodules of A are the subsets closed under addition and multiplication by an integer (which amounts to addition also). Thus the submodules of A are simply the subgroups. If R is a ring, hence a module over itself, the submodules are those subsets closed under addition and also under multiplication by any $r \in R$, in other words, the left ideals. (If we take R to be a right R-module, then the submodules are the right ideals.)

We can produce many examples of subspaces of vector spaces by considering kernels and images of linear transformations. A similar idea applies to modules.

4.1.7 Definitions and Comments

Let M and N be R-modules. A *module homomorphism* (also called an *R-homomorphism)* from M to N is a map $f\colon M \to N$ such that

$$f(rx + sy) = rf(x) + sf(y) \text{ for all } x, y \in M \text{ and } r, s \in R.$$

Equivalently, $f(x + y) = f(x) + f(y)$ and $f(rx) = rf(x)$ for all $x, y \in M$ and $r \in R$.

The *kernel* of a homomorphism f is $\ker f = \{x \in M\colon f(x) = 0\}$, and the *image* of f is $\{f(x)\colon x \in M\}$.

If follows from the definitions that the kernel of f is a submodule of M, and the image of f is a submodule of N.

If M and N are R-algebras, an *algebra homomorphism* or *homomorphism of algebras* from M to N is an R-module homomorphism that is also a ring homomorphism.

4.1.8 Another Way to Describe an Algebra

Assume that A is an algebra over the commutative ring R, and consider the map $r \to r1$ of R into A. The commutativity of R and the compatibility of the ring and module operations imply that the map is a ring homomorphism. To see this, note that if $r, s \in R$ then

$$(rs)1 = (sr)1 = s(r1) = s[(r1)1] = (r1)(s1).$$

Furthermore, if $y \in A$ then

$$(r1)y = r(1y) = r(y1) = y(r1)$$

so that $r1$ belongs to the *center* of A, i.e., the set of elements that commute with everything in A.

Conversely, if f is a ring homomorphism from the commutative ring R to the center of the ring A, we can make A into an R-module via $rx = f(r)x$. The compatibility conditions are satisfied because

$$r(xy) = f(r)(xy) = (f(r)x)y = (rx)y$$

and

$$(f(r)x)y = (xf(r))y = x(f(r)y) = x(ry).$$

Because of this result, the definition of an R-algebra is sometimes given as follows. The ring A is an algebra over the commutative ring R if there exists a ring homomorphism of R into the center of A. For us at this stage, such a definition would be a severe overdose of abstraction.

Notational Convention: We will often write the module $\{0\}$ (and the ideal $\{0\}$ in a ring) simply as 0.

Problems For Section 4.1

1. If I is an ideal of the ring R, show how to make the quotient ring R/I into a left R-module, and also show how to make R/I into a right R-module.
2. Let A be a commutative ring and F a field. Show that A is an algebra over F if and only if A contains (an isomorphic copy of) F as a subring.

Problems 3, 4 and 5 illustrate that familiar properties of vector spaces need not hold for modules.

3. Give an example of an R-module M with nonzero elements $r \in R$ and $x \in M$ such that $rx = 0$.
4. Let M be the additive group of rational numbers. Show that any two elements of M are linearly dependent (over the integers $\mathbb{Z}$).
5. Continuing Problem 4, show that M cannot have a basis, that is, a linearly independent spanning set over $\mathbb{Z}$.
6. Prove the *modular law* for subgroups of a given group G: With the group operation written multiplicatively,

$$A(B \cap C) = (AB) \cap C$$

if $A \subseteq C$. Switching to additive notation, we have, for submodules of a given R-module,

$$A + (B \cap C) = (A + B) \cap C,$$

again if $A \subseteq C$.

7. Let T be a linear transformation on the vector space V over the field F. Show how to make V into an R-module in a natural way, where R is the polynomial ring $F[X]$.

4.2 The Isomorphism Theorems For Modules

If N is a submodule of the R-module M (notation $N \le M$), then in particular N is an additive subgroup of M, and we may form the quotient group M/N in the usual way. In fact M/N becomes an R-module if we define $r(x + N) = rx + N$. (This makes sense because if x belongs to the submodule N, so does rx.) Since scalar multiplication in the quotient module M/N is carried out via scalar multiplication in the original module M, we can check the module axioms without difficulty. The canonical map $\pi\colon M \to M/N$ is a module homomorphism with kernel N. Just as with groups and rings, we can establish the basic isomorphism theorems for modules.

4.2.1 Factor Theorem For Modules

Any module homomorphism $f\colon M \to M'$ whose kernel contains N can be factored through M/N. In other words, there is a unique module homomorphism $\overline{f}\colon M/N \to M'$ such that $\overline{f}(x+N) = f(x)$. Furthermore, (i) $\overline{f}$ is an epimorphism if and only if f is an epimorphism; (ii) $\overline{f}$ is a monomorphism if and only if ker $f = N$; (iii) $\overline{f}$ is an isomorphism if and only if f is a epimorphism and ker$f = N$.

Proof. Exactly as in (2.3.1), with appropriate notational changes. (In Figure 2.3.1, replace R by M, S by M' and I by N.) ♣

4.2.2 First Isomorphism Theorem For Modules

If $f\colon M \to M'$ is a module homomorphism with kernel N, then the image of f is isomorphic to M/N.

Proof. Apply the factor theorem, and note that f is an epimorphism onto its image. ♣

4.2.3 Second Isomorphism Theorem For Modules

Let S and T be submodules of M, and let $S + T = \{x + y\colon x \in S, y \in T\}$. Then $S + T$ and $S \cap T$ are submodules of M and

$$(S + T)/T \cong S/(S \cap T).$$

Proof. The module axioms for $S + T$ and $S \cap T$ can be checked in routine fashion. Define a map $f\colon S \to M/T$ by $f(x) = x + T$. Then f is a module homomorphism whose kernel is $S \cap T$ and whose image is $\{x + T\colon x \in S\} = (S + T)/T$. The first isomorphism theorem for modules gives the desired result. ♣

4.2.4 Third Isomorphism Theorem For Modules

If $N \leq L \leq M$, then

$$M/L \cong (M/N)/(L/N).$$

Proof. Define $f\colon M/N \to M/L$ by $f(x+N) = x+L$. As in (2.3.4), the kernel of f is $\{x+N\colon x \in L\} = L/N$, and the image of f is $\{x+L\colon x \in M\} = M/L$. The result follows from the first isomorphism theorem for modules. ♣

4.2.5 Correspondence Theorem For Modules

Let N be a submodule of the R-module M. The map $S \to S/N$ sets up a one-to-one correspondence between the set of all submodules of M containing N and the set of all submodules of M/N. The inverse of the map is $T \to \pi^{-1}(T)$, where π is the canonical map: $M \to M/N$.

Proof. The correspondence theorem for groups yields a one-to-one correspondence between additive subgroups of M containing N and additive subgroups of M/N. We must check that submodules correspond to submodules, and it is sufficient to show that if $S_1/N \leq S_2/N$, then $S_1 \leq S_2$ (the converse is immediate). If $x \in S_1$, then $x+N \in S_1/N \subseteq S_2/N$, so $x+N = y+N$ for some $y \in S_2$. Thus $x-y \in N \subseteq S_2$, and since $y \in S_2$ we must have $x \in S_2$ as well. Therefore $S_1 \leq S_2$. ♣

We now look at modules that have a particularly simple structure, and can be used as building blocks for more complicated modules.

4.2.6 Definitions and Comments

An R-module M is *cyclic* if it is generated by a single element x. In other words,

$$M = Rx = \{rx\colon r \in R\}.$$

Thus every element of M is a scalar multiple of x. (If $x = 0$, then $M = \{0\}$, which is called the *zero module* and is often written simply as 0.) A cyclic vector space over a field is a one-dimensional space, assuming that $x \neq 0$.

The *annihilator* of an *element* y in the R-module M is $I_y = \{r \in R\colon ry = 0\}$, a left ideal of R. If R is commutative, and M is cyclic with generator x, then $M \cong R/I_x$. To see this, apply the first isomorphism theorem for modules to the map $r \to rx$ of R onto M. The *annihilator* of the *module* M is $I_o = \{r \in R\colon ry = 0 \text{ for every } y \in M\}$. Note that I_o is a two-sided ideal, because if $r \in I_0$ and $s \in R$, then for every $y \in M$ we have $(rs)y = r(sy) = 0$. When R is commutative, annihilating the generator of a cyclic module is equivalent to annihilating the entire module.

4.2.7 Lemma

(a) If x generates a cyclic module M over the commutative ring R, then $I_x = I_o$, so that $M \cong R/I_o$. (In this situation, I_o is frequently referred to as the *order ideal* of M.)
(b) Two cyclic R-modules over a commutative ring are isomorphic if and only if they have the same annihilator.

Proof. (a) If $rx = 0$ and $y \in M$, then $y = sx$ for some $s \in R$, so $ry = r(sx) = s(rx) = s0 = 0$. Conversely, if r annihilates M, then in particular, $rx = 0$.

(b) The "if" part follows from (a), so assume that $g\colon Rx \to Ry$ is an isomorphism of cyclic R-modules. Since g is an isomorphism, $g(x)$ must be a generator of Ry, so we may as well take $g(x) = y$. Then $g(rx) = rg(x) = ry$, so rx corresponds to ry under the isomorphism. Therefore r belongs to the annihilator of Rx if and only if r belongs to the annihilator of Ry. ♣

Problems For Section 4.2

1. Show that every submodule of the quotient module M/N can be expressed as $(L+N)/N$ for some submodule L of M.
2. In Problem 1, must L contain N?
3. In the matrix ring $M_n(R)$, let M be the submodule generated by E_{11}, the matrix with 1 in row 1, column 1, and 0's elsewhere. Thus $M = \{AE_{11}\colon A \in M_n(R)\}$. Show that M consists of all matrices whose entries are zero except perhaps in column 1.
4. Continuing Problem 3, show that the annihilator of E_{11} consists of all matrices whose first column is zero, but the annihilator of M is $\{0\}$.
5. If I is an ideal of the ring R, show that R/I is a cyclic R-module.
6. Let M be an R-module, and let I be an ideal of R. We wish to make M into an R/I-module via $(r + I)m = rm, r \in R, m \in M$. When will this be legal?
7. Assuming legality in Problem 6, let M_1 be the resulting R/I-module, and note that as sets, $M_1 = M$. Let N be a subset of M and consider the following two statements:

   (a) N is an R-submodule of M;

   (b) N is an R/I-submodule of M_1.

   Can one of these statements be true and the other false?

4.3 Direct Sums and Free Modules

4.3.1 Direct Products

In Section 1.5, we studied direct products of groups, and the basic idea seems to carry over to modules. Suppose that we have an R-module M_i for each i in some index set I (possibly infinite). The members of the *direct product* of the M_i, denoted by $\prod_{i\in I} M_i$, are all families $(a_i, i \in I)$, where $a_i \in M_i$. (A family is just a function on I whose value at the element i is a_i.) Addition is described by $(a_i) + (b_i) = (a_i + b_i)$ and scalar multiplication by $r(a_i) = (ra_i)$.

There is nothing wrong with this definition, but the resulting mathematical object has some properties that are undesirable. If (e_i) is the family with 1 in position i and zeros elsewhere, then (thinking about vector spaces) it would be useful to express an arbitrary element of the direct product in terms of the e_i. But if the index set I is infinite, we will need concepts of limit and convergence, and this will take us out of algebra and into analysis. Another approach is to modify the definition of direct product.

4.3.2 Definitions

The *external direct sum* of the modules $M_i, i \in I$, denoted by $\oplus_{i \in I} M_i$, consists of all families $(a_i, i \in I)$ with $a_i \in M_i$, such that $a_i = 0$ for all but finitely many i. Addition and scalar multiplication are defined exactly as for the direct product, so that the external direct sum coincides with the direct product when the index set I is finite.

The R-module M is the *internal direct sum* of the submodules M_i if each $x \in M$ can be expressed uniquely as $x_{i_1} + \cdots + x_{i_n}$ where $0 \neq x_{i_k} \in M_{i_k}, k = 1, \ldots, n$. (The positive integer n and the elements x_{i_k} depend on x. In any expression of this type, the indices i_k are assumed distinct.)

Just as with groups, the internal and external direct sums are isomorphic. To see this without a lot of formalism, let the element $x_{i_k} \in M_{i_k}$ correspond to the family that has x_{i_k} in position i_k and zeros elsewhere. We will follow standard practice and refer to the "direct sum" without the qualifying adjective. Again as with groups, the next result may help us to recognize when a module can be expressed as a direct sum.

4.3.3 Proposition

The module M is the direct sum of submodules M_i if and only if both of the following conditions are satisfied:

(1) $M = \sum_i M_i$, that is, each $x \in M$ is a finite sum of the form $x_{i_1} + \cdots + x_{i_n}$, where $x_{i_k} \in M_{i_k}$;

(2) For each i, $M_i \cap \sum_{j \neq i} M_j = 0$.

(Note that in condition (1), we do *not* assume that the representation is unique. Observe also that another way of expressing (2) is that if $x_{i_1} + \cdots + x_{i_n} = 0$, with $x_{i_k} \in M_{i_k}$, then $x_{i_k} = 0$ for all k.)

Proof. The necessity of the conditions follows from the definition of external direct sum, so assume that (1) and (2) hold. If $x \in M$ then by (1), x is a finite sum of elements from various M_i's. For convenience in notation, say $x = x_1 + x_2 + x_3 + x_4$ with $x_i \in M_i$, $i = 1, 2, 3, 4$. If the representation is not unique, say $x = y_1 + y_2 + y_4 + y_5 + y_6$ with $y_i \in M_i$, $i = 1, 2, 4, 5, 6$. Then x_3 is a sum of terms from modules other than M_3, so by (2), $x_3 = 0$. Similarly, $y_5 = y_6 = 0$ and we have $x_1 + x_2 + x_4 = y_1 + y_2 + y_4$. But then $x_1 - y_1$ is a sum of terms from modules other than M_1, so by (2), $x_1 = y_1$. Similarly $x_2 = y_2$, $x_4 = y_4$, and the result follows. ♣

A basic property of the direct sum $M = \oplus_{i \in I} M_i$ is that homomorphisms $f_i \colon M_i \to N$ can be "lifted" to M. In other words, there is a unique homomorphism $f \colon M \to N$ such that for each i, $f = f_i$ on M_i. Explicitly,

$$f(x_{i_1} + \cdots + x_{i_r}) = f_{i_1}(x_{i_1}) + \cdots + f_{i_r}(x_{i_r}).$$

[No other choice is possible for f, and since each f_i is a homomorphism, so is f.]

We know that every vector space has a basis, but not every module is so fortunate; see Section 4.1, Problem 5. We now examine modules that have this feature.

4.3.4 Definitions and Comments

Let S be a subset of the R-module M. We say that S is *linearly independent over* R if $\lambda_1 x_1 + \cdots + \lambda_k x_k = 0$ implies that all $\lambda_i = 0$ $(\lambda_i \in R, x_i \in S, k = 1, 2, \dots)$. We say that S is a *spanning* (or *generating*) *set for* M *over* R, or that S *spans (generates)* M *over* R if each $x \in M$ can be written as a finite linear combination of elements of S with coefficients in R. We will usually omit "over R" if the underlying ring R is clearly identified. A *basis* is a linearly independent spanning set, and a module that has a basis is said to be *free*.

Suppose that M is a free module with basis $(b_i, i \in I)$, and we look at the submodule M_i spanned by the basis element b_i. (In general, the submodule spanned (or generated) by a subset T of M consists of all finite linear combinations of elements of T with coefficients in R. Thus the submodule spanned by b_i is the set of all $rb_i, r \in R$.) If R is regarded as a module over itself, then the map $r \to rb_i$ is an R-module isomorphism of R and M_i, because $\{b_i\}$ is a linearly independent set. Since the b_i span M, it follows that M is the sum of the submodules M_i, and by linear independence of the b_i, the sum is direct. Thus we have an illuminating interpretation of a free module:

A free module is a direct sum of isomorphic copies of the underlying ring R.

Conversely, a direct sum of copies of R is a free R-module. If e_i has 1 as its i^{th} component and zeros elsewhere, the e_i form a basis.

This characterization allows us to recognize several examples of free modules.

1. For any positive integer n, R^n is a free R-module.

2. The matrix ring $M_{mn}(R)$ is a free R-module with basis E_{ij}, $i = 1, \dots, m$, $j = 1, \dots, n$.

3. The polynomial ring $R[X]$ is a free R-module with basis $1, X, X^2, \dots$.

We will adopt the standard convention that the zero module is free with the empty set as basis.

Any two bases for a vector space over a field have the same cardinality. This property does not hold for arbitrary free modules, but the following result covers quite a few cases.

4.3.5 Theorem

Any two bases for a free module M over a commutative ring R have the same cardinality.

Proof. If I is a maximal ideal of R, then $k = R/I$ is a field, and $V = M/IM$ is a vector space over k. [By IM we mean all finite sums $\sum a_i x_i$ with $a_i \in I$ and $x_i \in M$; thus IM is a submodule of M. If $r + I \in k$ and $x + IM \in M/IM$, we take $(r + I)(x + IM)$ to be $rx + IM$. The scalar multiplication is well-defined because $r \in I$ or $x \in IM$ implies that $rx \in IM$. We can express this in a slightly different manner by saying that I annihilates M/IM. The requirements for a vector space can be checked routinely.]

Now if (x_i) is a basis for M, let $\overline{x}_i = x_i + IM$. Since the x_i span M, the $\overline{x}_i$ span M/IM. If $\sum \overline{a}_i \overline{x}_i = 0$, where $\overline{a}_i = a_i + I, a_i \in R$, then $\sum a_i x_i \in IM$. Thus $\sum a_i x_i = \sum b_j x_j$ with $b_j \in I$. Since the x_i form a basis, we must have $a_i = b_j$ for some j. Consequently $a_i \in I$, so that $\overline{a}_i = 0$ in k. We conclude that the $\overline{x}_i$ form a basis for V over k, and since the dimension of V over k depends only on M, R and I, and not on a particular basis for M, the result follows. ♣

4.3.6 Some Key Properties of Free Modules

Suppose that M is a free module with basis (x_i), and we wish to construct a module homomorphism f from M to an arbitrary module N. Just as with vector spaces, we can specify $f(x_i) = y_i \in N$ arbitrarily on basis elements, and extend by linearity. Thus if $x = \sum a_i x_i \in M$, we have $f(x) = \sum a_i y_i$. (The idea should be familiar; for example, a linear transformation on Euclidean 3-space is determined by what it does to the three standard basis vectors.) Now let's turn this process around:

> If N is an arbitrary module, we can express N as a homomorphic image of a free module.

All we need is a set $(y_i, i \in I)$ of generators for N. (If all else fails, we can take the y_i to be all the elements of N.) We then construct a free module with basis $(x_i, i \in I)$. (To do this, take the direct sum of copies of R, as many copies as there are elements of I.) Then map x_i to y_i for each i.

Note that by the first isomorphism theorem, every module is a quotient of a free module.

Problems For Section 4.3

1. Show that in Proposition 4.3.3, (2) can be replaced by the weaker condition that for each $i, M_i \cap \sum_{j<i} M_j = 0$. (Assume a fixed total ordering on the index set.)
2. Let A be a finite abelian group. Is it possible for A to be a free $\mathbb{Z}$-module?
3. Let r and s be elements in the ideal I of the commutative ring R. Show that r and s are linearly dependent over R.
4. In Problem 3, regard I as an R-module. Can I be free?
5. Give an example of an infinite abelian group that is a free $\mathbb{Z}$-module, and an example of an infinite abelian group that is not free.
6. Show that a module M is free if and only if M has a subset S such that any function f from S to a module N can be extended uniquely to a module homomorphism from M to N.

7. Let M be a free module, expressed as the direct sum of α copies of the underlying ring R, where α and $|R|$ are infinite cardinals. Find the cardinality of M.
8. In Problem 7, assume that all bases B have the same cardinality, e.g., R is commutative. Find the cardinality of B.

4.4 Homomorphisms and Matrices

Suppose that M is a free R-module with a finite basis of n elements $v_1, \dots, v_n$, sometimes called a free module of *rank n*. We know from Section 4.3 that M is isomorphic to the direct sum of n copies of R. Thus we can regard M as R^n, the set of all n-tuples with components in R. Addition and scalar multiplication are performed componentwise, as in (4.1.3), Example 3. Note also that the direct sum coincides with the direct product, since we are summing only finitely many modules.

Let N be a free R-module of rank m, with basis $w_1, \dots, w_m$, and suppose that f is a module homomorphism from M to N. Just as in the familiar case of a linear transformation on a finite-dimensional vector space, we are going to represent f by a matrix. For each j, $f(v_j)$ is a linear combination of the basis elements w_j, so that

$$f(v_j) = \sum_{i=1}^{m} a_{ij} w_i, \ j = 1, \dots, n \tag{1}$$

where the a_{ij} belong to R.

It is natural to associate the $m \times n$ matrix A with the homomorphism f, and it appears that we have an isomorphism of some sort, but an isomorphism of what? If f and g are homomorphisms of M into N, then f and g can be added (and subtracted): $(f+g)(x) = f(x) + g(x)$. If f is represented by the matrix A and g by B, then $f+g$ corresponds to $A+B$. This gives us an abelian group isomorphism of $\operatorname{Hom}_R(M, N)$, the set of all R-module homomorphisms from M to N, and $M_{mn}(R)$, the set of all $m \times n$ matrices with entries in R. In addition, $M_{mn}(R)$ is an R-module, so it is tempting to say "obviously, we have an R-module isomorphism". But we must be very careful here. If $f \in \operatorname{Hom}_R(M, N)$ and $s \in R$, we can define sf in the natural way: $(sf)(x) = sf(x)$. However, if we carry out the "routine" check that $sf \in \operatorname{Hom}_R(M, N)$, there is one step that causes alarm bells to go off:

$$(sf)(rx) = sf(rx) = srf(x), \text{ but } r(sf)(x) = rsf(x)$$

and the two expressions can disagree if R is not commutative. Thus $\operatorname{Hom}_R(M, N)$ need not be an R-module. Let us summarize what we have so far.

4.4.1 The Correspondence Between Homomorphisms and Matrices

Associate with each $f \in \operatorname{Hom}_R(M, N)$ a matrix A as in (1) above. This yields an abelian group isomorphism, and also an R-module isomorphism if R is commutative.

Now let $m = n$, so that the dimensions are equal and the matrices are square, and take $v_i = w_i$ for all i. A homomorphism from M to itself is called an *endomorphism* of

M, and we use the notation $\operatorname{End}_R(M)$ for $\operatorname{Hom}_R(M,M)$. Since $\operatorname{End}_R(M)$ is a ring under composition of functions, and $M_n(R)$ is a ring under matrix multiplication, it is plausible to conjecture that we have a ring isomorphism. If f corresponds to A and g to B, then we can apply g to both sides of (1) to obtain

$$g(f(v_j)) = \sum_{i=1}^{n} a_{ij} \sum_{k=1}^{n} b_{ki} v_k = \sum_{k=1}^{n} (\sum_{i=1}^{n} a_{ij} b_{ki}) v_k. \tag{2}$$

If R is commutative, then $a_{ij}b_{ki} = b_{ki}a_{ij}$, and the matrix corresponding to $gf = g \circ f$ is BA, as we had hoped. In the noncommutative case, we will not be left empty-handed if we define the *opposite ring* R^o, which has exactly the same elements as R and the same addition structure. However, multiplication is done backwards, i.e., ab in R^o is ba in R. It is convenient to attach a superscript ^{o} to the elements of R^o, so that

$$a^o b^o = ba \text{ (more precisely, } a^o b^o = (ba)^o).$$

Thus in (2) we have $a_{ij}b_{ki} = b^o_{ki}a^o_{ij}$. To summarize,

> The endomorphism ring $\operatorname{End}_R(M)$ is isomorphic to the ring of $n \times n$ matrices with coefficients in the opposite ring R^o. If R is commutative, then $\operatorname{End}_R(M)$ is ring-isomorphic to $M_n(R)$.

4.4.2 Preparation For The Smith Normal Form

We now set up some basic machinery to be used in connection with the Smith normal form and its applications. Assume that M is a free $\mathbb{Z}$-module of rank n, with basis $x_1, \dots, x_n$, and that K is a submodule of M with finitely many generators $u_1, \dots, u_m$. (We say that K is *finitely generated*.) We change to a new basis $y_1, \dots, y_n$ via $Y = PX$, where X [resp. Y] is a column vector with components x_i [resp. y_i]. Since X and Y are bases, the $n \times n$ matrix P must be invertible, and we need to be very clear on what this means. If the determinant of P is nonzero, we can construct P^{-1}, for example by the "adjoint divided by determinant" formula given in Cramer's rule. But the underlying ring is $\mathbb{Z}$, not $\mathbb{Q}$, so we require that the coefficients of P^{-1} be integers. (For a more transparent equivalent condition, see Problem 1.) Similarly, we are going to change generators of K via $V = QU$, where Q is an invertible $m \times m$ matrix and U is a column vector with components u_i.

The generators of K are linear combinations of basis elements, so we have an equation of the form $U = AX$, where A is an $m \times n$ matrix called the *relations matrix*. Thus

$$V = QU = QAX = QAP^{-1}Y.$$

so the new relations matrix is

$$B = QAP^{-1}.$$

Thus B is obtained from A by pre-and postmultiplying by invertible matrices, and we say that A and B are *equivalent*. We will see that two matrices are equivalent iff they have the same Smith normal form. The point we wish to emphasize now is that if we know the matrix P, we can compute the new basis Y, and if we know the matrix Q, we can compute the new system of generators V. In our applications, P and Q will be constructed by elementary row and column operations.

Problems For Section 4.4

1. Show that a square matrix P over the integers has an inverse with integer entries if and only if P is *unimodular*, that is, the determinant of P is ± 1.
2. Let V be the direct sum of the R-modules $V_1, \dots, V_n$, and let W be the direct sum of R-modules $W_1, \dots, W_m$. Indicate how a module homomorphism from V to W can be represented by a matrix. (The entries of the matrix need not be elements of R.)
3. Continuing Problem 2, show that if V^n is the direct sum of n copies of the R-module V, then we have a ring isomorphism
$$\operatorname{End}_R(V^n) \cong M_n(\operatorname{End}_R(V)).$$
4. Show that if R is regarded as an R-module, then $\operatorname{End}_R(R)$ is isomorphic to the opposite ring R^o.
5. Let R be a ring, and let $f \in \operatorname{End}_R(R)$. Show that for some $r \in R$ we have $f(x) = xr$ for all $x \in R$.
6. Let M be a free R-module of rank n. Show that $\operatorname{End}_R(M) \cong M_n(R^o)$, a ring isomorphism.
7. Continuing Problem 6, if R is commutative, show that the ring isomorphism is in fact an R-algebra isomorphism.

4.5 Smith Normal Form

We are going to describe a procedure that is very similar to reduction of a matrix to echelon form. The result is that every matrix over a principal ideal domain is equivalent to a matrix in Smith normal form. Explicitly, the Smith matrix has nonzero entries only on the main diagonal. The main diagonal entries are, from the top, $a_1, \dots, a_r$ (possibly followed by zeros), where the a_i are nonzero and a_i divides a_{i+1} for all i.

We will try to convey the basic ideas via a numerical example. This will allow us to give informal but convincing proofs of some major theorems. A formal development is given in Jacobson, Basic Algebra I, Chapter 3. All our computations will be in the ring of integers, but we will indicate how the results can be extended to an arbitrary principal ideal domain. Let's start with the following matrix:

$$\begin{bmatrix} 0 & 0 & 22 & 0 \\ -2 & 2 & -6 & -4 \\ 2 & 2 & 6 & 8 \end{bmatrix}$$

As in (4.4.2), we assume a free $\mathbb{Z}$-module M with basis x_1, x_2, x_3, x_4, and a submodule K generated by u_1, u_2, u_3, where $u_1 = 22x_3, u_2 = -2x_1 + 2x_2 - 6x_3 - 4x_4, u_3 = 2x_1 + 2x_2 + 6x_3 + 8x_4$. The first step is to bring the smallest positive integer to the 1-1 position. Thus interchange rows 1 and 3 to obtain

$$\begin{bmatrix} 2 & 2 & 6 & 8 \\ -2 & 2 & -6 & -4 \\ 0 & 0 & 22 & 0 \end{bmatrix}$$

Since all entries in column 1, and similarly in row 1, are divisible by 2, we can pivot about the 1-1 position, in other words, use the 1-1 entry to produce zeros. Thus add row 1 to row 2 to get

$$\begin{bmatrix}2 & 2 & 6 & 8\\0 & 4 & 0 & 4\\0 & 0 & 22 & 0\end{bmatrix}$$

Add -1 times column 1 to column 2, then add -3 times column 1 to column 3, and add -4 times column 1 to column 4. The result is

$$\begin{bmatrix}2 & 0 & 0 & 0\\0 & 4 & 0 & 4\\0 & 0 & 22 & 0\end{bmatrix}$$

Now we have "peeled off" the first row and column, and we bring the smallest positive integer to the 2-2 position. It's already there, so no action is required. Furthermore, the 2-2 element is a multiple of the 1-1 element, so again no action is required. Pivoting about the 2-2 position, we add -1 times column 2 to column 4, and we have

$$\begin{bmatrix}2 & 0 & 0 & 0\\0 & 4 & 0 & 0\\0 & 0 & 22 & 0\end{bmatrix}$$

Now we have peeled off the first two rows and columns, and we bring the smallest positive integer to the 3-3 position; again it's already there. But 22 is not a multiple of 4, so we have more work to do. Add row 3 to row 2 to get

$$\begin{bmatrix}2 & 0 & 0 & 0\\0 & 4 & 22 & 0\\0 & 0 & 22 & 0\end{bmatrix}$$

Again we pivot about the 2-2 position; 4 does not divide 22, but if we add -5 times column 2 to column 3, we have

$$\begin{bmatrix}2 & 0 & 0 & 0\\0 & 4 & 2 & 0\\0 & 0 & 22 & 0\end{bmatrix}$$

Interchange columns 2 and 3 to get

$$\begin{bmatrix}2 & 0 & 0 & 0\\0 & 2 & 4 & 0\\0 & 22 & 0 & 0\end{bmatrix}$$

Add -11 times row 2 to row 3 to obtain

$$\begin{bmatrix}2 & 0 & 0 & 0\\0 & 2 & 4 & 0\\0 & 0 & -44 & 0\end{bmatrix}$$

Finally, add -2 times column 2 to column 3, and then (as a convenience to get rid of the minus sign) multiply row (or column) 3 by -1; the result is

$$\begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 44 & 0 \end{bmatrix}$$

which is the Smith normal form of the original matrix. Although we had to backtrack to produce a new pivot element in the 2-2 position, the new element is smaller than the old one (since it is a remainder after division by the original number). Thus we cannot go into an infinite loop, and the algorithm will indeed terminate in a finite number of steps. In view of (4.4.2), we have the following interpretation.

We have a new basis y_1, y_2, y_3, y_4 for M, and new generators v_1, v_2, v_3 for K, where $v_1 = 2y_1, v_2 = 2y_2$, and $v_3 = 44y_3$. In fact since the v_j's are nonzero multiples of the corresponding y_j's, they are linearly independent, and consequently form a basis of K. The new basis and set of generators can be expressed in terms of the original sets; see Problems 1–3 for the technique.

The above discussion indicates that the Euclidean algorithm guarantees that the Smith normal form can be computed in finitely many steps. Therefore the Smith procedure can be carried out in any Euclidean domain. In fact we can generalize to a principal ideal domain. Suppose that at a particular stage of the computation, the element a occupies the 1-1 position of the Smith matrix S, and the element b is in row 1, column 2. To use a as a pivot to eliminate b, let d be the greatest common divisor of a and b, and let r and s be elements of R such that $ar + bs = d$ (see (2.7.2)). We postmultiply the Smith matrix by a matrix T of the following form (to aid in the visualization, we give a concrete 5×5 example):

$$\begin{bmatrix} r & b/d & 0 & 0 & 0 \\ s & -a/d & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

The 2×2 matrix in the upper left hand corner has determinant -1, and is therefore invertible over R. The element in the 1-1 position of ST is $ar + bs = d$, and the element in the 1-2 position is $ab/d - ba/d = 0$, as desired. We have replaced the pivot element a by a divisor d, and this will decrease the number of prime factors, guaranteeing the finite termination of the algorithm. Similarly, if b were in the 2-1 position, we would premultiply S by the transpose of T; thus in the upper left hand corner we would have

$$\begin{bmatrix} r & s \\ b/d & -a/d \end{bmatrix}$$

Problems For Section 4.5

1. Let A be the matrix

$$\begin{bmatrix} -2 & 3 & 0 \\ -3 & 3 & 0 \\ -12 & 12 & 6 \end{bmatrix}$$

over the integers. Find the Smith normal form of A. (It is convenient to begin by adding column 2 to column 1.)

2. Continuing Problem 1, find the matrices P and Q, and verify that QAP^{-1} is the Smith normal form.
3. Continuing Problem 2, if the original basis for M is $\{x_1, x_2, x_3\}$ and the original set of generators of K is $\{u_1, u_2, u_3\}$, find the new basis and set of generators.

It is intuitively reasonable, but a bit messy to prove, that if a matrix A over a PID is multiplied by an invertible matrix, then the greatest common divisor of all the $i \times i$ minors of A is unchanged. Accept this fact in doing Problems 4 and 5.

4. The nonzero components a_i of the Smith normal form S of A are called the *invariant factors* of A. Show that the invariant factors of A are unique (up to associates).
5. Show that two $m \times n$ matrices are equivalent if and only if they have the same invariant factors, i.e. (by Problem 4), if and only if they have the same Smith normal form.
6. Recall that when a matrix over a field is reduced to row-echelon form (only row operations are involved), a pivot column is followed by non-pivot columns whose entries are zero in all rows below the pivot element. When a similar computation is carried out over the integers, or more generally over a Euclidean domain, the resulting matrix is said to be in *Hermite normal form.* We indicate the procedure in a typical example. Let

$$A = \begin{bmatrix} 6 & 4 & 13 & 5 \\ 9 & 6 & 0 & 7 \\ 12 & 8 & -1 & 12 \end{bmatrix}.$$

Carry out the following sequence of steps:

1. Add -1 times row 1 to row 2
2. Interchange rows 1 and 2
3. Add -2 times row 1 to row 2, and then add -4 times row 1 to row 3
4. Add -1 times row 2 to row 3
5. Interchange rows 2 and 3
6. Add -3 times row 2 to row 3
7. Interchange rows 2 and 3
8. Add -4 times row 2 to row 3
9. Add 5 times row 2 to row 1 (this corresponds to choosing $0, 1, \dots, m-1$ as a complete system of residues mod m)
10. Add 2 times row 3 to row 1, and then add row 3 to row 2

We now have reduced A to Hermite normal form.

7. Continuing Problem 6, consider the simultaneous equations

$$6x + 4y + 13z \equiv 5, \; 9x + 6y \equiv 7, \; 12x + 8y - z \equiv 12 \pmod{m}$$

For which values of $m \geq 2$ will the equations be consistent?

4.6 Fundamental Structure Theorems

The Smith normal form yields a wealth of information about modules over a principal ideal domain. In particular, we will be able to see exactly what finitely generated abelian groups must look like.

Before we proceed, we must mention a result that we will use now but not prove until later (see (7.5.5), Example 1, and (7.5.9)). If M is a finitely generated module over a PID R, then every submodule of M is finitely generated. [R is a Noetherian ring, hence M is a Noetherian R-module.] To avoid gaps in the current presentation, we can restrict our attention to finitely generated submodules.

4.6.1 Simultaneous Basis Theorem

Let M be a free module of finite rank $n \geq 1$ over the PID R, and let K be a submodule of M. Then there is a basis $\{y_1, \dots, y_n\}$ for M and nonzero elements $a_1, \dots, a_r \in R$ such that $r \leq n$, a_i divides a_{i+1} for all i, and $\{a_1 y_1, \dots, a_r y_r\}$ is a basis for K.

Proof. This is a corollary of the construction of the Smith normal form, as explained in Section 4.5. ♣

4.6.2 Corollary

Let M be a free module of finite rank n over the PID R. Then every submodule of M is free of rank at most n.

Proof. By (4.6.1), the submodule K has a basis with $r \leq n$ elements. ♣

In (4.6.2), the hypothesis that M has finite rank can be dropped, as the following sketch suggests. We can well-order the generators u_α of K, and assume as a transfinite induction hypothesis that for all $\beta < \alpha$, the submodule K_β spanned by all the generators up to u_β is free of rank at most that of M, and that if $\gamma < \beta$, then the basis of K_γ is contained in the basis of K_β. The union of the bases S_β of the K_β is a basis S_α for K_α. Furthermore, the inductive step preserves the bound on rank. This is because $|S_\beta| \leq$ rank M for all $\beta < \alpha$, and $|S_\alpha|$ is the smallest cardinal bounded below by all $|S_\beta|, \beta < \alpha$. Thus $|S_\alpha| \leq$ rank M.

4.6.3 Fundamental Decomposition Theorem

Let M be a finitely generated module over the PID R. Then there are ideals $I_1 = \langle a_1 \rangle$, $I_2 = \langle a_2 \rangle, \dots, I_n = \langle a_n \rangle$ of R such that $I_1 \supseteq I_2 \supseteq \cdots \supseteq I_n$ (equivalently, $a_1 \mid a_2 \mid \cdots \mid a_n$) and

$$M \cong R/I_1 \oplus R/I_2 \oplus \cdots \oplus R/I_n.$$

Thus M is a direct sum of cyclic modules.

Proof. By (4.3.6), M is the image of a free module R^n under a homomorphism f. If K is the kernel of f, then by (4.6.1) we have a basis $y_1, \dots, y_n$ for R^n and a corresponding basis $a_1y_1, \dots, a_ry_r$ for K. We set $a_i = 0$ for $r < i \leq n$. Then

$$M \cong R^n/K \cong \frac{Ry_1 \oplus \cdots \oplus Ry_n}{Ra_1y_1 \oplus \cdots \oplus Ra_ny_n} \cong \bigoplus_{i=1}^{n} Ry_i/Ra_iy_i.$$

(To justify the last step, apply the first isomorphism theorem to the map

$$r_1y_1 + \cdots + r_ny_n \to (r_1y_1 + Ra_1y_1, \dots, r_ny_n + Ra_ny_n.)$$

But

$$Ry_i/Ra_iy_i \cong R/Ra_i,$$

as can be seen via an application of the first isomorphism theorem to the map $r \to ry_i + Ra_iy_i$. Thus if $I_i = Ra_i$, $i = 1, \dots, n$, we have

$$M \cong \bigoplus_{i=1}^{n} R/I_i$$

and the result follows. ♣

Remark It is plausible, and can be proved formally, that the uniqueness of invariant factors in the Smith normal form implies the uniqueness of the decomposition (4.6.3). Intuitively, the decomposition is completely specified by the sequence $a_1, \dots, a_n$, as the proof of (4.6.3) indicates.

4.6.4 Finite Abelian Groups

Suppose that G is a finite abelian group of order 1350; what can we say about G? In the decomposition theorem (4.6.3), the components of G are of the form $\mathbb{Z}/\mathbb{Z}a_i$, that is, cyclic groups of order a_i. We must have $a_i \mid a_{i+1}$ for all i, and since the order of a direct sum is the product of the orders of the components, we have $a_1 \cdots a_r = 1350$.

The first step in the analysis is to find the prime factorization of 1350, which is $(2)(3^3)(5^2)$. One possible choice of the a_i is $a_1 = 3$, $a_2 = 3$, $a_3 = 150$. It is convenient to display the prime factors of the a_i, which are called *elementary divisors*, as follows:

$$\begin{aligned} a_1 &= 3 = 2^0 3^1 5^0 \\ a_2 &= 3 = 2^0 3^1 5^0 \\ a_3 &= 150 = 2^1 3^1 5^2 \end{aligned}$$

Since $a_1a_2a_3 = 2^1 3^3 5^2$, the sum of the exponents of 2 must be 1, the sum of the exponents of 3 must be 3, and the sum of the exponents of 5 must be 2. A particular distribution of exponents of a prime p corresponds to a partition of the sum of the exponents. For example, if the exponents of p were 0, 1, 1 and 2, this would correspond to a partition of 4 as $1 + 1 + 2$. In the above example, the partitions are $1 = 1$, $3 = 1 + 1 + 1$, $2 = 2$. We can count the number of abelian groups of order 1350 (up to isomorphism) by counting partitions. There is only one partition of 1, there are two partitions of 2 (2 and $1+1$) and three partitions of 3 (3, $1+2$ and $1+1+1$). [This pattern does not continue; there are five partitions of 4, namely 4, $1+3$, $1+1+2$, $1+1+1+1$, $2+2$, and seven partitions of 5, namely 5, $1+4$, $1+1+3$, $1+1+1+2$, $1+1+1+1+1$, $1+2+2$, $2+3$.] We specify a group by choosing a partition of 1, a partition of 3 and a partition of 2, and the number of possible choices is $(1)(3)(2) = 6$. Each choice of a sequence of partitions produces a different sequence of invariant factors. Here is the entire list; the above example appears as entry (5).

(1) $a_1 = 2^1 3^3 5^2 = 1350$, $G \cong \mathbb{Z}_{1350}$

(2) $a_1 = 2^0 3^0 5^1 = 5$, $a_2 = 2^1 3^3 5^1 = 270$, $G \cong \mathbb{Z}_5 \oplus \mathbb{Z}_{270}$

(3) $a_1 = 2^0 3^1 5^0 = 3$, $a_2 = 2^1 3^2 5^2 = 450$, $G \cong \mathbb{Z}_3 \oplus \mathbb{Z}_{450}$

(4) $a_1 = 2^0 3^1 5^1 = 15$, $a_2 = 2^1 3^2 5^1 = 90$, $G \cong \mathbb{Z}_{15} \oplus \mathbb{Z}_{90}$

(5) $a_1 = 2^0 3^1 5^0 = 3$, $a_2 = 2^0 3^1 5^0 = 3$, $a_3 = 2^1 3^1 5^2 = 150$, $G \cong \mathbb{Z}_3 \oplus \mathbb{Z}_3 \oplus \mathbb{Z}_{150}$

(6) $a_1 = 2^0 3^1 5^0 = 3$, $a_2 = 2^0 3^1 5^1 = 15$, $a_3 = 2^1 3^1 5^1 = 30$, $G \cong \mathbb{Z}_3 \oplus \mathbb{Z}_{15} \oplus \mathbb{Z}_{30}$.

In entry (6) for example, the maximum number of summands in a partition is 3 ($= 1 + 1 + 1$), and this reveals that there will be three invariant factors. The partition $2 = 1 + 1$ has only two summands, and it is "pushed to the right" so that 5^1 appears in a_2 and a_3 but not a_1. (Remember that we must have $a_1 \mid a_2 \mid a_3$.). Also, we can continue to decompose some of the components in the direct sum representation of G. (If m and n are relatively prime, then $\mathbb{Z}_{mn} \cong \mathbb{Z}_m \oplus \mathbb{Z}_n$ by the Chinese remainder theorem.) However, this does not change the conclusion that there are only 6 mutually nonisomorphic abelian groups of order 1350.

Before examining infinite abelian groups, let's come back to the fundamental decomposition theorem.

4.6.5 Definitions and Comments

If x belongs to the R-module M, where R is any integral domain, then x is a *torsion element* if $rx = 0$ for some nonzero $r \in R$. The *torsion submodule* T of M is the set of torsion elements. (T is indeed a submodule; if $rx = 0$ and $sy = 0$, then $rs(x + y) = 0$.) M is a *torsion module* if T is all of M, and M is *torsion-free* if T consists of 0 alone, in other words, $rx = 0$ implies that either $r = 0$ or $x = 0$. A free module must be torsion-free, by definition of linear independence. Now assume that R is a PID, and decompose M as in (4.6.3), where $a_1, \dots, a_r$ are nonzero and $a_{r+1} = \cdots = a_n = 0$. Each module $R/\langle a_i \rangle$, $1 \le i \le r$, is torsion (it is annihilated by a_i), and the $R/\langle a_i \rangle$, $r + 1 \le i \le n$, are copies of R. Thus $\oplus_{i=r+1}^n R/\langle a_i \rangle$ is free. We conclude that

(*) every finitely generated module over a PID is the direct sum of its torsion submodule and a free module

and

(**) every finitely generated torsion-free module over a PID is free.

In particular, a finitely generated abelian group is the direct sum of a number (possibly zero) of finite cyclic groups and a free abelian group (possibly $\{0\}$).

4.6.6 Abelian Groups Specified by Generators and Relations

Suppose that we have a free abelian group F with basis x_1, x_2, x_3, and we impose the following constraints on the x_i:

$$2x_1 + 2x_2 + 8x_3 = 0, \quad -2x_1 + 2x_2 + 4x_3 = 0. \tag{1}$$

What we are doing is forming a "submodule of relations" K with generators

$$u_1 = 2x_1 + 2x_2 + 8x_3 \quad \text{and} \quad u_2 = -2x_1 + 2x_2 + 4x_3 \tag{2}$$

and we are identifying every element in K with zero. This process yields the abelian group $G = F/K$, which is generated by $x_1 + K$, $x_2 + K$ and $x_3 + K$. The matrix associated with (2) is

$$\begin{bmatrix} 2 & 2 & 8 \\ -2 & 2 & 4 \end{bmatrix}$$

and a brief computation gives the Smith normal form

$$\begin{bmatrix} 2 & 0 & 0 \\ 0 & 4 & 0 \end{bmatrix}.$$

Thus we have a new basis y_1, y_2, y_3 for F and new generators $2y_1, 4y_2$ for K. The quotient group F/K is generated by $y_1 + K, y_2 + K$ and $y_3 + K$, with $2(y_1 + K) = 4(y_2 + K) = 0 + K$. In view of (4.6.3) and (4.6.5), we must have

$$F/K \cong \mathbb{Z}_2 \oplus \mathbb{Z}_4 \oplus \mathbb{Z}.$$

Canonical forms of a square matrix A can be developed by reducing the matrix $xI - A$ to Smith normal form. In this case, R is the polynomial ring $F[X]$ where F is a field. But the analysis is quite lengthy, and I prefer an approach in which the Jordan canonical form is introduced at the very beginning, and then used to prove some basic results in the theory of linear operators; see Ash, A Primer of Abstract Mathematics, MAA 1998.

Problems For Section 4.6

1. Classify all abelian groups of order 441.
2. Classify all abelian groups of order 40.
3. Identify the abelian group given by generators x_1, x_2, x_3 and relations

$$x_1 + 5x_2 + 3x_3 = 0, \; 2x_1 - x_2 + 7x_3 = 0, \; 3x_1 + 4x_2 + 2x_3 = 0.$$

4. In (4.6.6), suppose we cancel a factor of 2 in Equation (1). This changes the matrix associated with (2) to

$$\begin{bmatrix} 1 & 1 & 4 \\ -1 & 1 & 2 \end{bmatrix},$$

whose Smith normal form differs from that given in the text. What's wrong?

5. Let M, N and P be abelian groups. If $M \oplus N \cong M \oplus P$, show by example that N need not be isomorphic to P.

6. In Problem 5, show that $M \oplus N \cong M \oplus P$ does imply $N \cong P$ if M, N and P are finitely generated.

4.7 Exact Sequences and Diagram Chasing

4.7.1 Definitions and Comments

Suppose that the R-module M is the direct sum of the submodules A and B. Let f be the *inclusion* or *injection* map of A into M (simply the identity function on A), and let g be the *natural projection* of M on B, given by $g(a+b) = b$, $a \in A$, $b \in B$. The image of f, namely A, coincides with the kernel of g, and we say that the sequence

$$\begin{tikzcd} A \arrow[r, "f"] & M \arrow[r, "g"] & B \end{tikzcd} \tag{1}$$

is *exact* at M. A longer (possibly infinite) sequence of homomorphisms is said to be exact if it is exact at each junction, that is, everywhere except at the left and right endpoints, if they exist.

There is a natural exact sequence associated with any module homomorphism $g\colon M \to N$, namely

$$\begin{tikzcd} 0 \arrow[r] & A \arrow[r, "f"] & M \arrow[r, "g"] & B \arrow[r] & 0 \end{tikzcd} \tag{2}$$

In the diagram, A is the kernel of g, f is the injection map, and B is the image of g. A five term exact sequence with zero modules at the ends, as in (2), is called a *short exact sequence*. Notice that exactness at A is equivalent to ker $f = 0$, i.e., injectivity of f. Exactness at B is equivalent to im $g = B$, i.e., surjectivity of g. Notice also that by the first isomorphism theorem, we may replace B by M/A and g by the canonical map of M onto M/A, while preserving exactness.

Now let's come back to (1), where M is the direct sum of A and B, and attach zero modules to produce the short exact sequence (2). If we define h as the injection of B into M and e as the projection of M on A, we have (see (3) below) $g \circ h = 1$ and $e \circ f = 1$, where 1 stands for the identity map.

$$\begin{tikzcd} 0 \arrow[r] & A \arrow[r, shift right, "f"'] & M \arrow[l, shift right, "e"'] \arrow[r, shift right, "g"'] & B \arrow[l, shift right, "h"'] \arrow[r] & 0 \end{tikzcd} \tag{3}$$

The short exact sequence (2) is said to *split on the right* if there is a homomorphism $h\colon B \to M$ such that $g \circ h = 1$, and *split on the left* if there is a homomorphism $e\colon M \to A$ such that $e \circ f = 1$. These conditions turn out to be equivalent, and both are equivalent to the statement that M is essentially the direct sum of A and B. "Essentially" means that not only is M isomorphic to $A \oplus B$, but f can be identified with the injection of A into the direct sum, and g with the projection of the direct sum on B. We will see how to make this statement precise, but first we must turn to *diagram chasing*, which is a technique for proving assertions about commutative diagrams by sliding from one vertex to another. The best way to get accustomed to the method is to do examples. We will work one out in great detail in the text, and there will be more practice in the exercises, with solutions provided.

We will use the shorthand gf for $g \circ f$ and fm for $f(m)$.

4.7.2 The Five Lemma

Consider the following commutative diagram with exact rows.

$$\begin{tikzcd}
D \arrow[r, "e"] \arrow[d, "s"'] & A \arrow[r, "f"] \arrow[d, "t"'] & M \arrow[r, "g"] \arrow[d, "u"'] & B \arrow[r, "h"] \arrow[d, "v"'] & C \arrow[d, "w"'] \\
D' \arrow[r, "e'"'] & A' \arrow[r, "f'"'] & M' \arrow[r, "g'"'] & B' \arrow[r, "h'"'] & C'
\end{tikzcd}$$

If s, t, v and w are isomorphisms, so is u. (In fact, the hypotheses on s and w can be weakened to s surjective and w injective.)

Proof. The two parts of the proof are of interest in themselves, and are frequently called the "four lemma", since they apply to diagrams with four rather than five modules in each row.

(i) If t and v are surjective and w is injective, then u is surjective.

(ii) If s is surjective and t and v are injective, then u is injective.

[The pattern suggests a "duality" between injective and surjective maps. This idea will be explored in Chapter 10; see (10.1.4).] The five lemma follows from (i) and (ii). To prove (i), let $m' \in M'$. Then $g'm' \in B'$, and since v is surjective, we can write $g'm' = vb$ for some $b \in B$. By commutativity of the square on the right, $h'vb = whb$. But $h'vb = h'g'm' = 0$ by exactness of the bottom row at B', and we then have $whb = 0$. Thus $hb \in \ker w$, and since w is injective, we have $hb = 0$, so that $b \in \ker h = \operatorname{im} g$ by exactness of the top row at B. So we can write $b = gm$ for some $m \in M$. Now $g'm' = vb$ (see above) $= vgm = g'um$ by commutativity of the square $MBB'M'$. Therefore $m' - um \in \ker g' = imf'$ by exactness of the bottom row at M'. Let $m' - um = f'a'$ for some $a' \in A'$. Since t is surjective, $a' = ta$ for some $a \in A$, and by commutativity of the square $AMM'A'$, $f'ta = ufa$, so $m' - um = ufa$, so $m' = u(m + fa)$. Consequently, m' belongs to the image of u, proving that u is surjective.

To prove (ii), suppose $m \in \ker u$. By commutativity, $g'um = vgm$, so $vgm = 0$. Since v is injective, $gm = 0$. Thus $m \in \ker g = \operatorname{im} f$ by exactness, say $m = fa$. Then

$0 = um = ufa = f'ta$ by commutativity. Thus $ta \in \ker f' = \operatorname{im} e'$ by exactness. If $ta = e'd'$, then since s is surjective, we can write $d' = sd$, so $ta = e'sd$. By commutativity, $e'sd = ted$, so $ta = ted$. By injectivity of t, $a = ed$. Therefore $m = fa = fed = 0$ by exactness. We conclude that u is injective. ♣

4.7.3 Corollary: The Short Five Lemma

Consider the following commutative diagram with exact rows. (Throughout this section, all maps in commutative diagrams and exact sequences are assumed to be R-module homomorphisms.)

$$\begin{tikzcd}
0 \arrow[r] & A \arrow[r, "f"] \arrow[d, "t"'] & M \arrow[r, "g"] \arrow[d, "u"'] & B \arrow[r] \arrow[d, "v"'] & 0 \\
0 \arrow[r] & A' \arrow[r, "f'"'] & M' \arrow[r, "g'"'] & B' \arrow[r] & 0
\end{tikzcd}$$

If t and v are isomorphisms, so is u.

Proof. Apply the five lemma with $C = D = C' = D' = 0$, and s and w the identity maps. ♣

We can now deal with splitting of short exact sequences.

4.7.4 Proposition

Let

$$\begin{tikzcd}
0 \arrow[r] & A \arrow[r, "f"] & M \arrow[r, "g"] & B \arrow[r] & 0
\end{tikzcd}$$

be a short exact sequence. The following conditions are equivalent, and define a *split exact sequence.*

(i) The sequence splits on the right.

(ii) The sequence splits on the left.

(iii) There is an isomorphism u of M and $A \oplus B$ such that the following diagram is commutative.

$$\begin{tikzcd}
0 \arrow[r] & A \arrow[r, "f"] \arrow[d, equal] & M \arrow[r, "g"] \arrow[d, "u"'] & B \arrow[r] \arrow[d, equal] & 0 \\
0 \arrow[r] & A \arrow[r, "i"'] & A \oplus B \arrow[r, "\pi"'] & B \arrow[r] & 0
\end{tikzcd}$$

Thus M is isomorphic to the direct sum of A and B, and in addition, f can be identified with the injection i of A into $A \oplus B$, and g with the projection π of the direct sum onto B. (The double vertical bars indicate the identity map.)

Proof. It follows from our earlier discussion of diagram (3) that (iii) implies (i) and (ii). To show that (i) implies (iii), let h be a homomorphism of B into M such that $gh = 1$. We claim that

$$M = \ker g \oplus h(B).$$

First, suppose that $m \in M$. Write $m = (m - hgm) + hgm$; then $hgm \in h(B)$ and $g(m-hgm) = gm-ghgm = gm-1gm = gm-gm = 0$. Second, suppose $m \in \ker g \cap h(B)$, with $m = hb$. Then $0 = gm = ghb = 1b = b$, so $m = hb = h0 = 0$, proving the claim. Now since $\ker g = \operatorname{im} f$ by exactness, we may express any $m \in M$ in the form $m = fa + hb$. We take $um = a + b$, which makes sense because both f and h are injective and $f(A) \cap h(B) = 0$. This forces the diagram of (iii) to be commutative, and u is therefore an isomorphism by the short five lemma. Finally, we show that (ii) implies (iii). Let e be a homomorphism of M into A such that $ef = 1$. In this case, we claim that

$$M = f(A) \oplus \ker e.$$

If $m \in M$ then $m = fem + (m - fem)$ and $fem \in f(A)$, $e(m - fem) = em - efem = em - em = 0$. If $m \in f(A) \cap \ker e$, then, with $m = fa$, we have $0 = em = efa = a$, so $m = 0$, and the claim is verified. Now if $m \in M$ we have $m = fa + m'$ with $a \in A$ and $m' \in \ker e$. We take $u(m) = a + g(m') = a + gm$ since $gf = 0$. (The definition of u is unambiguous because f is injective and $f(A) \cap \ker e = 0$.) The choice of u forces the diagram to be commutative, and again u is an isomorphism by the short five lemma. ♣

4.7.5 Corollary

If the sequence

$$\begin{tikzcd} 0 \arrow[r] & A \arrow[r, "f"] & M \arrow[r, "g"] & B \arrow[r] & 0 \end{tikzcd}$$

is split exact with splitting maps e and h as in (3), then the "backwards" sequence

$$\begin{tikzcd} 0 & A \arrow[l] & M \arrow[l, "e"'] & B \arrow[l, "h"'] & 0 \arrow[l] \end{tikzcd}$$

is also split exact, with splitting maps g and f.

Proof. Simply note that $gh = 1$ and $ef = 1$. ♣

A device that I use to remember which way the splitting maps go (i.e., it's $ef = 1$, not $fe = 1$) is that the map that is applied first points inward toward the "center" M.

Problems For Section 4.7

Consider the following commutative diagram with exact rows:

$$\begin{tikzcd} 0 \arrow[r] & A \arrow[r, "f"] & B \arrow[r, "g"] \arrow[d, "v"'] & C \arrow[d, "w"'] \\ 0 \arrow[r] & A' \arrow[r, "f'"'] & B' \arrow[r, "g'"'] & C' \end{tikzcd}$$

Our objective in Problems 1–3 is to find a homomorphism $u\colon A \to A'$ such that the square $ABB'A'$, hence the entire diagram, is commutative.

1. Show that if u exists, it is unique.
2. If $a \in A$, show that $vfa \in \operatorname{im} f'$.
3. If $vfa = f'a'$, define ua appropriately.

Now consider another commutative diagram with exact rows:

$$\begin{tikzcd}
A \arrow[r, "f"] \arrow[d, "w"'] & B \arrow[r, "g"] \arrow[d, "v"'] & C \arrow[r] & 0 \\
A' \arrow[r, "f'"'] & B' \arrow[r, "g'"'] & C' \arrow[r] & 0
\end{tikzcd}$$

In Problems 4 and 5 we are to define $u\colon C \to C'$ so that the diagram will commute.

4. If $c \in C$, then since g is surjective, $c = gb$ for some $b \in B$. Write down the only possible definition of uc.
5. In Problem 4, b is not unique. Show that your definition of u does not depend on the particular b.

Problems 6–11 refer to the diagram of the short five lemma (4.7.3). Application of the four lemma is very efficient, but a direct attack is also good practice.

6. If t and v are injective, so is u.
7. If t and v are surjective, so is u.
8. If t is surjective and u is injective, then v is injective.
9. If u is surjective, so is v.

By Problems 8 and 9, if t and u are isomorphisms, so is v.

10. If u is injective, so is t.
11. If u is surjective and v is injective, then t is surjective.

Note that by Problems 10 and 11, if u and v are isomorphisms, so is t.

12. If you have not done so earlier, do Problem 8 directly, without appealing to the four lemma.
13. If you have not done so earlier, do Problem 11 directly, without appealing to the four lemma.