typical members of $\mathbb{Z}D_8$. Their sum and product are then

$$\begin{aligned}\alpha + \beta &= r - 2r^2 - 2s + rs \\ \alpha\beta &= (r + r^2 - 2s)(-3r^2 + rs) \\ &= r(-3r^2 + rs) + r^2(-3r^2 + rs) - 2s(-3r^2 + rs) \\ &= -3r^3 + r^2s - 3 + r^3s + 6r^2s - 2r^3 \\ &= -3 - 5r^3 + 7r^2s + r^3s.\end{aligned}$$

The ring R appears in RG as the "constant" formal sums i.e., the R-multiples of the identity of G (note that the definition of the addition and multiplication in RG restricted to these elements is just the addition and multiplication in R). These elements of R commute with all elements of RG. The identity of R is the identity of RG.

The group G also appears in RG (the element g_i appears as $1g_i$ — for example, $r, s \in D_8$ are also elements of the group ring $\mathbb{Z}D_8$ above) — multiplication in the ring RG restricted to G is just the group operation. In particular, each element of G has a multiplicative inverse in the ring RG (namely, its inverse in G). This says that G is a *subgroup of the group of units of* RG.

If $|G| > 1$ then RG always has zero divisors. For example, let g be any element of G of order $m > 1$. Then

$$(1 - g)(1 + g + \cdots + g^{m-1}) = 1 - g^m = 1 - 1 = 0$$

so $1 - g$ is a zero divisor (note that by definition of RG neither of the formal sums in the above product is zero).

If S is a subring of R then SG is a subring of RG. For instance, $\mathbb{Z}G$ (called the *integral group ring* of G) is a subring of $\mathbb{Q}G$ (the *rational group ring* of G). Furthermore, if H is a subgroup of G then RH is a subring of RG. The set of all elements of RG whose coefficients sum to zero is a subring (without identity). If $|G| > 1$, the set of elements with zero "constant term" (i.e., the coefficient of the identity of G is zero) is *not* a subring (it is not closed under multiplication).

Note that the group ring $\mathbb{R}Q_8$ is *not* the same ring as the Hamilton Quaternions $\mathbb{H}$ even though the latter contains a copy of the quaternion group Q_8 (under multiplication). One difference is that the unique element of order 2 in Q_8 (usually denoted by -1) is not the additive inverse of 1 in $\mathbb{R}Q_8$. In other words, if we temporarily denote the identity of the group Q_8 by g_1 and the unique element of order 2 by g_2, then $g_1 + g_2$ is not zero in $\mathbb{R}Q_8$, whereas $1 + (-1)$ is zero in $\mathbb{H}$. Furthermore, as noted above, the group ring $\mathbb{R}Q_8$ contains zero divisors hence is not a division ring.

Group rings over fields will be studied extensively in Chapter 18.

EXERCISES

Let R be a commutative ring with 1.

1. Let $p(x) = 2x^3 - 3x^2 + 4x - 5$ and let $q(x) = 7x^3 + 33x - 4$. In each of parts (a), (b) and (c) compute $p(x) + q(x)$ and $p(x)q(x)$ under the assumption that the coefficients of the two given polynomials are taken from the specified ring (where the integer coefficients are taken mod n in parts (b) and (c)):
(a) $R = \mathbb{Z}$, **(b)** $R = \mathbb{Z}/2\mathbb{Z}$, **(c)** $R = \mathbb{Z}/3\mathbb{Z}$.

2. Let $p(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$ be an element of the polynomial ring $R[x]$. Prove that $p(x)$ is a zero divisor in $R[x]$ if and only if there is a nonzero $b \in R$ such that $bp(x) = 0$. [Let $g(x) = b_m x^m + b_{m-1}x^{m-1} + \cdots + b_0$ be a nonzero polynomial of minimal degree such that $g(x)p(x) = 0$. Show that $b_m a_n = 0$ and so $a_n g(x)$ is a polynomial of degree less than m that also gives 0 when multiplied by $p(x)$. Conclude that $a_n g(x) = 0$. Apply a similar argument to show by induction on i that $a_{n-i}g(x) = 0$ for $i = 0, 1, \ldots, n$, and show that this implies $b_m p(x) = 0$.]

3. Define the set $R[[x]]$ of *formal power series* in the indeterminate x with coefficients from R to be all formal infinite sums

$$\sum_{n=0}^{\infty} a_n x^n = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \cdots.$$

Define addition and multiplication of power series in the same way as for power series with real or complex coefficients i.e., extend polynomial addition and multiplication to power series as though they were "polynomials of infinite degree":

$$\sum_{n=0}^{\infty} a_n x^n \quad + \quad \sum_{n=0}^{\infty} b_n x^n = \sum_{n=0}^{\infty} (a_n + b_n) x^n$$

$$\sum_{n=0}^{\infty} a_n x^n \quad \times \quad \sum_{n=0}^{\infty} b_n x^n = \sum_{n=0}^{\infty} \Big(\sum_{k=0}^{n} a_k b_{n-k} \Big) x^n.$$

(The term "formal" is used here to indicate that convergence is not considered, so that formal power series need not represent functions on R.)

(a) Prove that $R[[x]]$ is a commutative ring with 1.

(b) Show that $1 - x$ is a unit in $R[[x]]$ with inverse $1 + x + x^2 + \cdots$.

(c) Prove that $\sum_{n=0}^{\infty} a_n x^n$ is a unit in $R[[x]]$ if and only if a_0 is a unit in R.

4. Prove that if R is an integral domain then the ring of formal power series $R[[x]]$ is also an integral domain.

5. Let F be a field and define the ring $F((x))$ of *formal Laurent series* with coefficients from F by

$$F((x)) = \{ \sum_{n \geq N}^{\infty} a_n x^n \mid a_n \in F \text{ and } N \in \mathbb{Z} \}.$$

(Every element of $F((x))$ is a power series in x plus a polynomial in $1/x$, i.e., each element of $F((x))$ has only a finite number of terms with negative powers of x.)

(a) Prove that $F((x))$ is a field.

(b) Define the map

$$v : F((x))^{\times} \to \mathbb{Z} \qquad \text{by} \qquad v\Big(\sum_{n \geq N}^{\infty} a_n x^n \Big) = N$$

where a_N is the first nonzero coefficient of the series (i.e., N is the "order of zero or pole of the series at 0"). Prove that v is a discrete valuation on $F((x))$ whose discrete valuation ring is $F[[x]]$, the ring of formal power series (cf. Exercise 26, Section 1).

6. Let S be a ring with identity $1 \neq 0$. Let $n \in \mathbb{Z}^+$ and let A be an $n \times n$ matrix with entries from S whose i, j entry is a_{ij}. Let E_{ij} be the element of $M_n(S)$ whose i, j entry is 1 and whose other entries are all 0.

(a) Prove that $E_{ij}A$ is the matrix whose i^{th} row equals the j^{th} row of A and all other rows are zero.

(b) Prove that AE_{ij} is the matrix whose j^{th} column equals the i^{th} column of A and all other columns are zero.

(c) Deduce that $E_{pq}AE_{rs}$ is the matrix whose p, s entry is a_{qr} and all other entries are zero.

7. Prove that the center of the ring $M_n(R)$ is the set of scalar matrices (cf. Exercise 7, Section 1). [Use the preceding exercise.]

8. Let S be any ring and let $n \geq 2$ be an integer. Prove that if A is any strictly upper triangular matrix in $M_n(S)$ then $A^n = 0$ (a strictly upper triangular matrix is one whose entries on and below the main diagonal are all zero).

9. Let $\alpha = r + r^2 - 2s$ and $\beta = -3r^2 + rs$ be the two elements of the integral group ring $\mathbb{Z}D_8$ described in this section. Compute the following elements of $\mathbb{Z}D_8$:
(a) $\beta\alpha$, **(b)** α^2, **(c)** $\alpha\beta - \beta\alpha$, **(d)** $\beta\alpha\beta$.

10. Consider the following elements of the integral group ring $\mathbb{Z}S_3$:

$$\alpha = 3(1\ 2) - 5(2\ 3) + 14(1\ 2\ 3) \qquad \text{and} \qquad \beta = 6(1) + 2(2\ 3) - 7(1\ 3\ 2)$$

(where (1) is the identity of S_3). Compute the following elements:
(a) $\alpha + \beta$, **(b)** $2\alpha - 3\beta$, **(c)** $\alpha\beta$, **(d)** $\beta\alpha$, **(e)** α^2.

11. Repeat the preceding exercise under the assumption that the coefficients of α and β are in $\mathbb{Z}/3\mathbb{Z}$ (i.e., $\alpha, \beta \in \mathbb{Z}/3\mathbb{Z}S_3$).

12. Let $G = \{g_1, \ldots, g_n\}$ be a finite group. Prove that the element $N = g_1 + g_2 + \ldots + g_n$ is in the center of the group ring RG (cf. Exercise 7, Section 1).

13. Let $\mathcal{K} = \{k_1, \ldots, k_m\}$ be a conjugacy class in the finite group G.
(a) Prove that the element $K = k_1 + \ldots + k_m$ is in the center of the group ring RG (cf. Exercise 7, Section 1). [Check that $g^{-1}Kg = K$ for all $g \in G$.]
(b) Let $\mathcal{K}_1, \ldots, \mathcal{K}_r$ be the conjugacy classes of G and for each $\mathcal{K}_i$ let K_i be the element of RG that is the sum of the members of $\mathcal{K}_i$. Prove that an element $\alpha \in RG$ is in the center of RG if and only if $\alpha = a_1K_1 + a_2K_2 + \cdots + a_rK_r$ for some $a_1, a_2, \ldots, a_r \in R$.

7.3 RING HOMOMORPHISMS AND QUOTIENT RINGS

A ring homomorphism is a map from one ring to another that respects the additive and multiplicative structures:

Definition. Let R and S be rings.

(1) A *ring homomorphism* is a map $\varphi : R \to S$ satisfying
  (i) $\varphi(a + b) = \varphi(a) + \varphi(b)$ for all $a, b \in R$ (so φ is a group homomorphism on the additive groups) and
  (ii) $\varphi(ab) = \varphi(a)\varphi(b)$ for all $a, b \in R$.

(2) The *kernel* of the ring homomorphism φ, denoted $\ker\varphi$, is the set of elements of R that map to 0 in S (i.e., the kernel of φ viewed as a homomorphism of additive groups).

(3) A bijective ring homomorphism is called an *isomorphism.*

If the context is clear we shall simply use the term "homomorphism" instead of "ring homomorphism." Similarly, if A and B are rings, $A \cong B$ will always mean an isomorphism of rings unless otherwise stated.

Examples

(1) The map $\varphi : \mathbb{Z} \to \mathbb{Z}/2\mathbb{Z}$ defined by sending an even integer to 0 and an odd integer to 1 is a ring homomorphism. The map is additive since the sum of two even or odd integers is even and the sum of an even integer and an odd integer is odd. The map is multiplicative since the product of two odd integers is odd and the product of an even integer with any integer is even. The kernel of φ (the fiber of φ above $0 \in \mathbb{Z}/2\mathbb{Z}$) is the set of even integers. The fiber of φ above $1 \in \mathbb{Z}/2\mathbb{Z}$ is the set of odd integers.

(2) For $n \in \mathbb{Z}$ the maps $\varphi_n : \mathbb{Z} \to \mathbb{Z}$ defined by $\varphi_n(x) = nx$ are *not* in general ring homomorphisms because $\varphi_n(xy) = nxy$ whereas $\varphi_n(x)\varphi_n(y) = nxny = n^2xy$. Hence φ_n is a ring homomorphism only when $n^2 = n$, i.e., $n = 0, 1$. Note however that φ_n is always a *group homomorphism* on the additive groups. Thus care should be exercised when dealing with rings to be sure to check that *both* ring operations are preserved. Note that φ_0 is the zero homomorphism and φ_1 is the identity homomorphism.

(3) Let $\varphi : \mathbb{Q}[x] \to \mathbb{Q}$ be the map from the ring of polynomials in x with rational coefficients to the rationals defined by $\varphi(p(x)) = p(0)$ (i.e., mapping the polynomial to its constant term). Then φ is a ring homomorphism since the constant term of the sum of two polynomials is the sum of their constant terms and the constant term of the product of two polynomials is the product of their constant terms. The fiber above $a \in \mathbb{Q}$ consists of the set of polynomials with a as constant term. In particular, the kernel of φ consists of the polynomials with constant term 0.

Proposition 5. Let R and S be rings and let $\varphi : R \to S$ be a homomorphism.

(1) The image of φ is a subring of S.

(2) The kernel of φ is a subring of R. Furthermore, if $\alpha \in \ker\varphi$ then $r\alpha$ and $\alpha r \in \ker\varphi$ for every $r \in R$, i.e., $\ker\varphi$ is closed under multiplication by elements from R.

Proof: (1) If $s_1, s_2 \in \operatorname{im}\varphi$ then $s_1 = \varphi(r_1)$ and $s_2 = \varphi(r_2)$ for some $r_1, r_2 \in R$. Then $\varphi(r_1 - r_2) = s_1 - s_2$ and $\varphi(r_1r_2) = s_1s_2$. This shows $s_1 - s_2, s_1s_2 \in \operatorname{im}\varphi$, so the image of φ is closed under subtraction and under multiplication, hence is a subring of S.

(2) If $\alpha, \beta \in \ker\varphi$ then $\varphi(\alpha) = \varphi(\beta) = 0$. Hence $\varphi(\alpha - \beta) = 0$ and $\varphi(\alpha\beta) = 0$, so $\ker\varphi$ is closed under subtraction and under multiplication, so is a subring of R. Similarly, for any $r \in R$ we have $\varphi(r\alpha) = \varphi(r)\varphi(\alpha) = \varphi(r)\ 0 = 0$, and also $\varphi(\alpha r) = \varphi(\alpha)\varphi(r) = 0\ \varphi(r) = 0$, so $r\alpha, \alpha r \in \ker\varphi$.

In the case of a homomorphism φ of groups we saw that the fibers of the homomorphism have the structure of a group naturally isomorphic to the image of φ, which led to the notion of a quotient group by a normal subgroup. An analogous result is true for a homomorphism of rings.

Let $\varphi : R \to S$ be a ring homomorphism with kernel I. Since R and S are in particular additive abelian groups, φ is in particular a homomorphism of abelian groups

and the fibers of φ are the additive cosets $r + I$ of the kernel I (more precisely, if r is any element of R mapping to $a \in S$, $\varphi(r) = a$, then the fiber of φ over a is the coset $r + I$ of the kernel I). These fibers have the structure of a ring naturally isomorphic to the image of φ: if X is the fiber over $a \in S$ and Y is the fiber over $b \in S$, then $X + Y$ is the fiber over $a + b$ and XY is the fiber over ab. In terms of cosets of the kernel I this addition and multiplication is

$$(r + I) + (s + I) = (r + s) + I \tag{7.1}$$

$$(r + I) \times (s + I) = (rs) + I. \tag{7.2}$$

As in the case for groups, the verification that these operations define a ring structure on the collection of cosets of the kernel I ultimately rests on the corresponding ring properties of S. This ring of cosets is called the *quotient ring* of R by $I = \ker \varphi$ and is denoted R/I. Note that the additive structure of the ring R/I is just the additive quotient group of the additive abelian group R by the (necessarily normal) subgroup I. When I is the kernel of some homomorphism φ this additive abelian quotient group also has a multiplicative structure, defined by (7.2), which makes R/I into a ring.

As in the case for groups, we can also consider whether (1) and (2) can be used to define a ring structure on the collection of cosets of an *arbitrary* subgroup I of R. Note that since R is an abelian additive group, the subgroup I is necessarily normal so that the quotient R/I of cosets of I is automatically an additive abelian group. The question then is whether this quotient group also has a *multiplicative* structure induced from the multiplication in R, defined by (2). The answer is no in general (just as the answer is no in trying to form the quotient by an arbitrary subgroup of a group), which leads to the notion of an *ideal* in R (the analogue for rings of a normal subgroup of a group). We shall then see that the ideals of R are exactly the kernels of the ring homomorphisms of R (the analogue for rings of the characterization of normal subgroups as the kernels of group homomorphisms).

Let I be an arbitrary subgroup of the additive group R. We consider when the multiplication of cosets in (2) is well defined and makes the additive abelian group R/I into a ring. The statement that the multiplication in (2) is well defined is the statement that the multiplication is independent of the particular representatives r and s chosen, i.e., that we obtain the same coset on the right if instead we use the representatives $r + \alpha$ and $s + \beta$ for any $\alpha, \beta \in I$. In other words, we must have

$$(r + \alpha)(s + \beta) + I = rs + I \tag{$*$}$$

for all $r, s \in R$ and all $\alpha, \beta \in I$.

Letting $r = s = 0$, we see that I must be closed under multiplication, i.e., I must be a *subring* of R.

Next, by letting $s = 0$ and letting r be arbitrary, we see that we must have $r\beta \in I$ for every $r \in R$ and every $\beta \in I$, i.e., that I must be closed under multiplication on the left by elements from R. Letting $r = 0$ and letting s be arbitrary, we see similarly that I must be closed under multiplication on the right by elements from R.

Conversely, if I is closed under multiplication on the left and on the right by elements from R then the relation $(*)$ is satisfied for all $\alpha, \beta \in I$. Hence this is a necessary and sufficient condition for the multiplication in (2) to be well defined.

Finally, if the multiplication of cosets defined by (2) is well defined, then this multiplication makes the additive quotient group R/I into a ring. Each ring axiom in the quotient follows directly from the corresponding axiom in R. For example, one of the distributive laws is verified as follows:

$$\begin{aligned}(r+I)[(s+I)+(t+I)] &= (r+I)[(s+t)+I] \\ &= r(s+t)+I = (rs+rt)+I \\ &= (rs+I)+(rt+I) \\ &= [(r+I)(s+I)]+[(r+I)(t+I)].\end{aligned}$$

This shows that the quotient R/I of the ring R by a subgroup I has a natural ring structure if and only if I is also closed under multiplication on the left and on the right by elements from R (so in particular must be a subring of R since it is closed under multiplication). As mentioned, such subrings I are called the *ideals* of R:

Definition. Let R be a ring, let I be a subset of R and let $r \in R$.

(1) $rI = \{ra \mid a \in I\}$ and $Ir = \{ar \mid a \in I\}$.

(2) A subset I of R is a *left ideal* of R if

(i) I is a subring of R, and

(ii) I is closed under left multiplication by elements from R, i.e., $rI \subseteq I$ for all $r \in R$.

Similarly I is a *right ideal* if (i) holds and in place of (ii) one has

(ii)′ I is closed under right multiplication by elements from R, i.e., $Ir \subseteq I$ for all $r \in R$.

(3) A subset I that is both a left ideal and a right ideal is called an *ideal* (or, for added emphasis, a *two-sided ideal*) of R.

For commutative rings the notions of left, right and two-sided ideal coincide. We emphasize that to prove a subset I of a ring R is an ideal it is necessary to prove that I is nonempty, closed under subtraction and closed under multiplication by all the elements of R (and not just by elements of I). If R has a 1 then $(-1)a = -a$ so in this case I is an ideal if it is nonempty, closed under addition and closed under multiplication by all the elements of R.

Note also that the last part of Proposition 5 proves that the kernel of any ring homomorphism is an ideal.

We summarize the preceding discussion in the following proposition.

Proposition 6. Let R be a ring and let I be an ideal of R. Then the (additive) quotient group R/I is a ring under the binary operations:

$$(r+I)+(s+I) = (r+s)+I \qquad \text{and} \qquad (r+I)\times(s+I) = (rs)+I$$

for all $r, s \in R$. Conversely, if I is any subgroup such that the above operations are well defined, then I is an ideal of R.

Definition. When I is an ideal of R the ring R/I with the operations in the previous proposition is called the *quotient ring* of R by I.

Theorem 7.

(1) *(The First Isomorphism Theorem for Rings)* If $\varphi : R \to S$ is a homomorphism of rings, then the kernel of φ is an ideal of R, the image of φ is a subring of S and $R/\ker\varphi$ is isomorphic as a ring to $\varphi(R)$.

(2) If I is any ideal of R, then the map

$$R \to R/I \qquad \text{defined by} \qquad r \mapsto r + I$$

is a surjective ring homomorphism with kernel I (this homomorphism is called the *natural projection* of R onto R/I). Thus every ideal is the kernel of a ring homomorphism and vice versa.

Proof: This is just a matter of collecting previous calculations. If I is the kernel of φ, then the cosets (under addition) of I are precisely the fibers of φ. In particular, the cosets $r+I$, $s+I$ and $rs+I$ are the fibers of φ over $\varphi(r)$, $\varphi(s)$ and $\varphi(rs)$, respectively. Since φ is a ring homomorphism $\varphi(r)\varphi(s) = \varphi(rs)$, hence $(r+I)(s+I) = rs+I$. Multiplication of cosets is well defined and so I is an ideal and R/I is a ring. The correspondence $r + I \mapsto \varphi(r)$ is a bijection between the rings R/I and $\varphi(R)$ which respects addition and multiplication, hence is a ring isomorphism.

If I is any ideal, then R/I is a ring (in particular is an abelian group) and the map $\pi : r \mapsto r + I$ is a group homomorphism with kernel I. It remains to check that π is a ring homomorphism. This is immediate from the definition of multiplication in R/I:

$$\pi : rs \mapsto rs + I = (r + I)(s + I) = \pi(r)\pi(s).$$

As with groups we shall often use the bar notation for reduction mod I: $\bar{r} = r + I$. With this notation the addition and multiplication in the quotient ring R/I become simply $\bar{r} + \bar{s} = \overline{r+s}$ and $\bar{r}\,\bar{s} = \overline{rs}$.

Examples

Let R be a ring.

(1) The subrings R and $\{0\}$ are ideals. An ideal I is *proper* if $I \neq R$. The ideal $\{0\}$ is called the *trivial ideal* and is denoted by 0.

(2) It is immediate that $n\mathbb{Z}$ is an ideal of $\mathbb{Z}$ for any $n \in \mathbb{Z}$ and these are the only ideals of $\mathbb{Z}$ since in particular these are the only subgroups of $\mathbb{Z}$. The associated quotient ring is $\mathbb{Z}/n\mathbb{Z}$ (which explains the choice of notation and which we have now proved is a ring), introduced in Chapter 0. For example, if $n = 15$ then the elements of $\mathbb{Z}/15\mathbb{Z}$ are the cosets $\bar{0}, \bar{1}, \ldots, \overline{13}, \overline{14}$. To add (or multiply) in the quotient, simply choose any representatives for the two cosets, add (multiply, respectively) these representatives in the integers $\mathbb{Z}$, and take the corresponding coset containing this sum (product, respectively). For example, $\bar{7} + \overline{11} = \overline{18}$ and $\overline{18} = \bar{3}$, so $\bar{7} + \overline{11} = \bar{3}$ in $\mathbb{Z}/15\mathbb{Z}$. Similarly, $\bar{7}\,\overline{11} = \overline{77} = \bar{2}$ in $\mathbb{Z}/15\mathbb{Z}$. We could also express this by writing $7 + 11 \equiv 3 \bmod 15$, $7(11) \equiv 2 \bmod 15$.

The natural projection $\mathbb{Z} \to \mathbb{Z}/n\mathbb{Z}$ is called *reduction mod n* and will be discussed further at the end of these examples.

(3) Let $R = \mathbb{Z}[x]$ be the ring of polynomials in x with integer coefficients. Let I be the collection of polynomials whose terms are of degree at least 2 (i.e., having no terms of degree 0 or degree 1) together with the zero polynomial. Then I is an ideal: the sum of two such polynomials again has terms of degree at least 2 and the product of a polynomial whose terms are of degree at least 2 with *any* polynomial again only has terms of degree at least 2. Two polynomials $p(x), q(x)$ are in the same coset of I if and only if they differ by a polynomial whose terms are of degree at least 2, i.e., if and only if $p(x)$ and $q(x)$ have the same constant and first degree terms. For example, the polynomials $3 + 5x + x^3 + x^5$ and $3 + 5x - x^4$ are in the same coset of I. It follows easily that a complete set of representatives for the quotient R/I is given by the polynomials $a + bx$ of degree at most 1.

Addition and multiplication in the quotient are again performed by representatives. For example,

$$(\overline{1 + 3x}) + (\overline{-4 + 5x}) = \overline{-3 + 8x}$$

and

$$(\overline{1 + 3x})(\overline{-4 + 5x}) = \overline{(-4 - 7x + 15x^2)} = \overline{-4 - 7x}\,.$$

Note that in this quotient ring R/I we have $\bar{x}\,\bar{x} = \overline{x^2} = \bar{0}$, for example, so that R/I has zero divisors, even though $R = \mathbb{Z}[x]$ does not.

(4) Let A be a ring, let X be any nonempty set and let R be the ring of all functions from X to A. For each fixed $c \in X$ the map

$$E_c : R \to A \quad \text{defined by} \quad E_c(f) = f(c)$$

(called *evaluation at c*) is a ring homomorphism because the operations in R are pointwise addition and multiplication of functions. The kernel of E_c is given by $\{f \in R \mid f(c) = 0\}$ (the set of functions from X to A that vanish at c). Also, E_c is surjective: given any $a \in A$ the constant function $f(x) = a$ maps to a under evaluation at c. Thus $R/\ker E_c \cong A$.

Similarly, let X be the closed interval $[0,1]$ in $\mathbb{R}$ and let R be the ring of all continuous real valued functions on $[0,1]$. For each $c \in [0, 1]$, evaluation at c is a surjective ring homomorphism (since R contains the constant functions) and so $R/\ker E_c \cong \mathbb{R}$. The kernel of E_c is the ideal of all continuous functions whose graph crosses the x-axis at c. More generally, the fiber of E_c above the real number y_0 is the set of all continuous functions that pass through the point (c, y_0).

(5) The map from the polynomial ring $R[x]$ to R defined by $p(x) \mapsto p(0)$ (evaluation at 0) is a ring homomorphism whose kernel is the set of all polynomials whose constant term is zero, i.e., $p(0) = 0$. We can compose this homomorphism with any homomorphism from R to another ring S to obtain a ring homomorphism from $R[x]$ to S. For example, let $R = \mathbb{Z}$ and consider the homomorphism $\mathbb{Z}[x] \to \mathbb{Z}/2\mathbb{Z}$ defined by the composition $p(x) \mapsto p(0) \mapsto p(0) \bmod 2 \in \mathbb{Z}/2\mathbb{Z}$. The kernel of this composite map is given by $\{p(x) \in \mathbb{Z}[x] \mid p(0) \in 2\mathbb{Z}\}$, i.e., the set of all polynomials with integer coefficients whose constant term is even. The other fiber of this homomorphism is the coset of polynomials whose constant term is odd, as we determined earlier. Since the homomorphism is clearly surjective, the quotient ring is $\mathbb{Z}/2\mathbb{Z}$.

(6) Fix some $n \in \mathbb{Z}$ with $n \geq 2$ and consider the noncommutative ring $M_n(R)$. If J is any ideal of R then $M_n(J)$, the $n \times n$ matrices whose entries come from J, is a two-sided ideal of $M_n(R)$. This ideal is the kernel of the surjective homomorphism $M_n(R) \to M_n(R/J)$ which reduces each entry of a matrix mod J, i.e., which maps each entry a_{ij} to $\overline{a_{ij}}$ (here bar denotes passage to R/J). For instance, when $n = 3$ and $R = \mathbb{Z}$, the 3×3 matrices whose entries are all even is the two-sided ideal $M_3(2\mathbb{Z})$

of $M_3(\mathbb{Z})$ and the quotient $M_3(\mathbb{Z})/M_3(2\mathbb{Z})$ is isomorphic to $M_3(\mathbb{Z}/2\mathbb{Z})$. If the ring R has an identity then the exercises below show that every two-sided ideal of $M_n(R)$ is of the form $M_n(J)$ for some two-sided ideal J of R.

(7) Let R be a commutative ring with 1 and let $G = \{g_1, \ldots, g_n\}$ be a finite group. The map from the group ring RG to R defined by $\sum_{i=1}^n a_i g_i \mapsto \sum_{i=1}^n a_i$ is easily seen to be a homomorphism, called the *augmentation map* . The kernel of the augmentation map, the *augmentation ideal*, is the set of elements of RG whose coefficients sum to 0. For example, $g_i - g_j$ is an element of the augmentation ideal for all i, j. Since the augmentation map is surjective, the quotient ring is isomorphic to R.

Another ideal in RG is $\{\sum_{i=1}^n a g_i \mid a \in R\}$, i.e., the formal sums whose coefficients are all equal (equivalently, all R-multiples of the element $g_1 + \cdots + g_n$).

(8) Let R be a commutative ring with identity $1 \neq 0$ and let $n \in \mathbb{Z}$ with $n \geq 2$. We exhibit some one-sided ideals in the ring $M_n(R)$. For each $j \in \{1, 2, \ldots, n\}$ let L_j be the set of all $n \times n$ matrices in $M_n(R)$ with arbitrary entries in the j^{th} column and zeros in all other columns. It is clear that L_j is closed under subtraction. It follows directly from the definition of matrix multiplication that for any matrix $T \in M_n(R)$ and any $A \in L_j$ the product TA has zero entries in the i^{th} column for all $i \neq j$. This shows L_j is a *left ideal* of $M_n(R)$. Moreover, L_j is *not* a *right* ideal (hence is not a two-sided ideal). To see this, let E_{pq} be the matrix with 1 in the p^{th} row and q^{th} column and zeros elsewhere $(p, q \in \{1, \ldots, n\})$. Then $E_{1j} \in L_j$ but $E_{1j}E_{ji} = E_{1i} \notin L_j$ if $i \neq j$, so L_j is not closed under right multiplication by arbitrary ring elements. An analogous argument shows that if R_j is the set of all $n \times n$ matrices in $M_n(R)$ with arbitrary entries in the j^{th} row and zeros in all other rows, then R_j is a *right* ideal which is not a *left* ideal. These one-sided ideals will play an important role in Part VI.

Example: (The Reduction Homomorphism)

The canonical projection map from $\mathbb{Z}$ to $\mathbb{Z}/n\mathbb{Z}$ obtained by factoring out by the ideal $n\mathbb{Z}$ of $\mathbb{Z}$ is usually referred to as "reducing modulo n." The fact that this is a *ring homomorphism* has important consequences for elementary number theory. For example, suppose we are trying to solve the equation

$$x^2 + y^2 = 3z^2$$

in *integers* x, y and z (such problems are frequently referred to as *Diophantine equations* after Diophantus, who was one of the first to systematically examine the existence of *integer* solutions of equations). Suppose such integers exist. Observe first that we may assume x, y and z have no factors in common, since otherwise we could divide through this equation by the square of this common factor and obtain another set of integer solutions smaller than the initial ones. This equation simply states a relation between these elements in the *ring* $\mathbb{Z}$. As such, the same relation must also hold in any *quotient* ring as well. In particular, this relation must hold in $\mathbb{Z}/n\mathbb{Z}$ for any integer n. The choice $n = 4$ is particularly efficacious, for the following reason: the squares mod 4 are just $0^2, 1^2, 2^2, 3^2$, i.e., 0, 1 (mod 4). Reading the above equation mod 4 (that is, considering this equation in the quotient ring $\mathbb{Z}/4\mathbb{Z}$), we must have

$$\begin{Bmatrix} 0 \\ 1 \end{Bmatrix} + \begin{Bmatrix} 0 \\ 1 \end{Bmatrix} \equiv 3 \begin{Bmatrix} 0 \\ 1 \end{Bmatrix} \equiv \begin{Bmatrix} 0 \\ 3 \end{Bmatrix} \pmod 4$$

where the $\begin{Bmatrix} 0 \\ 1 \end{Bmatrix}$, for example, indicates that either a 0 or a 1 may be taken. Checking the few possibilities shows that we must take the 0 each time. This means that each

of x, y and z must be even integers (squares of the odd integers gave us 1 mod 4). But this contradicts the assumption of no common factors for these integers, and shows that this equation has *no solutions in nonzero integers.*

Note that even had solutions existed, this technique gives information about the possible residues of the solutions mod n (since we could just as well have examined the possibilities mod n as mod 4) and note that for each choice of n we have only a *finite* problem to solve because there are only finitely many residue classes mod n. Together with the Chinese Remainder Theorem (described in Section 6), we can then determine the possible solutions modulo very large integers, which greatly assists in finding them numerically (when they exist). We also observe that this technique has a number of limitations — for example, there are equations which have solutions modulo every integer, but which do not have integer solutions. An easy example (but extremely hard to verify that it does indeed have this property) is the equation

$$3x^3 + 4y^3 + 5z^3 = 0.$$

As a final example of this technique, we mention that the map from the ring $\mathbb{Z}[x]$ of polynomials with integer coefficients to the ring $\mathbb{Z}/p\mathbb{Z}[x]$ of polynomials with coefficients in $\mathbb{Z}/p\mathbb{Z}$ for a prime p given by *reducing the coefficients modulo p* is a ring homomorphism. This example of reduction will be used in Chapter 9 in trying to determine whether polynomials can be factored.

The following theorem gives the remaining Isomorphism Theorems for rings. Each of these may be proved as follows: first use the corresponding theorem from group theory to obtain an isomorphism of *additive groups* (or correspondence of groups, in the case of the Fourth Isomorphism Theorem) and then check that this group isomorphism (or correspondence, respectively) is a multiplicative map, and so defines a *ring* isomorphism. In each case the verification is immediate from the definition of multiplication in quotient rings. For example, the map that gives the isomorphism in (2) below is defined by $\varphi : r + I \mapsto r + J$. This map is multiplicative since $(r_1 + I)(r_2 + I) = r_1r_2 + I$ by the definition of the multiplication in the quotient ring R/I, and $r_1r_2 + I \mapsto r_1r_2 + J = (r_1 + J)(r_2 + J)$ by the definition of the multiplication in the quotient ring R/J, i.e., $\varphi(r_1r_2) = \varphi(r_1)\varphi(r_2)$. The proofs for the other parts of the theorem are similar.

Theorem 8. Let R be a ring.

(1) *(The Second Isomorphism Theorem for Rings)* Let A be a subring and let B be an ideal of R. Then $A + B = \{a + b \mid a \in A,\ b \in B\}$ is a subring of R, $A \cap B$ is an ideal of A and $(A + B)/B \cong A/(A \cap B)$.

(2) *(The Third Isomorphism Theorem for Rings)* Let I and J be ideals of R with $I \subseteq J$. Then J/I is an ideal of R/I and $(R/I)/(J/I) \cong R/J$.

(3) *(The Fourth or Lattice Isomorphism Theorem for Rings)* Let I be an ideal of R. The correspondence $A \leftrightarrow A/I$ is an inclusion preserving bijection between the set of subrings A of R that contain I and the set of subrings of R/I. Furthermore, A (a subring containing I) is an ideal of R if and only if A/I is an ideal of R/I.

Let $R = \mathbb{Z}$ and let I be the ideal $12\mathbb{Z}$. The quotient ring $\overline{R} = R/I = \mathbb{Z}/12\mathbb{Z}$ has ideals $\overline{R}$, $2\mathbb{Z}/12\mathbb{Z}$, $3\mathbb{Z}/12\mathbb{Z}$, $4\mathbb{Z}/12\mathbb{Z}$, $6\mathbb{Z}/12\mathbb{Z}$, and $\overline{0} = 12\mathbb{Z}/12\mathbb{Z}$ corresponding to the ideals $R = \mathbb{Z}$, $2\mathbb{Z}$, $3\mathbb{Z}$, $4\mathbb{Z}$, $6\mathbb{Z}$ and $12\mathbb{Z} = I$ of R containing I, respectively.

If I and J are ideals in the ring R then the set of sums $a + b$ with $a \in I$ and $b \in J$ is not only a subring of R (as in the Second Isomorphism Theorem for Rings), but is an *ideal* in R (the set is clearly closed under sums and $r(a + b) = ra + rb \in I + J$ since $ra \in I$ and $rb \in J$). We can also define the product of two ideals:

Definition. Let I and J be ideals of R.

(1) Define the *sum* of I and J by $I + J = \{a + b \mid a \in I,\ b \in J\}$.

(2) Define the *product* of I and J, denoted by IJ, to be the set of all finite sums of elements of the form ab with $a \in I$ and $b \in J$.

(3) For any $n \geq 1$, define the n^{th} *power* of I, denoted by I^n, to be the set consisting of all finite sums of elements of the form $a_1 a_2 \cdots a_n$ with $a_i \in I$ for all i. Equivalently, I^n is defined inductively by defining $I^1 = I$, and $I^n = II^{n-1}$ for $n = 2, 3, \ldots$.

It is easy to see that the sum $I + J$ of the ideals I and J is the smallest ideal of R containing both I and J and that the product IJ is an ideal contained in $I \cap J$ (but may be strictly smaller, cf. the exercises). Note also that the elements of the product ideal IJ are *finite sums* of products of elements ab from I and J. The set $\{ab \mid a \in I,\ b \in J\}$ consisting just of products of elements from I and J is in general not closed under addition, hence is not in general an ideal.

Examples

(1) Let $I = 6\mathbb{Z}$ and $J = 10\mathbb{Z}$ in $\mathbb{Z}$. Then $I + J$ consists of all integers of the form $6x + 10y$ with $x, y \in \mathbb{Z}$. Since every such integer is divisible by 2, the ideal $I + J$ is contained in $2\mathbb{Z}$. On the other hand, $2 = 6(2) + 10(-1)$ shows that the ideal $I + J$ contains the ideal $2\mathbb{Z}$, so that $6\mathbb{Z} + 10\mathbb{Z} = 2\mathbb{Z}$. In general, $m\mathbb{Z} + n\mathbb{Z} = d\mathbb{Z}$, where d is the greatest common divisor of m and n. The product IJ consists of all finite sums of elements of the form $(6x)(10y)$ with $x, y \in \mathbb{Z}$, which clearly gives the ideal $60\mathbb{Z}$.

(2) Let I be the ideal in $\mathbb{Z}[x]$ consisting of the polynomials with integer coefficients whose constant term is even (cf. Example 5). The two polynomials 2 and x are contained in I, so both $4 = 2 \cdot 2$ and $x^2 = x \cdot x$ are elements of the product ideal $I^2 = II$, as is their sum $x^2 + 4$. It is easy to check, however, that $x^2 + 4$ cannot be written as a single product $p(x)q(x)$ of two elements of I.

EXERCISES

Let R be a ring with identity $1 \neq 0$.

1. Prove that the rings $2\mathbb{Z}$ and $3\mathbb{Z}$ are not isomorphic.

2. Prove that the rings $\mathbb{Z}[x]$ and $\mathbb{Q}[x]$ are not isomorphic.

3. Find all homomorphic images of $\mathbb{Z}$.

4. Find all ring homomorphisms from $\mathbb{Z}$ to $\mathbb{Z}/30\mathbb{Z}$. In each case describe the kernel and the image.
5. Describe all ring homomorphisms from the ring $\mathbb{Z} \times \mathbb{Z}$ to $\mathbb{Z}$. In each case describe the kernel and the image.
6. Decide which of the following are ring homomorphisms from $M_2(\mathbb{Z})$ to $\mathbb{Z}$:
   (a) $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \mapsto a$ (projection onto the 1,1 entry)
   (b) $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \mapsto a + d$ (the *trace* of the matrix)
   (c) $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \mapsto ad - bc$ (the *determinant* of the matrix).
7. Let $R = \{\begin{pmatrix} a & b \\ 0 & d \end{pmatrix} \mid a, b, d \in \mathbb{Z}\}$ be the subring of $M_2(\mathbb{Z})$ of upper triangular matrices. Prove that the map
$$\varphi : R \to \mathbb{Z} \times \mathbb{Z} \quad \text{defined by} \quad \varphi : \begin{pmatrix} a & b \\ 0 & d \end{pmatrix} \mapsto (a, d)$$
is a surjective homomorphism and describe its kernel.
8. Decide which of the following are ideals of the ring $\mathbb{Z} \times \mathbb{Z}$:
   (a) $\{(a, a) \mid a \in \mathbb{Z}\}$
   (b) $\{(2a, 2b) \mid a, b \in \mathbb{Z}\}$
   (c) $\{(2a, 0) \mid a \in \mathbb{Z}\}$
   (d) $\{(a, -a) \mid a \in \mathbb{Z}\}$.
9. Decide which of the sets in Exercise 6 of Section 1 are ideals of the ring of all functions from [0,1] to $\mathbb{R}$.
10. Decide which of the following are ideals of the ring $\mathbb{Z}[x]$:
    (a) the set of all polynomials whose constant term is a multiple of 3
    (b) the set of all polynomials whose coefficient of x^2 is a multiple of 3
    (c) the set of all polynomials whose constant term, coefficient of x and coefficient of x^2 are zero
    (d) $\mathbb{Z}[x^2]$ (i.e., the polynomials in which only even powers of x appear)
    (e) the set of polynomials whose coefficients sum to zero
    (f) the set of polynomials $p(x)$ such that $p'(0) = 0$, where $p'(x)$ is the usual first derivative of $p(x)$ with respect to x.
11. Let R be the ring of all continuous real valued functions on the closed interval $[0, 1]$. Prove that the map $\varphi : R \to \mathbb{R}$ defined by $\varphi(f) = \int_0^1 f(t)dt$ is a homomorphism of additive groups but not a ring homomorphism.
12. Let D be an integer that is not a perfect square in $\mathbb{Z}$ and let $S = \{\begin{pmatrix} a & b \\ Db & a \end{pmatrix} \mid a, b \in \mathbb{Z}\}$.
    (a) Prove that S is a subring of $M_2(\mathbb{Z})$.
    (b) If D is not a perfect square in $\mathbb{Z}$ prove that the map $\varphi : \mathbb{Z}[\sqrt{D}] \to S$ defined by $\varphi(a + b\sqrt{D}) = \begin{pmatrix} a & b \\ Db & a \end{pmatrix}$ is a ring isomorphism.
    (c) If $D \equiv 1 \bmod 4$ is squarefree, prove that the set $\{\begin{pmatrix} a & b \\ (D-1)b/4 & a+b \end{pmatrix} \mid a, b \in \mathbb{Z}\}$
    is a subring of $M_2(\mathbb{Z})$ and is isomorphic to the quadratic integer ring $\mathcal{O}$.

13. Prove that the ring $M_2(\mathbb{R})$ contains a subring that is isomorphic to $\mathbb{C}$.

14. Prove that the ring $M_4(\mathbb{R})$ contains a subring that is isomorphic to the real Hamilton Quaternions, $\mathbb{H}$.

15. Let X be a nonempty set and let $\mathcal{P}(X)$ be the Boolean ring of all subsets of X defined in Exercise 21 of Section 1. Let R be the ring of all functions from X into $\mathbb{Z}/2\mathbb{Z}$. For each $A \in \mathcal{P}(X)$ define the function

$$\chi_A : X \to \mathbb{Z}/2\mathbb{Z} \qquad \text{by} \qquad \chi_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases}$$

(χ_A is called the *characteristic function of* A with values in $\mathbb{Z}/2\mathbb{Z}$). Prove that the map $\mathcal{P}(X) \to R$ defined by $A \mapsto \chi_A$ is a ring isomorphism.

16. Let $\varphi : R \to S$ be a surjective homomorphism of rings. Prove that the image of the center of R is contained in the center of S (cf. Exercise 7 of Section 1).

17. Let R and S be nonzero rings with identity and denote their respective identities by 1_R and 1_S. Let $\varphi : R \to S$ be a nonzero homomorphism of rings.

(a) Prove that if $\varphi(1_R) \neq 1_S$ then $\varphi(1_R)$ is a zero divisor in S. Deduce that if S is an integral domain then every ring homomorphism from R to S sends the identity of R to the identity of S.

(b) Prove that if $\varphi(1_R) = 1_S$ then $\varphi(u)$ is a unit in S and that $\varphi(u^{-1}) = \varphi(u)^{-1}$ for each unit u of R.

18. **(a)** If I and J are ideals of R prove that their intersection $I \cap J$ is also an ideal of R.

(b) Prove that the intersection of an arbitrary nonempty collection of ideals is again an ideal (do not assume the collection is countable).

19. Prove that if $I_1 \subseteq I_2 \subseteq \cdots$ are ideals of R then $\cup_{n=1}^{\infty} I_n$ is an ideal of R.

20. Let I be an ideal of R and let S be a subring of R. Prove that $I \cap S$ is an ideal of S. Show by example that not every ideal of a subring S of a ring R need be of the form $I \cap S$ for some ideal I of R.

21. Prove that every (two-sided) ideal of $M_n(R)$ is equal to $M_n(J)$ for some (two-sided) ideal J of R. [Use Exercise 6(c) of Section 2 to show first that the set of entries of matrices in an ideal of $M_n(R)$ form an ideal in R.]

22. Let a be an element of the ring R.

(a) Prove that $\{x \in R \mid ax = 0\}$ is a right ideal and $\{y \in R \mid ya = 0\}$ is a left ideal (called respectively the right and left *annihilators* of a in R).

(b) Prove that if L is a left ideal of R then $\{x \in R \mid xa = 0 \text{ for all } a \in L\}$ is a two-sided ideal (called the left *annihilator* of L in R).

23. Let S be a subring of R and let I be an ideal of R. Prove that if $S \cap I = 0$ then $\overline{S} \cong S$, where the bar denotes passage to R/I.

24. Let $\varphi : R \to S$ be a ring homomorphism.

(a) Prove that if J is an ideal of S then $\varphi^{-1}(J)$ is an ideal of R. Apply this to the special case when R is a subring of S and φ is the inclusion homomorphism to deduce that if J is an ideal of S then $J \cap R$ is an ideal of R.

(b) Prove that if φ is surjective and I is an ideal of R then $\varphi(I)$ is an ideal of S. Give an example where this fails if φ is not surjective.

25. Assume R is a commutative ring with 1. Prove that the Binomial Theorem

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

holds in R, where the binomial coefficient $\binom{n}{k}$ is interpreted in R as the sum $1+1+\cdots+1$ of the identity 1 in R taken $\binom{n}{k}$ times.

26. The *characteristic* of a ring R is the smallest positive integer n such that $1+1+\cdots+1=0$ (n times) in R; if no such integer exists the characteristic of R is said to be 0. For example, $\mathbb{Z}/n\mathbb{Z}$ is a ring of characteristic n for each positive integer n and $\mathbb{Z}$ is a ring of characteristic 0.

(a) Prove that the map $\mathbb{Z} \to R$ defined by

$$k \mapsto \begin{cases} 1+1+\cdots+1 \ (k \text{ times}) & \text{if } k > 0 \\ 0 & \text{if } k = 0 \\ -1-1-\cdots-1 \ (-k \text{ times}) & \text{if } k < 0 \end{cases}$$

is a ring homomorphism whose kernel is $n\mathbb{Z}$, where n is the characteristic of R (this explains the use of the terminology "characteristic 0" instead of the archaic phrase "characteristic ∞" for rings in which no sum of 1's is zero).

(b) Determine the characteristics of the rings $\mathbb{Q}$, $\mathbb{Z}[x]$, $\mathbb{Z}/n\mathbb{Z}[x]$.

(c) Prove that if p is a prime and if R is a commutative ring of characteristic p, then $(a+b)^p = a^p + b^p$ for all $a, b \in R$.

27. Prove that a nonzero Boolean ring has characteristic 2 (cf. Exercise 15, Section 1).

28. Prove that an integral domain has characteristic p, where p is either a prime or 0 (cf. Exercise 26).

29. Let R be a commutative ring. Recall (cf. Exercise 13, Section 1) that an element $x \in R$ is nilpotent if $x^n = 0$ for some $n \in \mathbb{Z}^+$. Prove that the set of nilpotent elements form an ideal — called the *nilradical* of R and denoted by $\mathfrak{N}(R)$. [Use the Binomial Theorem to show $\mathfrak{N}(R)$ is closed under addition.]

30. Prove that if R is a commutative ring and $\mathfrak{N}(R)$ is its nilradical (cf. the preceding exercise) then zero is the only nilpotent element of $R/\mathfrak{N}(R)$ i.e., prove that $\mathfrak{N}(R/\mathfrak{N}(R)) = 0$.

31. Prove that the elements $\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ and $\begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$ are nilpotent elements of $M_2(\mathbb{Z})$ whose sum is not nilpotent (note that these two matrices do not commute). Deduce that the set of nilpotent elements in the noncommutative ring $M_2(\mathbb{Z})$ is not an ideal.

32. Let $\varphi : R \to S$ be a homomorphism of rings. Prove that if x is a nilpotent element of R then $\varphi(x)$ is nilpotent in S.

33. Assume R is commutative. Let $p(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$ be an element of the polynomial ring $R[x]$.

(a) Prove that $p(x)$ is a unit in $R[x]$ if and only if a_0 is a unit and $a_1, a_2, \ldots, a_n$ are nilpotent in R. [See Exercise 14 of Section 1.]

(b) Prove that $p(x)$ is nilpotent in $R[x]$ if and only if $a_0, a_1, \ldots, a_n$ are nilpotent elements of R.

34. Let I and J be ideals of R.

(a) Prove that $I + J$ is the smallest ideal of R containing both I and J.

(b) Prove that IJ is an ideal contained in $I \cap J$.

(c) Give an example where $IJ \neq I \cap J$.

(d) Prove that if R is commutative and if $I + J = R$ then $IJ = I \cap J$.

35. Let I, J, K be ideals of R.

(a) Prove that $I(J+K) = IJ + IK$ and $(I+J)K = IK + JK$.

(b) Prove that if $J \subseteq I$ then $I \cap (J+K) = J + (I \cap K)$.

36. Show that if I is the ideal of all polynomials in $\mathbb{Z}[x]$ with zero constant term then $I^n = \{a_n x^n + a_{n+1}x^{n+1} + \cdots + a_{n+m}x^{n+m} \mid a_i \in \mathbb{Z},\ m \geq 0\}$ is the set of polynomials whose first nonzero term has degree at least n.

37. An ideal N is called *nilpotent* if N^n is the zero ideal for some $n \geq 1$. Prove that the ideal $p\mathbb{Z}/p^m\mathbb{Z}$ is a nilpotent ideal in the ring $\mathbb{Z}/p^m\mathbb{Z}$.

7.4 PROPERTIES OF IDEALS

Throughout this section R is a ring with identity $1 \neq 0$.

Definition. Let A be any subset of the ring R.

(1) Let (A) denote the smallest ideal of R containing A, called *the ideal generated by* A.

(2) Let RA denote the set of all finite sums of elements of the form ra with $r \in R$ and $a \in A$ i.e., $RA = \{r_1a_1 + r_2a_2 + \cdots + r_na_n \mid r_i \in R,\ a_i \in A,\ n \in \mathbb{Z}^+\}$ (where the convention is $RA = 0$ if $A = \emptyset$).
Similarly, $AR = \{a_1r_1 + a_2r_2 + \cdots + a_nr_n \mid r_i \in R,\ a_i \in A,\ n \in \mathbb{Z}^+\}$ and $RAR = \{r_1a_1r_1' + r_2a_2r_2' + \cdots + r_na_nr_n' \mid r_i, r_i' \in R,\ a_i \in A,\ n \in \mathbb{Z}^+\}$.

(3) An ideal generated by a single element is called a *principal ideal.*

(4) An ideal generated by a finite set is called a *finitely generated ideal.*

When $A = \{a\}$ or $\{a_1, a_2, \ldots\}$, etc., we shall drop the set brackets and simply write (a), $(a_1, a_2, \ldots)$ for (A), respectively.

The notion of ideals generated by subsets of a ring is analogous to that of subgroups generated by subsets of a group (Section 2.4). Since the intersection of any nonempty collection of ideals of R is also an ideal (cf. Exercise 18, Section 3) and A is always contained in at least one ideal (namely R), we have

$$(A) = \bigcap_{\substack{I \text{ an ideal} \\ A \subseteq I}} I\ ,$$

i.e., (A) is the intersection of all ideals of R that contain the set A.

The *left ideal generated by* A is the intersection of all left ideals of R that contain A. This left ideal is obtained from A by closing A under all the operations that define a left ideal. It is immediate from the definition that RA is closed under addition and under left multiplication by any ring element. Since R has an identity, RA contains A. Thus RA is a left ideal of R which contains A. Conversely, any left ideal which contains A must contain all finite sums of elements of the form ra, $r \in R$ and $a \in A$ and so must contain RA. Thus *RA is precisely the left ideal generated by A.* Similarly, *AR is the right ideal generated by A* and *RAR is the (two-sided) ideal generated by A.* In particular,

$$\textit{if } R \textit{ is commutative then } RA = AR = RAR = (A).$$

When R is a commutative ring and $a \in R$, the principal ideal (a) generated by a is just the set of all R-multiples of a. If R is not commutative, however, the set

$\{ras \mid r, s \in R\}$ is not necessarily the two-sided ideal generated by a since it need not be closed under addition (in this case the ideal generated by a is the ideal RaR, which consists of all *finite sums* of elements of the form ras, $r, s \in R$).

The formation of principal ideals in a commutative ring is a particularly simple way of creating ideals, similar to generating cyclic subgroups of a group. Notice that the element $b \in R$ belongs to the ideal (a) if and only if $b = ra$ for some $r \in R$, i.e., if and only if *b is a multiple of a* or, put another way, *a divides b in R*. Also, $b \in (a)$ if and only if $(b) \subseteq (a)$. Thus containment relations between ideals, in particular between principal ideals, is seen to capture some of the arithmetic of general commutative rings. Commutative rings in which all ideals are principal are among the easiest to study and these will play an important role in Chapters 8 and 9.

Examples

(1) The trivial ideal 0 and the ideal R are both principal: $0 = (0)$ and $R = (1)$.

(2) In $\mathbb{Z}$ we have $n\mathbb{Z} = \mathbb{Z}n = (n) = (-n)$ for all integers n. Thus our notation for aR is consistent with the definition of $n\mathbb{Z}$ we have been using. As noted in the preceding section, these are all the ideals of $\mathbb{Z}$ so *every* ideal of $\mathbb{Z}$ is principal. For positive integers n and m, $n\mathbb{Z} \subseteq m\mathbb{Z}$ if and only if m divides n in $\mathbb{Z}$, so the lattice of ideals containing $n\mathbb{Z}$ is the same as the lattice of divisors of n. Furthermore, the ideal generated by two nonzero integers n and m is the principal ideal generated by their greatest common divisor, d: $(n, m) = (d)$. The notation for (n, m) as the greatest common divisor of n and m is thus consistent with the same notation for the ideal generated by n and m (although a principal generator for the ideal generated by n and m is determined only up to a $\pm$ sign — we could make it unique by choosing a nonnegative generator). In particular, n and m are relatively prime if and only if $(n, m) = (1)$.

(3) We show that the ideal $(2, x)$ generated by 2 and x in $\mathbb{Z}[x]$ is *not* a principal ideal. Observe that $(2, x) = \{2p(x) + xq(x) \mid p(x), q(x) \in \mathbb{Z}[x]\}$ and so this ideal consists precisely of the polynomials with integer coefficients whose constant term is even (as discussed in Example 5 in the preceding section) — in particular, this is a proper ideal. Assume by way of contradiction that $(2, x) = (a(x))$ for some $a(x) \in \mathbb{Z}[x]$. Since $2 \in (a(x))$ there must be some $p(x)$ such that $2 = p(x)a(x)$. The degree of $p(x)a(x)$ equals degree $p(x)$ + degree $a(x)$, hence both $p(x)$ and $a(x)$ must be constant polynomials, i.e., integers. Since 2 is a prime number, $a(x)$, $p(x) \in \{\pm 1, \pm 2\}$. If $a(x)$ were ± 1 then every polynomial would be a multiple of $a(x)$, contrary to $(a(x))$ being a proper ideal. The only possibility is $a(x) = \pm 2$. But now $x \in (a(x)) = (2) = (-2)$ and so $x = 2q(x)$ for some polynomial $q(x)$ with integer coefficients, clearly impossible. This contradiction proves that $(2, x)$ is not principal.

Note that the symbol (A) is ambiguous if the ring is not specified: the ideal generated by 2 and x in $\mathbb{Q}[x]$ is the entire ring (1) since it contains the element $\frac{1}{2}2 = 1$.

We shall see in Chapter 9 that for any *field* F, all ideals of $F[x]$ *are* principal.

(4) If R is the ring of all functions from the closed interval $[0,1]$ into $\mathbb{R}$ let M be the ideal $\{f \mid f(\frac{1}{2}) = 0\}$ (the kernel of evaluation at $\frac{1}{2}$). Let $g(x)$ be the function which is zero at $x = \frac{1}{2}$ and 1 at all other points. Then $f = fg$ for all $f \in M$ so M is a principal ideal with generator g. In fact, *any* function which is zero at $\frac{1}{2}$ and nonzero at all other points is another generator for the same ideal M.

On the other hand, if R is the ring of all *continuous* functions from $[0,1]$ to $\mathbb{R}$ then $\{f \mid f(\frac{1}{2}) = 0\}$ is *not* principal nor is it even finitely generated (cf. the exercises).

(5) If G is a finite group and R is a commutative ring with 1 then the augmentation ideal is generated by the set $\{g - 1 \mid g \in G\}$, although this need not be a minimal set of generators. For example, if G is a cyclic group with generator σ, then the augmentation ideal is a principal ideal with generator $\sigma - 1$.

Proposition 9. Let I be an ideal of R.

(1) $I = R$ if and only if I contains a unit.

(2) Assume R is commutative. Then R is a field if and only if its only ideals are 0 and R.

Proof: (1) If $I = R$ then I contains the unit 1. Conversely, if u is a unit in I with inverse v, then for any $r \in R$

$$r = r \cdot 1 = r(vu) = (rv)u \in I$$

hence $R = I$.

(2) The ring R is a field if and only if every nonzero element is a unit. If R is a field every nonzero ideal contains a unit, so by the first part R is the only nonzero ideal. Conversely, if 0 and R are the only ideals of R let u be any nonzero element of R. By hypothesis $(u) = R$ and so $1 \in (u)$. Thus there is some $v \in R$ such that $1 = vu$, i.e., u is a unit. Every nonzero element of R is therefore a unit and so R is a field.

Corollary 10. If R is a field then any nonzero ring homomorphism from R into another ring is an injection.

Proof: The kernel of a ring homomorphism is an ideal. The kernel of a nonzero homomorphism is a proper ideal hence is 0 by the proposition.

These results show that the ideal structure of fields is trivial. Our approach to studying an algebraic structure through its homomorphisms will still play a fundamental role in field theory (Part IV) when we study injective homomorphisms (embeddings) of one field into another and automorphisms of fields (isomorphisms of a field to itself).

If D is a ring with identity $1 \neq 0$ in which the only left ideals and the only right ideals are 0 and D, then D is a division ring. Conversely, the only (left, right or two-sided) ideals in a division ring D are 0 and D, which gives an analogue of Proposition 9(2) if R is not commutative (see the exercises). However, if F is a field, then for any $n \geq 2$ the only two-sided ideals in the matrix ring $M_n(F)$ are 0 and $M_n(F)$, even though this is not a division ring (it does have proper, nontrivial, left and right ideals: cf. Section 3), which shows that Proposition 9(2) does not hold for noncommutative rings. Rings whose only two-sided ideals are 0 and the whole ring (which are called *simple rings*) will be studied in Chapter 18.

One important class of ideals are those which are not contained in any other proper ideal:

Definition. An ideal M in an arbitrary ring S is called a *maximal ideal* if $M \neq S$ and the only ideals containing M are M and S.

A general ring need not have maximal ideals. For example, take any abelian group which has no maximal subgroups (for example, $\mathbb{Q}$ — cf. Exercise 16, Section 6.1) and make it into a trivial ring by defining $ab = 0$ for all a, b. In such a ring the ideals are simply the subgroups and so there are no maximal ideals. The zero ring has no maximal ideals, hence any result involving maximal ideals forces a ring to be nonzero. The next proposition shows that rings with an identity $1 \neq 0$ always possess maximal ideals. Like many such general existence theorems (e.g., the result that a finitely generated group has maximal subgroups or that every vector space has a basis) the proof relies on Zorn's Lemma (see Appendix I). In many specific rings, however, the presence of maximal ideals is often obvious, independent of Zorn's Lemma.

Proposition 11. In a ring with identity every proper ideal is contained in a maximal ideal.

Proof: Let R be a ring with identity and let I be a proper ideal (so R cannot be the zero ring, i.e., $1 \neq 0$). Let $\mathcal{S}$ be the set of all proper ideals of R which contain I. Then $\mathcal{S}$ is nonempty ($I \in \mathcal{S}$) and is partially ordered by inclusion. If $\mathcal{C}$ is a chain in $\mathcal{S}$, define J to be the union of all ideals in $\mathcal{C}$:

$$J = \bigcup_{A \in \mathcal{C}} A.$$

We first show that J is an ideal. Certainly J is nonempty because $\mathcal{C}$ is nonempty — specifically, $0 \in J$ since 0 is in every ideal A. If $a, b \in J$, then there are ideals $A, B \in \mathcal{C}$ such that $a \in A$ and $b \in B$. By definition of a chain either $A \subseteq B$ or $B \subseteq A$. In either case $a - b \in J$, so J is closed under subtraction. Since each $A \in \mathcal{C}$ is closed under left and right multiplication by elements of R, so is J. This proves J is an ideal.

If J is not a proper ideal then $1 \in J$. In this case, by definition of J we must have $1 \in A$ for some $A \in \mathcal{C}$. This is a contradiction because each A is a proper ideal ($A \in \mathcal{C} \subseteq \mathcal{S}$). This proves that each chain has an upper bound in $\mathcal{S}$. By Zorn's Lemma $\mathcal{S}$ has a maximal element which is therefore a maximal (proper) ideal containing I.

For commutative rings the next result characterizes maximal ideals by the structure of their quotient rings.

Proposition 12. Assume R is commutative. The ideal M is a maximal ideal if and only if the quotient ring R/M is a field.

Proof: This follows from the Lattice Isomorphism Theorem together with Proposition 9(2). The ideal M is maximal if and only if there are no ideals I with $M \subset I \subset R$. By the Lattice Isomorphism Theorem the ideals of R containing M correspond bijectively with the ideals of R/M, so M is maximal if and only if the only ideals of R/M are 0 and R/M. By Proposition 9(2) we see that M is maximal if and only if R/M is a field.

The proposition above indicates how to *construct* some fields: take the quotient of any commutative ring R with identity by a maximal ideal in R. We shall use this in Part IV to construct all finite fields by taking quotients of the ring $\mathbb{Z}[x]$ by maximal ideals.

Examples

(1) Let n be a nonnegative integer. The ideal $n\mathbb{Z}$ of $\mathbb{Z}$ is a maximal ideal if and only if $\mathbb{Z}/n\mathbb{Z}$ is a field. We saw in Section 3 that this is the case if and only if n is a prime number. This also follows directly from the containment of ideals of $\mathbb{Z}$ described in Example 2 above.

(2) The ideal $(2, x)$ is a maximal ideal in $\mathbb{Z}[x]$ because its quotient ring is the field $\mathbb{Z}/2\mathbb{Z}$ — cf. Example 3 above and Example 5 at the end of Section 3.

(3) The ideal (x) in $\mathbb{Z}[x]$ is not a maximal ideal because $(x) \subset (2, x) \subset \mathbb{Z}[x]$. The quotient ring $\mathbb{Z}[x]/(x)$ is isomorphic to $\mathbb{Z}$ (the ideal (x) in $\mathbb{Z}[x]$ is the kernel of the surjective ring homomorphism from $\mathbb{Z}[x]$ to $\mathbb{Z}$ given by evaluation at 0). Since $\mathbb{Z}$ is not a field, we see again that (x) is not a maximal ideal in $\mathbb{Z}[x]$.

(4) Let R be the ring of all functions from [0,1] to $\mathbb{R}$ and for each $a \in [0, 1]$ let M_a be the kernel of evaluation at a. Since evaluation is a surjective homomorphism from R to $\mathbb{R}$, we see that $R/M_a \cong \mathbb{R}$ and hence M_a is a maximal ideal. Similarly, the kernel of evaluation at any fixed point is a maximal ideal in the ring of continuous real valued functions on [0, 1].

(5) If F is a field and G is a finite group, then the augmentation ideal I is a maximal ideal of the group ring FG (cf. Example 7 at the end of the preceding section). The augmentation ideal is the kernel of the augmentation map which is a surjective homomorphism onto the field F (i.e., $FG/I \cong F$, a field). Note that Proposition 12 does not apply directly since FG need not be commutative, however, the implication in Proposition 12 that I is a maximal ideal if R/I is a field holds for arbitrary rings.

Definition. Assume R is commutative. An ideal P is called a *prime ideal* if $P \neq R$ and whenever the product ab of two elements $a, b \in R$ is an element of P, then at least one of a and b is an element of P.

The notion of a maximal ideal is fairly intuitive but the definition of a prime ideal may seem a little strange. It is, however, a natural generalization of the notion of a "prime" in the integers $\mathbb{Z}$. Let n be a nonnegative integer. According to the above definition the ideal $n\mathbb{Z}$ is a *prime* ideal provided $n \neq 1$ (to ensure that the ideal is proper) and provided every time the product ab of two integers is an element of $n\mathbb{Z}$, at least one of a, b is an element of $n\mathbb{Z}$. Put another way, if $n \neq 0$, it must have the property that whenever n divides ab, n must divide a or divide b. This is equivalent to the usual definition that n is a prime number. Thus *the prime ideals of $\mathbb{Z}$ are just the ideals $p\mathbb{Z}$ of $\mathbb{Z}$ generated by prime numbers p together with the ideal 0.*

For the integers $\mathbb{Z}$ there is no difference between the maximal ideals and the nonzero prime ideals. This is not true in general, but we shall see shortly that every maximal ideal is a prime ideal. First we translate the notion of prime ideals into properties of quotient rings as we did for maximal ideals in Proposition 12. Recall that an integral domain is a commutative ring with identity $1 \neq 0$ that has no zero divisors.

Proposition 13. Assume R is commutative. Then the ideal P is a prime ideal in R if and only if the quotient ring R/P is an integral domain.

Proof: This proof is simply a matter of translating the definition of a prime ideal into the language of quotients. The ideal P is prime if and only if $P \neq R$ and whenever

$ab \in P$, then either $a \in P$ or $b \in P$. Use the bar notation for elements of R/P: $\bar{r} = r + P$. Note that $r \in P$ if and only if the element $\bar{r}$ is zero in the quotient ring R/P. Thus in the terminology of quotients P is a prime ideal if and only if $\bar{R} \neq \bar{0}$ and whenever $\overline{ab} = \bar{a}\bar{b} = \bar{0}$, then either $\bar{a} = \bar{0}$ or $\bar{b} = \bar{0}$, i.e., R/P is an integral domain.

It follows in particular that a commutative ring with identity is an integral domain if and only if 0 is a prime ideal.

Corollary 14. Assume R is commutative. Every maximal ideal of R is a prime ideal.

Proof: If M is a maximal ideal then R/M is a field by Proposition 12. A field is an integral domain so the corollary follows from Proposition 13.

Examples

(1) The principal ideals generated by primes in $\mathbb{Z}$ are both prime and maximal ideals. The zero ideal in $\mathbb{Z}$ is prime but not maximal.

(2) The ideal (x) is a prime ideal in $\mathbb{Z}[x]$ since $\mathbb{Z}[x]/(x) \cong \mathbb{Z}$. This ideal is not a maximal ideal. The ideal 0 is a prime ideal in $\mathbb{Z}[x]$, but is not a maximal ideal.

EXERCISES

Let R be a ring with identity $1 \neq 0$.

1. Let L_j be the left ideal of $M_n(R)$ consisting of arbitrary entries in the j^{th} column and zero in all other entries and let E_{ij} be the element of $M_n(R)$ whose i, j entry is 1 and whose other entries are all 0. Prove that $L_j = M_n(R)E_{ij}$ for any i. [See Exercise 6, Section 2.]

2. Assume R is commutative. Prove that the augmentation ideal in the group ring RG is generated by $\{g - 1 \mid g \in G\}$. Prove that if $G = \langle \sigma \rangle$ is cyclic then the augmentation ideal is generated by $\sigma - 1$.

3. **(a)** Let p be a prime and let G be an abelian group of order p^n. Prove that the nilradical of the group ring $\mathbb{F}_p G$ is the augmentation ideal (cf. Exercise 29, Section 3). [Use the preceding exercise.]

(b) Let $G = \{g_1, \ldots, g_n\}$ be a finite group and assume R is commutative. Prove that if r is any element of the augmentation ideal of RG then $r(g_1 + \cdots + g_n) = 0$. [Use the preceding exercise.]

4. Assume R is commutative. Prove that R is a field if and only if 0 is a maximal ideal.

5. Prove that if M is an ideal such that R/M is a field then M is a maximal ideal (do not assume R is commutative).

6. Prove that R is a division ring if and only if its only left ideals are (0) and R. (The analogous result holds when "left" is replaced by "right.")

7. Let R be a commutative ring with 1. Prove that the principal ideal generated by x in the polynomial ring $R[x]$ is a prime ideal if and only if R is an integral domain. Prove that (x) is a maximal ideal if and only if R is a field.

8. Let R be an integral domain. Prove that $(a) = (b)$ for some elements $a, b \in R$, if and only if $a = ub$ for some unit u of R.

9. Let R be the ring of all continuous functions on $[0, 1]$ and let I be the collection of functions $f(x)$ in R with $f(1/3) = f(1/2) = 0$. Prove that I is an ideal of R but is not a prime ideal.

10. Assume R is commutative. Prove that if P is a prime ideal of R and P contains no zero divisors then R is an integral domain.

11. Assume R is commutative. Let I and J be ideals of R and assume P is a prime ideal of R that contains IJ (for example, if P contains $I \cap J$). Prove either I or J is contained in P.

12. Assume R is commutative and suppose $I = (a_1, a_2, \ldots, a_n)$ and $J = (b_1, b_2, \ldots, b_m)$ are two finitely generated ideals in R. Prove that the product ideal IJ is finitely generated by the elements $a_i b_j$ for $i = 1, 2, \ldots, n$, and $j = 1, 2, \ldots, m$.

13. Let $\varphi : R \to S$ be a homomorphism of commutative rings.

(a) Prove that if P is a prime ideal of S then either $\varphi^{-1}(P) = R$ or $\varphi^{-1}(P)$ is a prime ideal of R. Apply this to the special case when R is a subring of S and φ is the inclusion homomorphism to deduce that if P is a prime ideal of S then $P \cap R$ is either R or a prime ideal of R.

(b) Prove that if M is a maximal ideal of S and φ is surjective then $\varphi^{-1}(M)$ is a maximal ideal of R. Give an example to show that this need not be the case if φ is not surjective.

14. Assume R is commutative. Let x be an indeterminate, let $f(x)$ be a monic polynomial in $R[x]$ of degree $n \geq 1$ and use the bar notation to denote passage to the quotient ring $R[x]/(f(x))$.

(a) Show that every element of $R[x]/(f(x))$ is of the form $\overline{p(x)}$ for some polynomial $p(x) \in R[x]$ of degree less than n, i.e.,

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

[If $f(x) = x^n + b_{n-1}x^{n-1} + \cdots + b_0$ then $\overline{x^n} = \overline{-(b_{n-1}x^{n-1} + \cdots + b_0)}$. Use this to reduce powers of $\overline{x}$ in the quotient ring.]

(b) Prove that if $p(x)$ and $q(x)$ are distinct polynomials in $R[x]$ which are both of degree less than n, then $\overline{p(x)} \neq \overline{q(x)}$. [Otherwise $p(x) - q(x)$ is an $R[x]$-multiple of the monic polynomial $f(x)$.]

(c) If $f(x) = a(x)b(x)$ where both $a(x)$ and $b(x)$ have degree less than n, prove that $\overline{a(x)}$ is a zero divisor in $R[x]/(f(x))$.

(d) If $f(x) = x^n - a$ for some nilpotent element $a \in R$, prove that $\overline{x}$ is nilpotent in $R[x]/(f(x))$.

(e) Let p be a prime, assume $R = \mathbb{F}_p$ and $f(x) = x^p - a$ for some $a \in \mathbb{F}_p$. Prove that $\overline{x - a}$ is nilpotent in $R[x]/(f(x))$. [Use Exercise 26(c) of Section 3.]

15. Let $x^2 + x + 1$ be an element of the polynomial ring $E = \mathbb{F}_2[x]$ and use the bar notation to denote passage to the quotient ring $\mathbb{F}_2[x]/(x^2 + x + 1)$.

(a) Prove that $\overline{E}$ has 4 elements: $\overline{0}$, $\overline{1}$, $\overline{x}$ and $\overline{x+1}$.

(b) Write out the 4×4 addition table for $\overline{E}$ and deduce that the additive group $\overline{E}$ is isomorphic to the Klein 4-group.

(c) Write out the 4×4 multiplication table for $\overline{E}$ and prove that $\overline{E}^\times$ is isomorphic to the cyclic group of order 3. Deduce that $\overline{E}$ is a field.

16. Let $x^4 - 16$ be an element of the polynomial ring $E = \mathbb{Z}[x]$ and use the bar notation to denote passage to the quotient ring $\mathbb{Z}[x]/(x^4 - 16)$.

(a) Find a polynomial of degree ≤ 3 that is congruent to $7x^{13} - 11x^9 + 5x^5 - 2x^3 + 3$ modulo $(x^4 - 16)$.

(b) Prove that $\overline{x-2}$ and $\overline{x+2}$ are zero divisors in $\overline{E}$.

17. Let $x^3 - 2x + 1$ be an element of the polynomial ring $E = \mathbb{Z}[x]$ and use the bar notation to denote passage to the quotient ring $\mathbb{Z}[x]/(x^3 - 2x + 1)$. Let $p(x) = 2x^7 - 7x^5 + 4x^3 - 9x + 1$ and let $q(x) = (x-1)^4$.

(a) Express each of the following elements of $\overline{E}$ in the form $\overline{f(x)}$ for some polynomial $f(x)$ of degree ≤ 2: $\overline{p(x)}$, $\overline{q(x)}$, $\overline{p(x)+q(x)}$ and $\overline{p(x)q(x)}$.
(b) Prove that $\overline{E}$ is not an integral domain.
(c) Prove that $\bar{x}$ is a unit in $\overline{E}$.

18. Prove that if R is an integral domain and $R[[x]]$ is the ring of formal power series in the indeterminate x then the principal ideal generated by x is a prime ideal (cf. Exercise 3, Section 2). Prove that the principal ideal generated by x is a maximal ideal if and only if R is a field.

19. Let R be a finite commutative ring with identity. Prove that every prime ideal of R is a maximal ideal.

20. Prove that a nonzero finite commutative ring that has no zero divisors is a field (if the ring has an identity, this is Corollary 3, so do not assume the ring has a 1).

21. Prove that a finite ring with identity $1 \neq 0$ that has no zero divisors is a field (you may quote Wedderburn's Theorem).

22. Let $p \in \mathbb{Z}^+$ be a prime and let the $\mathbb{F}_p$ Quaternions be defined by

$$a + bi + cj + dk \qquad a, b, c, d \in \mathbb{Z}/p\mathbb{Z}$$

where addition is componentwise and multiplication is defined using the same relations on i, j, k as for the real Quaternions.
(a) Prove that the $\mathbb{F}_p$ Quaternions are a homomorphic image of the integral Quaternions (cf. Section 1).
(b) Prove that the $\mathbb{F}_p$ Quaternions contain zero divisors (and so they cannot be a division ring). [Use the preceding exercise.]

23. Prove that in a Boolean ring (cf. Exercise 15, Section 1) every prime ideal is a maximal ideal.

24. Prove that in a Boolean ring every finitely generated ideal is principal.

25. Assume R is commutative and for each $a \in R$ there is an integer $n > 1$ (depending on a) such that $a^n = a$. Prove that every prime ideal of R is a maximal ideal.

26. Prove that a prime ideal in a commutative ring R contains every nilpotent element (cf. Exercise 13, Section 1). Deduce that the nilradical of R (cf. Exercise 29, Section 3) is contained in the intersection of all the prime ideals of R. (It is shown in Section 15.2 that the nilradical of R is equal to the intersection of all prime ideals of R.)

27. Let R be a commutative ring with $1 \neq 0$. Prove that if a is a nilpotent element of R then $1 - ab$ is a unit for all $b \in R$.

28. Prove that if R is a commutative ring and $N = (a_1, a_2, \ldots, a_m)$ where each a_i is a nilpotent element, then N is a nilpotent ideal (cf. Exercise 37, Section 3). Deduce that if the nilradical of R is finitely generated then it is a nilpotent ideal.

29. Let p be a prime and let G be a finite group of order a power of p (i.e., a p-group). Prove that the augmentation ideal in the group ring $\mathbb{Z}/p\mathbb{Z}G$ is a nilpotent ideal. (Note that this ring may be noncommutative.) [Use Exercise 2.]

30. Let I be an ideal of the commutative ring R and define

$$\text{rad } I = \{r \in R \mid r^n \in I \text{ for some } n \in \mathbb{Z}^+\}$$

called the *radical* of I. Prove that rad I is an ideal containing I and that $(\text{rad } I)/I$ is the nilradical of the quotient ring R/I, i.e., $(\text{rad } I)/I = \mathfrak{N}(R/I)$ (cf. Exercise 29, Section 3).

31. An ideal I of the commutative ring R is called a *radical ideal* if rad $I = I$.

(a) Prove that every prime ideal of R is a radical ideal.

(b) Let $n > 1$ be an integer. Prove that 0 is a radical ideal in $\mathbb{Z}/n\mathbb{Z}$ if and only if n is a product of distinct primes to the first power (i.e., n is square free). Deduce that (n) is a radical ideal of $\mathbb{Z}$ if and only if n is a product of distinct primes in $\mathbb{Z}$.

32. Let I be an ideal of the commutative ring R and define

$$\text{Jac } I \text{ to be the intersection of all maximal ideals of } R \text{ that contain } I$$

where the convention is that Jac $R = R$. (If I is the zero ideal, Jac 0 is called the *Jacobson radical* of the ring R, so Jac I is the preimage in R of the Jacobson radical of R/I.)

(a) Prove that Jac I is an ideal of R containing I.

(b) Prove that rad $I \subseteq$ Jac I, where rad I is the radical of I defined in Exercise 30.

(c) Let $n > 1$ be an integer. Describe Jac $n\mathbb{Z}$ in terms of the prime factorization of n.

33. Let R be the ring of all continuous functions from the closed interval [0,1] to $\mathbb{R}$ and for each $c \in [0, 1]$ let $M_c = \{f \in R \mid f(c) = 0\}$ (recall that M_c was shown to be a maximal ideal of R).

(a) Prove that if M is *any* maximal ideal of R then there is a real number $c \in [0, 1]$ such that $M = M_c$.

(b) Prove that if b and c are distinct points in [0,1] then $M_b \neq M_c$.

(c) Prove that M_c is not equal to the principal ideal generated by $x - c$.

(d) Prove that M_c is not a finitely generated ideal.

The preceding exercise shows that there is a bijection between the *points* of the closed interval [0,1] and the set of *maximal ideals* in the ring R of all of continuous functions on [0,1] given by $c \leftrightarrow M_c$. For any subset X of $\mathbb{R}$ or, more generally, for any completely regular topological space X, the map $c \mapsto M_c$ is an *injection* from X to the set of maximal ideals of R, where R is the ring of all bounded continuous real valued functions on X and M_c is the maximal ideal of functions that vanish at c. Let $\beta(X)$ be the set of maximal ideals of R. One can put a topology on $\beta(X)$ in such a way that if we identify X with its image in $\beta(X)$ then X (in its given topology) becomes a subspace of $\beta(X)$. Moreover, $\beta(X)$ is a compact space under this topology and is called the *Stone-Čech compactification* of X.

34. Let R be the ring of all continuous functions from $\mathbb{R}$ to $\mathbb{R}$ and for each $c \in \mathbb{R}$ let M_c be the maximal ideal $\{f \in R \mid f(c) = 0\}$.

(a) Let I be the collection of functions $f(x)$ in R with *compact support* (i.e., $f(x) = 0$ for $|x|$ sufficiently large). Prove that I is an ideal of R that is not a prime ideal.

(b) Let M be a maximal ideal of R containing I (properly, by (a)). Prove that $M \neq M_c$ for any $c \in \mathbb{R}$ (cf. the preceding exercise).

35. Let $A = (a_1, a_2, \ldots, a_n)$ be a nonzero finitely generated ideal of R. Prove that there is an ideal B which is maximal with respect to the property that it does not contain A. [Use Zorn's Lemma.]

36. Assume R is commutative. Prove that the set of prime ideals in R has a minimal element with respect to inclusion (possibly the zero ideal). [Use Zorn's Lemma.]

37. A commutative ring R is called a *local ring* if it has a unique maximal ideal. Prove that if R is a local ring with maximal ideal M then every element of $R - M$ is a unit. Prove conversely that if R is a commutative ring with 1 in which the set of nonunits forms an ideal M, then R is a local ring with unique maximal ideal M.

38. Prove that the ring of all rational numbers whose denominators is odd is a local ring whose unique maximal ideal is the principal ideal generated by 2.

39. Following the notation of Exercise 26 in Section 1, let K be a field, let v be a discrete

valuation on K and let R be the valuation ring of v. For each integer $k \geq 0$ define $A_k = \{r \in R \mid v(r) \geq k\} \cup \{0\}$.

(a) Prove that A_k is a principal ideal and that $A_0 \supseteq A_1 \supseteq A_2 \supseteq \cdots$.

(b) Prove that if I is any nonzero ideal of R, then $I = A_k$ for some $k \geq 0$. Deduce that R is a local ring with unique maximal ideal A_1.

40. Assume R is commutative. Prove that the following are equivalent: (see also Exercises 13 and 14 in Section 1)

(i) R has exactly one prime ideal

(ii) every element of R is either nilpotent or a unit

(iii) $R/\eta(R)$ is a field (cf. Exercise 29, Section 3).

41. A proper ideal Q of the commutative ring R is called *primary* if whenever $ab \in Q$ and $a \notin Q$ then $b^n \in Q$ for some positive integer n. (Note that the symmetry between a and b in this definition implies that if Q is a primary ideal and $ab \in Q$ with *neither* a nor b in Q, then a positive power of a and a positive power of b both lie in Q.) Establish the following facts about primary ideals.

(a) The primary ideals of $\mathbb{Z}$ are 0 and (p^n), where p is a prime and n is a positive integer.

(b) Every prime ideal of R is a primary ideal.

(c) An ideal Q of R is primary if and only if every zero divisor in R/Q is a nilpotent element of R/Q.

(d) If Q is a primary ideal then $\text{rad}(Q)$ is a prime ideal (cf. Exercise 30).

7.5 RINGS OF FRACTIONS

Throughout this section R is a commutative ring. Proposition 2 shows that if a is not zero nor a zero divisor and $ab = ac$ in R then $b = c$. Thus a nonzero element that is not a zero divisor enjoys some of the properties of a unit without necessarily possessing a multiplicative inverse in R. On the other hand, we saw in Section 1 that a zero divisor a cannot be a unit in R and, by definition, if a is a zero divisor we cannot always cancel the a's in the equation $ab = ac$ to obtain $b = c$ (take $c = 0$ for example). The aim of this section is to prove that a commutative ring R is always a subring of a larger ring Q in which every nonzero element of R that is not a zero divisor is a unit in Q. The principal application of this will be to integral domains, in which case this ring Q will be a field — called its *field of fractions* or *quotient field*. Indeed, the paradigm for the construction of Q from R is the one offered by the construction of the field of rational numbers from the integral domain $\mathbb{Z}$.

In order to see the essential features of the construction of the field $\mathbb{Q}$ from the integral domain $\mathbb{Z}$ we review the basic properties of fractions. Each rational number may be represented in many different ways as the quotient of two integers (for example, $\frac{1}{2} = \frac{2}{4} = \frac{3}{6} = \ldots$, etc.). These representations are related by

$$\frac{a}{b} = \frac{c}{d} \qquad \text{if and only if} \qquad ad = bc.$$

In more precise terms, the fraction $\frac{a}{b}$ is the equivalence class of ordered pairs (a, b) of integers with $b \neq 0$ under the equivalence relation: $(a, b) \sim (c, d)$ if and only if

$ad = bc$. The arithmetic operations on fractions are given by

$$\frac{a}{b} + \frac{c}{d} = \frac{ad + bc}{bd} \qquad \text{and} \qquad \frac{a}{b} \times \frac{c}{d} = \frac{ac}{bd}.$$

These are well defined (independent of choice of representatives of the equivalence classes) and make the set of fractions into a commutative ring (in fact, a field), $\mathbb{Q}$. The integers $\mathbb{Z}$ are identified with the subring $\{\frac{a}{1} \mid a \in \mathbb{Z}\}$ of $\mathbb{Q}$ and every nonzero integer a has an inverse $\frac{1}{a}$ in $\mathbb{Q}$.

It seems reasonable to attempt to follow the same steps for any commutative ring R, allowing arbitrary denominators. If, however, b is zero or a zero divisor in R, say $bd = 0$, and if we allow b as a denominator, then we should expect to have

$$d = \frac{d}{1} = \frac{bd}{b} = \frac{0}{b} = 0$$

in the "ring of fractions" (where, for convenience, we have assumed R has a 1). Thus if we allow zero or zero divisors as denominators there must be some collapsing in the sense that we cannot expect R to appear naturally as a subring of this "ring of fractions." A second restriction is more obviously imposed by the laws of addition and multiplication: if ring elements b and d are allowed as denominators, then bd must also be a denominator, i.e., the set of denominators must be closed under multiplication in R. The main result of this section shows that these two restrictions are sufficient to construct a ring of fractions for R. Note that this theorem includes the construction of $\mathbb{Q}$ from $\mathbb{Z}$ as a special case.

Theorem 15. Let R be a commutative ring. Let D be any nonempty subset of R that does not contain 0, does not contain any zero divisors and is closed under multiplication (i.e., $ab \in D$ for all $a, b \in D$). Then there is a commutative ring Q with 1 such that Q contains R as a subring and every element of D is a unit in Q. The ring Q has the following additional properties.

(1) every element of Q is of the form rd^{-1} for some $r \in R$ and $d \in D$. In particular, if $D = R - \{0\}$ then Q is a field.

(2) (uniqueness of Q) The ring Q is the *"smallest"* ring containing R in which all elements of D become units, in the following sense. Let S be any commutative ring with identity and let $\varphi : R \to S$ be any injective ring homomorphism such that $\varphi(d)$ is a unit in S for every $d \in D$. Then there is an injective homomorphism $\Phi : Q \to S$ such that $\Phi|_R = \varphi$. In other words, any ring containing an isomorphic copy of R in which all the elements of D become units must also contain an isomorphic copy of Q.

Remark: In Section 15.4 a more general construction is given. The proof of the general result is more technical but relies on the same basic rationale and steps as the proof of Theorem 15. Readers wishing greater generality may read the proof below and the beginning of Section 15.4 in concert.

Proof: Let $\mathcal{F} = \{(r, d) \mid r \in R,\ d \in D\}$ and define the relation $\sim$ on $\mathcal{F}$ by

$$(r, d) \sim (s, e) \qquad \text{if and only if} \qquad re = sd.$$

It is immediate that this relation is reflexive and symmetric. Suppose $(r, d) \sim (s, e)$ and $(s, e) \sim (t, f)$. Then $re - sd = 0$ and $sf - te = 0$. Multiplying the first of these equations by f and the second by d and adding them gives $(rf - td)e = 0$. Since $e \in D$ is neither zero nor a zero divisor we must have $rf - td = 0$, i.e., $(r, d) \sim (t, f)$. This proves $\sim$ is transitive, hence an equivalence relation. Denote the equivalence class of (r, d) by $\frac{r}{d}$:

$$\frac{r}{d} = \{(a, b) \mid a \in R,\ b \in D \text{ and } rb = ad\}.$$

Let Q be the set of equivalence classes under $\sim$. Note that $\frac{r}{d} = \frac{re}{de}$ in Q for all $e \in D$, since D is closed under multiplication.

We now define an additive and multiplicative structure on Q:

$$\frac{a}{b} + \frac{c}{d} = \frac{ad + bc}{bd} \qquad \text{and} \qquad \frac{a}{b} \times \frac{c}{d} = \frac{ac}{bd}.$$

In order to prove that Q is a commutative ring with identity there are a number of things to check:

(1) these operations are well defined (i.e., do not depend on the choice of representatives for the equivalence classes),

(2) Q is an abelian group under addition, where the additive identity is $\frac{0}{d}$ for any $d \in D$ and the additive inverse of $\frac{a}{d}$ is $\frac{-a}{d}$,

(3) multiplication is associative, distributive and commutative, and

(4) Q has an identity ($= \frac{d}{d}$ for any $d \in D$).

These are all completely straightforward calculations involving only arithmetic in R and the definition of $\sim$. Again we need D to be closed under multiplication for addition and multiplication to be defined.

For example, to check that addition is well defined assume $\frac{a}{b} = \frac{a'}{b'}$ (i.e., $ab' = a'b$) and $\frac{c}{d} = \frac{c'}{d'}$ (i.e., $cd' = c'd$). We must show that $\frac{ad + bc}{bd} = \frac{a'd' + b'c'}{b'd'}$, i.e.,

$$(ad + bc)(b'd') = (a'd' + b'c')(bd).$$

The left hand side of this equation is $ab'dd' + cd'bb'$ substituting $a'b$ for ab' and $c'd$ for cd' gives $a'bdd' + c'dbb'$, which is the right hand side. Hence addition of fractions is well defined. Checking the details in the other parts of (1) to (4) involves even easier manipulations and so is left as an exercise.

Next we embed R into Q by defining

$$\iota : R \to Q \qquad \text{by} \qquad \iota : r \mapsto \frac{rd}{d} \quad \text{where } d \text{ is any element of } D.$$

Since $\frac{rd}{d} = \frac{re}{e}$ for all $d, e \in D$, $\iota(r)$ does not depend on the choice of $d \in D$. Since D is closed under multiplication, one checks directly that ι is a ring homomorphism.

Furthermore, ι is injective because

$$\iota(r) = 0 \Leftrightarrow \frac{rd}{d} = \frac{0}{d} \Leftrightarrow rd^2 = 0 \Leftrightarrow r = 0$$

because d (hence also d^2) is neither zero nor a zero divisor. The subring $\iota(R)$ of Q is therefore isomorphic to R. We henceforth identify each $r \in R$ with $\iota(r)$ and so consider R as a subring of Q.

Next note that each $d \in D$ has a multiplicative inverse in Q: namely, if d is represented by the fraction $\frac{de}{e}$ then its multiplicative inverse is $\frac{e}{de}$. One then sees that every element of Q may be written as $r \cdot d^{-1}$ for some $r \in R$ and some $d \in D$. In particular, if $D = R - \{0\}$, every nonzero element of Q has a multiplicative inverse and Q is a field.

It remains to establish the uniqueness property of Q. Assume $\varphi : R \to S$ is an injective ring homomorphism such that $\varphi(d)$ is a unit in S for all $d \in D$. Extend φ to a map $\Phi : Q \to S$ by defining $\Phi(rd^{-1}) = \varphi(r)\varphi(d)^{-1}$ for all $r \in R$, $d \in D$. This map is well defined, since $rd^{-1} = se^{-1}$ implies $re = sd$, so $\varphi(r)\varphi(e) = \varphi(s)\varphi(d)$, and then

$$\Phi(rd^{-1}) = \varphi(r)\varphi(d)^{-1} = \varphi(s)\varphi(e)^{-1} = \Phi(se^{-1}).$$

It is straightforward to check that Φ is a ring homomorphism — the details are left as an exercise. Finally, Φ is injective because $rd^{-1} \in \ker \Phi$ implies $r \in \ker \Phi \cap R = \ker \varphi$; since φ is injective this forces r and hence also rd^{-1} to be zero. This completes the proof.

Definition. Let R, D and Q be as in Theorem 15.

(1) The ring Q is called the *ring of fractions* of D with respect to R and is denoted $D^{-1}R$.

(2) If R is an integral domain and $D = R - \{0\}$, Q is called the *field of fractions* or *quotient field* of R.

If A is a subset of a field F (for example, if A is a subring of F), then the intersection of all the subfields of F containing A is a subfield of F and is called the subfield *generated* by A. This subfield is the smallest subfield of F containing A (namely, any subfield of F containing A contains the subfield generated by A).

The next corollary shows that the smallest field containing an integral domain R is its field of fractions.

Corollary 16. Let R be an integral domain and let Q be the field of fractions of R. If a field F contains a subring R' isomorphic to R then the subfield of F generated by R' is isomorphic to Q.

Proof: Let $\varphi : R \cong R' \subseteq F$ be a (ring) isomorphism of R to R'. In particular, $\varphi : R \to F$ is an injective homomorphism from R into the field F. Let $\Phi : Q \to F$ be the extension of φ to Q as in the theorem. By Theorem 15, Φ is injective, so $\Phi(Q)$ is an isomorphic copy of Q in F containing $\varphi(R) = R'$. Now, any subfield of F containing $R' = \varphi(R)$ contains the elements $\varphi(r_1)\varphi(r_2)^{-1} = \varphi(r_1 r_2^{-1})$ for all $r_1, r_2 \in R$. Since

every element of Q is of the form $r_1 r_2^{-1}$ for some $r_1, r_2 \in R$, it follows that any subfield of F containing R' contains the field $\Phi(Q)$, so that $\Phi(Q)$ is the subfield of F generated by R', proving the corollary.

Examples

(1) If R is a field then its field of fractions is just R itself.

(2) The integers $\mathbb{Z}$ are an integral domain whose field of fractions is the field $\mathbb{Q}$ of rational numbers. The quadratic integer ring $\mathcal{O}$ of Section 1 is an integral domain whose field of fractions is the quadratic field $\mathbb{Q}(\sqrt{D})$.

(3) The subring $2\mathbb{Z}$ of $\mathbb{Z}$ also has no zero divisors (but has no identity). Its field of fractions is also $\mathbb{Q}$. Note how an identity "appears" in the field of fractions.

(4) If R is any integral domain, then the polynomial ring $R[x]$ is also an integral domain. The associated field of fractions is the field of *rational functions* in the variable x over R. The elements of this field are of the form $\dfrac{p(x)}{q(x)}$, where $p(x)$ and $q(x)$ are polynomials with coefficients in R with $q(x)$ not the zero polynomial. In particular, $p(x)$ and $q(x)$ may both be constant polynomials, so the field of rational functions contains the field of fractions of R: elements of the form $\dfrac{a}{b}$ such that $a, b \in R$ and $b \neq 0$. If F is a field, we shall denote the field of rational functions by $F(x)$. Thus if F is the field of fractions of the integral domain R then the field of rational functions over R is the same as the field of rational functions over F, namely $F(x)$.

For example, suppose $R = \mathbb{Z}$, so $F = \mathbb{Q}$. If $p(x), q(x)$ are polynomials in $\mathbb{Q}[x]$ then for some integer N, $Np(x), Nq(x)$ have integer coefficients (let N be a common denominator for all the coefficients in $p(x)$ and $q(x)$, for example). Then $\dfrac{p(x)}{q(x)} = \dfrac{Np(x)}{Nq(x)}$ can be written as the quotient of two polynomials with integer coefficients, so the field of fractions of $\mathbb{Q}[x]$ is the same as the field of fractions of $\mathbb{Z}[x]$.

(5) If R is any commutative ring with identity and d is neither zero nor a zero divisor in R we may form the ring $R[1/d]$ by setting $D = \{1, d, d^2, d^3, \dots\}$ and defining $R[1/d]$ to be the ring of fractions $D^{-1}R$. Note that R is the subring of elements of the form $\dfrac{r}{1}$. In this way any nonzero element of R that is not a zero divisor can be inverted in a larger ring containing R. Note that the elements of $R[1/d]$ look like polynomials in $1/d$ with coefficients in R, which explains the notation.

EXERCISES

Let R be a commutative ring with identity $1 \neq 0$.

1. Fill in all the details in the proof of Theorem 15.

2. Let R be an integral domain and let D be a nonempty subset of R that is closed under multiplication. Prove that the ring of fractions $D^{-1}R$ is isomorphic to a subring of the quotient field of R (hence is also an integral domain).

3. Let F be a field. Prove that F contains a unique smallest subfield F_0 and that F_0 is isomorphic to either $\mathbb{Q}$ or $\mathbb{Z}/p\mathbb{Z}$ for some prime p (F_0 is called the *prime subfield* of F). [See Exercise 26, Section 3.]

4. Prove that any subfield of $\mathbb{R}$ must contain $\mathbb{Q}$.

5. If F is a field, prove that the field of fractions of $F[[x]]$ (the ring of formal power series in the indeterminate x with coefficients in F) is the ring $F((x))$ of formal Laurent series (cf. Exercises 3 and 5 of Section 2). Show the field of fractions of the power series ring $\mathbb{Z}[[x]]$ is *properly* contained in the field of Laurent series $\mathbb{Q}((x))$. [Consider the series for e^x.]
6. Prove that the real numbers, $\mathbb{R}$, contain a subring A with $1 \in A$ and A maximal (under inclusion) with respect to the property that $\frac{1}{2} \notin A$. [Use Zorn's Lemma.] (Exercise 13 in Section 15.3 shows $\mathbb{R}$ is the quotient field of A, so $\mathbb{R}$ is the quotient field of a proper subring.)

7.6 THE CHINESE REMAINDER THEOREM

Throughout this section we shall assume unless otherwise stated that all rings are commutative with an identity $1 \neq 0$.

Given an arbitrary collection of rings (not necessarily satisfying the conventions above), their *(ring) direct product* is defined to be their direct product as (abelian) groups made into a ring by defining multiplication componentwise. In particular, if R_1 and R_2 are two rings, we shall denote by $R_1 \times R_2$ their direct product (as rings), that is, the set of ordered pairs (r_1, r_2) with $r_1 \in R_1$ and $r_2 \in R_2$ where addition and multiplication are performed componentwise:

$$(r_1, r_2) + (s_1, s_2) = (r_1 + s_1, r_2 + s_2) \qquad \text{and} \qquad (r_1, r_2)(s_1, s_2) = (r_1 s_1, r_2 s_2).$$

We note that a map φ from a ring R into a direct product ring is a homomorphism if and only if the induced maps into each of the components are homomorphisms.

There is a generalization to arbitrary rings of the notion in $\mathbb{Z}$ of two integers n and m being relatively prime (even to rings where the notion of greatest common divisor is not defined). In $\mathbb{Z}$ this is equivalent to being able to solve the equation $nx + my = 1$ in integers x and y (this fact was stated in Chapter 0 and will be proved in Chapter 8). This in turn is equivalent to $n\mathbb{Z} + m\mathbb{Z} = \mathbb{Z}$ as ideals (in general, $n\mathbb{Z} + m\mathbb{Z} = (m, n)\mathbb{Z}$). This motivates the following definition:

Definition. The ideals A and B of the ring R are said to be *comaximal* if $A + B = R$.

Recall that the *product*, AB, of the ideals A and B of R is the ideal consisting of all finite sums of elements of the form xy, $x \in A$ and $y \in B$ (cf. Exercise 34, Section 3). If $A = (a)$ and $B = (b)$, then $AB = (ab)$. More generally, the product of the ideals $A_1, A_2, \ldots, A_k$ is the ideal of all finite sums of elements of the form $x_1 x_2 \cdots x_k$ such that $x_i \in A_i$ for all i. If $A_i = (a_i)$, then $A_1 \cdots A_k = (a_1 \cdots a_k)$.

Theorem 17. *(Chinese Remainder Theorem)* Let $A_1, A_2, \ldots, A_k$ be ideals in R. The map

$$R \to R/A_1 \times R/A_2 \times \cdots \times R/A_k \qquad \text{defined by} \qquad r \mapsto (r + A_1, r + A_2, \ldots, r + A_k)$$

is a ring homomorphism with kernel $A_1 \cap A_2 \cap \cdots \cap A_k$. If for each $i, j \in \{1, 2, \ldots, k\}$ with $i \neq j$ the ideals A_i and A_j are comaximal, then this map is surjective and $A_1 \cap A_2 \cap \cdots \cap A_k = A_1 A_2 \cdots A_k$, so

$$R/(A_1 A_2 \cdots A_k) = R/(A_1 \cap A_2 \cap \cdots \cap A_k) \cong R/A_1 \times R/A_2 \times \cdots \times R/A_k.$$

Proof: We first prove this for $k = 2$; the general case will follow by induction. Let $A = A_1$ and $B = A_2$. Consider the map $\varphi : R \to R/A \times R/B$ defined by $\varphi(r) = (r \bmod A, r \bmod B)$, where mod A means the class in R/A containing r (that is, $r + A$). This map is a ring homomorphism because φ is just the natural projection of R into R/A and R/B for the two components. The kernel of φ consists of all the elements $r \in R$ that are in A and in B, i.e., $A \cap B$. To complete the proof in this case it remains to show that when A and B are comaximal, φ is surjective and $A \cap B = AB$. Since $A + B = R$, there are elements $x \in A$ and $y \in B$ such that $x + y = 1$. This equation shows that $\varphi(x) = (0, 1)$ and $\varphi(y) = (1, 0)$ since, for example, x is an element of A and $x = 1 - y \in 1 + B$. If now $(r_1 \bmod A, r_2 \bmod B)$ is an arbitrary element in $R/A \times R/B$, then the element $r_2x + r_1y$ maps to this element since

$$\begin{aligned}\varphi(r_2x + r_1y) &= \varphi(r_2)\varphi(x) + \varphi(r_1)\varphi(y)\\ &= (r_2 \bmod A, r_2 \bmod B)(0, 1) + (r_1 \bmod A, r_1 \bmod B)(1, 0)\\ &= (0, r_2 \bmod B) + (r_1 \bmod A, 0)\\ &= (r_1 \bmod A, r_2 \bmod B).\end{aligned}$$

This shows that φ is indeed surjective. Finally, the ideal AB is always contained in $A \cap B$. If A and B are comaximal and x and y are as above, then for any $c \in A \cap B$, $c = c1 = cx + cy \in AB$. This establishes the reverse inclusion $A \cap B \subseteq AB$ and completes the proof when $k = 2$.

The general case follows easily by induction from the case of two ideals using $A = A_1$ and $B = A_2 \cdots A_k$ once we show that A_1 and $A_2 \cdots A_k$ are comaximal. By hypothesis, for each $i \in \{2, 3, \ldots, k\}$ there are elements $x_i \in A_1$ and $y_i \in A_i$ such that $x_i + y_i = 1$. Since $x_i + y_i \equiv y_i \bmod A_1$, it follows that $1 = (x_2 + y_2) \cdots (x_k + y_k)$ is an element in $A_1 + (A_2 \cdots A_k)$. This completes the proof.

This theorem obtained its name from the special case $\mathbb{Z}/mn\mathbb{Z} \cong (\mathbb{Z}/m\mathbb{Z}) \times (\mathbb{Z}/n\mathbb{Z})$ *as rings* when m and n are relatively prime integers. We proved this isomorphism just for the additive groups earlier. This isomorphism, phrased in number-theoretic terms, relates to simultaneously solving two congruences modulo relatively prime integers (and states that such congruences can always be solved, and uniquely). Such problems were considered by the ancient Chinese, hence the name. Some examples are provided in the exercises.

Since the isomorphism in the Chinese Remainder Theorem is an isomorphism of *rings*, in particular the groups of *units* on both sides must be isomorphic. It is easy to see that the units in any direct product of rings are the elements that have units in each of the coordinates. In the case of $\mathbb{Z}/mn\mathbb{Z}$ the Chinese Remainder Theorem gives the following isomorphism on the groups of units:

$$(\mathbb{Z}/mn\mathbb{Z})^\times \cong (\mathbb{Z}/m\mathbb{Z})^\times \times (\mathbb{Z}/n\mathbb{Z})^\times.$$

More generally we have the following result.

Corollary 18. Let n be a positive integer and let $p_1^{\alpha_1} p_2^{\alpha_2} \dots p_k^{\alpha_k}$ be its factorization into powers of distinct primes. Then

$$\mathbb{Z}/n\mathbb{Z} \cong (\mathbb{Z}/p_1^{\alpha_1}\mathbb{Z}) \times (\mathbb{Z}/p_2^{\alpha_2}\mathbb{Z}) \times \cdots \times (\mathbb{Z}/p_k^{\alpha_k}\mathbb{Z}),$$

as rings, so in particular we have the following isomorphism of multiplicative groups:

$$(\mathbb{Z}/n\mathbb{Z})^\times \cong (\mathbb{Z}/p_1^{\alpha_1}\mathbb{Z})^\times \times (\mathbb{Z}/p_2^{\alpha_2}\mathbb{Z})^\times \times \cdots \times (\mathbb{Z}/p_k^{\alpha_k}\mathbb{Z})^\times.$$

If we compare orders on the two sides of this last isomorphism, we obtain the formula

$$\varphi(n) = \varphi(p_1^{\alpha_1})\varphi(p_2^{\alpha_2}) \dots \varphi(p_k^{\alpha_k})$$

for the Euler φ-function. This in turn implies that φ is what in elementary number theory is termed a *multiplicative function*, namely that $\varphi(ab) = \varphi(a)\varphi(b)$ whenever a and b are relatively prime positive integers. The value of φ on prime powers p^α is easily seen to be $\varphi(p^\alpha) = p^{\alpha-1}(p-1)$ (cf. Chapter 0). From this and the multiplicativity of φ we obtain its value on all positive integers.

Corollary 18 is also a step toward a determination of the decomposition of the abelian group $(\mathbb{Z}/n\mathbb{Z})^\times$ into a direct product of cyclic groups. The complete structure is derived at the end of Section 9.5.

EXERCISES

Let R be a ring with identity $1 \neq 0$.

1. An element $e \in R$ is called an *idempotent* if $e^2 = e$. Assume e is an idempotent in R and $er = re$ for all $r \in R$. Prove that Re and $R(1-e)$ are two-sided ideals of R and that $R \cong Re \times R(1-e)$. Show that e and $1-e$ are identities for the subrings Re and $R(1-e)$ respectively.
2. Let R be a finite Boolean ring with identity $1 \neq 0$ (cf. Exercise 15 of Section 1). Prove that $R \cong \mathbb{Z}/2\mathbb{Z} \times \cdots \times \mathbb{Z}/2\mathbb{Z}$. [Use the preceding exercise.]
3. Let R and S be rings with identities. Prove that every ideal of $R \times S$ is of the form $I \times J$ where I is an ideal of R and J is an ideal of S.
4. Prove that if R and S are nonzero rings then $R \times S$ is never a field.
5. Let $n_1, n_2, \dots, n_k$ be integers which are relatively prime in pairs: $(n_i, n_j) = 1$ for all $i \neq j$.
   (a) Show that the Chinese Remainder Theorem implies that for any $a_1, \dots, a_k \in \mathbb{Z}$ there is a solution $x \in \mathbb{Z}$ to the simultaneous congruences
   $$x \equiv a_1 \bmod n_1\,, \qquad x \equiv a_2 \bmod n_2\,, \quad \dots, \quad x \equiv a_k \bmod n_k$$
   and that the solution x is unique mod $n = n_1 n_2 \dots n_k$.
   (b) Let $n_i' = n/n_i$ be the quotient of n by n_i, which is relatively prime to n_i by assumption. Let t_i be the inverse of n_i' mod n_i. Prove that the solution x in (a) is given by
   $$x = a_1 t_1 n_1' + a_2 t_2 n_2' + \cdots + a_k t_k n_k' \bmod n.$$
   Note that the elements t_i can be quickly found by the Euclidean Algorithm as described in Section 2 of the Preliminaries chapter (writing $an_i + bn_i' = (n_i, n_i') = 1$ gives $t_i = b$) and that these then quickly give the solutions to the system of congruences above for any choice of $a_1, a_2, \dots, a_k$.

(c) Solve the simultaneous system of congruences

$$x \equiv 1 \bmod 8\,, \qquad x \equiv 2 \bmod 25\,, \quad \text{and} \quad x \equiv 3 \bmod 81$$

and the simultaneous system

$$y \equiv 5 \bmod 8\,, \qquad y \equiv 12 \bmod 25\,, \quad \text{and} \quad y \equiv 47 \bmod 81.$$

6. Let $f_1(x), f_2(x), \ldots, f_k(x)$ be polynomials with integer coefficients of the same degree d. Let $n_1, n_2, \ldots, n_k$ be integers which are relatively prime in pairs (i.e., $(n_i, n_j) = 1$ for all $i \neq j$). Use the Chinese Remainder Theorem to prove there exists a polynomial $f(x)$ with integer coefficients and of degree d with

$$f(x) \equiv f_1(x) \bmod n_1\,, \qquad f(x) \equiv f_2(x) \bmod n_2\,, \quad \ldots, \quad f(x) \equiv f_k(x) \bmod n_k$$

i.e., the coefficients of $f(x)$ agree with the coefficients of $f_i(x) \bmod n_i$. Show that if all the $f_i(x)$ are monic, then $f(x)$ may also be chosen monic. [Apply the Chinese Remainder Theorem in $\mathbb{Z}$ to each of the coefficients separately.]

7. Let m and n be positive integers with n dividing m. Prove that the natural surjective ring projection $\mathbb{Z}/m\mathbb{Z} \to \mathbb{Z}/n\mathbb{Z}$ is also surjective on the units: $(\mathbb{Z}/m\mathbb{Z})^\times \to (\mathbb{Z}/n\mathbb{Z})^\times$.

The next four exercises develop the concept of *direct limits* and the "dual" notion of *inverse limits*. In these exercises I is a nonempty index set with a partial order $\leq$ (cf. Appendix I). For each $i \in I$ let A_i be an additive abelian group. In Exercise 8 assume also that I is a *directed set*: for every $i, j \in I$ there is some $k \in I$ with $i \leq k$ and $j \leq k$.

8. Suppose for every pair of indices i, j with $i \leq j$ there is a map $\rho_{ij} : A_i \to A_j$ such that the following hold:

i. $\rho_{jk} \circ \rho_{ij} = \rho_{ik}$ whenever $i \leq j \leq k$, and
ii. $\rho_{ii} = 1$ for all $i \in I$.

Let B be the disjoint union of all the A_i. Define a relation $\sim$ on B by

$$a \sim b \text{ if and only if there exists } k \text{ with } i, j \leq k \text{ and } \rho_{ik}(a) = \rho_{jk}(b),$$

for $a \in A_i$ and $b \in A_j$.

(a) Show that $\sim$ is an equivalence relation on B. (The set of equivalence classes is called the *direct* or *inductive limit* of the directed system $\{A_i\}$, and is denoted $\varinjlim A_i$. In the remaining parts of this exercise let $A = \varinjlim A_i$.)

(b) Let $\overline{x}$ denote the class of x in A and define $\rho_i : A_i \to A$ by $\rho_i(a) = \overline{a}$. Show that if each ρ_{ij} is injective, then so is ρ_i for all i (so we may then identify each A_i as a subset of A).

(c) Assume all ρ_{ij} are group homomorphisms. For $a \in A_i, b \in A_j$ show that the operation

$$\overline{a} + \overline{b} = \overline{\rho_{ik}(a) + \rho_{jk}(b)}$$

where k is any index with $i, j \leq k$, is well defined and makes A into an abelian group. Deduce that the maps ρ_i in (b) are group homomorphisms from A_i to A.

(d) Show that if all A_i are commutative rings with 1 and all ρ_{ij} are ring homomorphisms that send 1 to 1, then A may likewise be given the structure of a commutative ring with 1 such that all ρ_i are ring homomorphisms.

(e) Under the hypotheses in (c) prove that the direct limit has the following *universal property*: if C is any abelian group such that for each $i \in I$ there is a homomorphism $\varphi_i : A_i \to C$ with $\varphi_i = \varphi_j \circ \rho_{ij}$ whenever $i \leq j$, then there is a unique homomorphism $\varphi : A \to C$ such that $\varphi \circ \rho_i = \varphi_i$ for all i.

9. Let I be the collection of open intervals $U = (a, b)$ on the real line containing a fixed real number p. Order these by reverse inclusion: $U \leq V$ if $V \subseteq U$ (note that I is a directed set). For each U let A_U be the ring of continuous real valued functions on U. For $V \subseteq U$ define the *restriction maps* $\rho_{UV} : A_U \to A_V$ by $f \mapsto f|_V$, the usual restriction of a function on U to a function on the subset V (which is easily seen to be a ring homomorphism). Let $A = \varinjlim A_U$ be the direct limit. In the notation of the preceding exercise, show that the maps $\rho_U : A_U \to A$ are *not* injective but are all surjective (A is called the ring of *germs of continuous functions* at p).

We now develop the notion of *inverse limits*. Continue to assume I is a partially ordered set (but not necessarily directed), and A_i is a group for all $i \in I$.

10. Suppose for every pair of indices i, j with $i \leq j$ there is a map $\mu_{ji} : A_j \to A_i$ such that the following hold:

i. $\mu_{ji} \circ \mu_{kj} = \mu_{ki}$ whenever $i \leq j \leq k$, and
ii. $\mu_{ii} = 1$ for all $i \in I$.

Let P be the subset of elements $(a_i)_{i \in I}$ in the direct product $\prod_{i \in I} A_i$ such that $\mu_{ji}(a_j) = a_i$ whenever $i \leq j$ (here a_i and a_j are the i^{th} and j^{th} components respectively of the element in the direct product). The set P is called the *inverse* or *projective limit* of the system $\{A_i\}$, and is denoted $\varprojlim A_i$.)

(a) Assume all μ_{ji} are group homomorphisms. Show that P is a subgroup of the direct product group (cf. Exercise 15, Section 5.1).
(b) Assume the hypotheses in (a), and let $I = \mathbb{Z}^+$ (usual ordering). For each $i \in I$ let $\mu_i : P \to A_i$ be the projection of P onto its i^{th} component. Show that if each μ_{ji} is surjective, then so is μ_i for all i (so each A_i is a quotient group of P).
(c) Show that if all A_i are commutative rings with 1 and all μ_{ji} are ring homomorphisms that send 1 to 1, then A may likewise be given the structure of a commutative ring with 1 such that all μ_i are ring homomorphisms.
(d) Under the hypotheses in (a) prove that the inverse limit has the following *universal property*: if D is any group such that for each $i \in I$ there is a homomorphism $\pi_i : D \to A_i$ with $\pi_i = \mu_{ji} \circ \pi_j$ whenever $i \leq j$, then there is a unique homomorphism $\pi : D \to P$ such that $\mu_i \circ \pi = \pi_i$ for all i.

11. Let p be a prime let $I = \mathbb{Z}^+$, let $A_i = \mathbb{Z}/p^i\mathbb{Z}$ and let μ_{ji} be the natural projection maps

$$\mu_{ji} : a \pmod{p^j} \longmapsto a \pmod{p^i}.$$

The inverse limit $\varprojlim \mathbb{Z}/p^i\mathbb{Z}$ is called the ring of *p-adic integers*, and is denoted by $\mathbb{Z}_p$.

(a) Show that every element of $\mathbb{Z}_p$ may be written uniquely as an infinite formal sum $b_0 + b_1 p + b_2 p^2 + b_3 p^3 + \cdots$ with each $b_i \in \{0, 1, \ldots, p-1\}$. Describe the rules for adding and multiplying such formal sums corresponding to addition and multiplication in the ring $\mathbb{Z}_p$. [Write a least residue in each $\mathbb{Z}/p^i\mathbb{Z}$ in its base p expansion and then describe the maps μ_{ji}.] (Note in particular that $\mathbb{Z}_p$ is uncountable.)
(b) Prove that $\mathbb{Z}_p$ is an integral domain that contains a copy of the integers.
(c) Prove that $b_0 + b_1 p + b_2 p^2 + b_3 p^3 + \cdots$ as in (a) is a unit in $\mathbb{Z}_p$ if and only if $b_0 \neq 0$.
(d) Prove that $p\mathbb{Z}_p$ is the unique maximal ideal of $\mathbb{Z}_p$ and $\mathbb{Z}_p/p\mathbb{Z}_p \cong \mathbb{Z}/p\mathbb{Z}$ (where $p = 0 + 1p + 0p^2 + 0p^3 + \cdots$). Prove that every ideal of $\mathbb{Z}_p$ is of the form $p^n\mathbb{Z}_p$ for some integer $n \geq 0$.
(e) Show that if $a_1 \not\equiv 0 \pmod{p}$ then there is an element $a = (a_i)$ in the direct limit $\mathbb{Z}_p$ satisfying $a_j^p \equiv 1 \pmod{p^j}$ and $\mu_{j1}(a_j) = a_1$ for all j. Deduce that $\mathbb{Z}_p$ contains $p - 1$ distinct $(p-1)^{\text{st}}$ roots of 1.

CHAPTER 8

Euclidean Domains, Principal Ideal Domains, and Unique Factorization Domains

There are a number of classes of rings with more algebraic structure than generic rings. Those considered in this chapter are rings with a division algorithm (Euclidean Domains), rings in which every ideal is principal (Principal Ideal Domains) and rings in which elements have factorizations into primes (Unique Factorization Domains). The principal examples of such rings are the ring $\mathbb{Z}$ of integers and polynomial rings $F[x]$ with coefficients in some field F. We prove here all the theorems on the integers $\mathbb{Z}$ stated in the Preliminaries chapter as special cases of results valid for more general rings. These results will be applied to the special case of the ring $F[x]$ in the next chapter.

All rings in this chapter are commutative.

8.1 EUCLIDEAN DOMAINS

We first define the notion of a *norm* on an integral domain R. This is essentially no more than a measure of "size" in R.

Definition. Any function $N : R \to \mathbb{Z}^+ \cup \{0\}$ with $N(0) = 0$ is called a *norm* on the integral domain R. If $N(a) > 0$ for $a \neq 0$ define N to be a *positive norm.*

We observe that this notion of a norm is fairly weak and that it is possible for the same integral domain R to possess several different norms.

Definition. The integral domain R is said to be a *Euclidean Domain* (or possess a *Division Algorithm*) if there is a norm N on R such that for any two elements a and b of R with $b \neq 0$ there exist elements q and r in R with

$$a = qb + r \qquad \text{with } r = 0 \text{ or } N(r) < N(b).$$

The element q is called the *quotient* and the element r the *remainder* of the division.

The importance of the existence of a Division Algorithm on an integral domain R is that it allows a *Euclidean Algorithm* for two elements a and b of R: by successive "divisions" (these actually *are* divisions in the field of fractions of R) we can write

$$a = q_0 b + r_0 \tag{0}$$

$$b = q_1 r_0 + r_1 \tag{1}$$

$$r_0 = q_2 r_1 + r_2 \tag{2}$$

$$\vdots$$

$$r_{n-2} = q_n r_{n-1} + r_n \tag{n}$$

$$r_{n-1} = q_{n+1} r_n \tag{n+1}$$

where r_n is the last nonzero remainder. Such an r_n exists since $N(b) > N(r_0) > N(r_1) > \cdots > N(r_n)$ is a decreasing sequence of nonnegative integers if the remainders are nonzero, and such a sequence cannot continue indefinitely. Note also that there is no guarantee that these elements are *unique*.

Examples

(0) Fields are trivial examples of Euclidean Domains where any norm will satisfy the defining condition (e.g., $N(a) = 0$ for all a). This is because for every a, b with $b \neq 0$ we have $a = qb + 0$, where $q = ab^{-1}$.

(1) The integers $\mathbb{Z}$ are a Euclidean Domain with norm given by $N(a) = |a|$, the usual absolute value. The existence of a Division Algorithm in $\mathbb{Z}$ (the familiar "long division" of elementary arithmetic) is verified as follows. Let a and b be two nonzero integers and suppose first that $b > 0$. The half open intervals $[nb, (n+1)b)$, $n \in \mathbb{Z}$ partition the real line and so a is in one of them, say $a \in [kb, (k+1)b)$. For $q = k$ we have $a - qb = r \in [0, |b|)$ as needed. If $b < 0$ (so $-b > 0$), by what we have just seen there is an integer q such that $a = q(-b) + r$ with either $r = 0$ or $|r| < |-b|$; then $a = (-q)b + r$ satisfies the requirements of the Division Algorithm for a and b. This argument can be made more formal by using induction on $|a|$.

Note that if a is not a multiple of b there are always two possibilities for the pair q, r: the proof above always produced a positive remainder r. If for example $b > 0$ and q, r are as above with $r > 0$, then $a = q'b + r'$ with $q' = q + 1$ and $r' = r - b$ also satisfy the conditions of the Division Algorithm applied to a, b. Thus $5 = 2 \cdot 2 + 1 = 3 \cdot 2 - 1$ are the two ways of applying the Division Algorithm in $\mathbb{Z}$ to $a = 5$ and $b = 2$. The quotient and remainder are unique if we require the remainder to be nonnegative.

(2) If F is a field, then the polynomial ring $F[x]$ is a Euclidean Domain with norm given by $N(p(x)) =$ the degree of $p(x)$. The Division Algorithm for polynomials is simply "long division" of polynomials which may be familiar for polynomials with real coefficients. The proof is very similar to that for $\mathbb{Z}$ and is given in the next chapter (although for polynomials the quotient and remainder are shown to be unique). In order for a polynomial ring to be a Euclidean Domain the coefficients must come from a field since the division algorithm ultimately rests on being able to divide arbitrary nonzero coefficients. We shall prove in Section 2 that $R[x]$ is not a Euclidean Domain if R is not a field.

(3) The quadratic integer rings $\mathcal{O}$ in Section 7.1 are integral domains with a norm defined by the absolute value of the field norm (to ensure the values taken are nonnegative;

when $D < 0$ the field norm is itself a norm), but in general $\mathcal{O}$ is not Euclidean with respect to this norm (or any other norm). The Gaussian integers $\mathbb{Z}[i]$ (where $D = -1$), however, are a Euclidean Domain with respect to the norm $N(a + bi) = a^2 + b^2$, as we now show (cf. also the end of Section 3).

Let $\alpha = a + bi$, $\beta = c + di$ be two elements of $\mathbb{Z}[i]$ with $\beta \neq 0$. Then in the field $\mathbb{Q}(i)$ we have $\dfrac{\alpha}{\beta} = r + si$ where $r = (ac + bd)/(c^2 + d^2)$ and $s = (bc - ad)/(c^2 + d^2)$ are rational numbers. Let p be an integer closest to the rational number r and let q be an integer closest to the rational number s, so that both $|r - p|$ and $|s - q|$ are at most $1/2$. The Division Algorithm follows immediately once we show

$$\alpha = (p + qi)\beta + \gamma \qquad \text{for some } \gamma \in \mathbb{Z}[i] \text{ with } \quad N(\gamma) \leq \frac{1}{2}N(\beta)$$

which is even stronger than necessary. Let $\theta = (r - p) + (s - q)i$ and set $\gamma = \beta\theta$. Then $\gamma = \alpha - (p + qi)\beta$, so that $\gamma \in \mathbb{Z}[i]$ is a Gaussian integer and $\alpha = (p + qi)\beta + \gamma$. Since $N(\theta) = (r - p)^2 + (s - q)^2$ is at most $1/4 + 1/4 = 1/2$, the multiplicativity of the norm N implies that $N(\gamma) = N(\theta)N(\beta) \leq \dfrac{1}{2}N(\beta)$ as claimed.

Note that the algorithm is quite explicit since a quotient $p + qi$ is quickly determined from the rational numbers r and s, and then the remainder $\gamma = \alpha - (p + qi)\beta$ is easily computed. Note also that the quotient need not be unique: if r (or s) is half of an odd integer then there are two choices for p (or for q, respectively).

This proof that $\mathbb{Z}[i]$ is a Euclidean Domain can also be used to show that $\mathcal{O}$ is a Euclidean Domain (with respect to the field norm defined in Section 7.1) for $D = -2, -3, -7, -11$ (cf. the exercises). We shall see shortly that $\mathbb{Z}[\sqrt{-5}]$ is not Euclidean with respect to any norm, and a proof that $\mathbb{Z}[(1 + \sqrt{-19})/2]$ is not a Euclidean Domain with respect to any norm appears at the end of this section.

(4) Recall (cf. Exercise 26 in Section 7.1) that a *discrete valuation ring* is obtained as follows. Let K be a field. A *discrete valuation* on K is a function $v : K^\times \to \mathbb{Z}$ satisfying

(i) $v(ab) = v(a) + v(b)$ (i.e., v is a homomorphism from the multiplicative group of nonzero elements of K to $\mathbb{Z}$),

(ii) v is surjective, and

(iii) $v(x + y) \geq \min\{v(x), v(y)\}$ for all $x, y \in K^\times$ with $x + y \neq 0$.

The set $\{x \in K^\times \mid v(x) \geq 0\} \cup \{0\}$ is a subring of K called the valuation ring of v. An integral domain R is called a discrete valuation ring if there is a valuation v on its field of fractions such that R is the valuation ring of v.

For example the ring R of all rational numbers whose denominators are relatively prime to the fixed prime $p \in \mathbb{Z}$ is a discrete valuation ring contained in $\mathbb{Q}$.

A discrete valuation ring is easily seen to be a Euclidean Domain with respect to the norm defined by $N(0) = 0$ and $N = v$ on the nonzero elements of R. This is because for $a, b \in R$ with $b \neq 0$

(a) if $N(a) < N(b)$ then $a = 0 \cdot b + a$, and

(b) if $N(a) \geq N(b)$ then it follows from property (i) of a discrete valuation that $q = ab^{-1} \in R$, so $a = qb + 0$.

The first implication of a Division Algorithm for the integral domain R is that it forces every ideal of R to be *principal.*

Proposition 1. Every ideal in a Euclidean Domain is principal. More precisely, if I is any nonzero ideal in the Euclidean Domain R then $I = (d)$, where d is any nonzero element of I of minimum norm.

Proof: If I is the zero ideal, there is nothing to prove. Otherwise let d be any nonzero element of I of minimum norm (such a d exists since the set $\{N(a) \mid a \in I\}$ has a minimum element by the Well Ordering of $\mathbb{Z}$). Clearly $(d) \subseteq I$ since d is an element of I. To show the reverse inclusion let a be any element of I and use the Division Algorithm to write $a = qd + r$ with $r = 0$ or $N(r) < N(d)$. Then $r = a - qd$ and both a and qd are in I, so r is also an element of I. By the minimality of the norm of d, we see that r must be 0. Thus $a = qd \in (d)$ showing $I = (d)$.

Proposition 1 shows that every ideal of $\mathbb{Z}$ is principal. This fundamental property of $\mathbb{Z}$ was previously determined (in Section 7.3) from the (additive) group structure of $\mathbb{Z}$, using the classification of the subgroups of cyclic groups in Section 2.3. Note that these are really the same proof, since the results in Section 2.3 ultimately relied on the Euclidean Algorithm in $\mathbb{Z}$.

Proposition 1 can also be used to prove that some integral domains R are *not* Euclidean Domains (with respect to *any* norm) by proving the existence of ideals of R that are not principal.

Examples

(1) Let $R = \mathbb{Z}[x]$. Since the ideal $(2, x)$ is not principal (cf. Example 3 at the beginning of Section 7.4), it follows that the ring $\mathbb{Z}[x]$ of polynomials with *integer* coefficients is *not* a Euclidean Domain (for any choice of norm), even though the ring $\mathbb{Q}[x]$ of polynomials with *rational* coefficients is a Euclidean Domain.

(2) Let R be the quadratic integer ring $\mathbb{Z}[\sqrt{-5}]$, let N be the associated field norm $N(a+b\sqrt{-5}) = a^2+5b^2$ and consider the ideal $I = (3, 2+\sqrt{-5})$ generated by 3 and $2+\sqrt{-5}$. Suppose $I = (a+b\sqrt{-5})$, $a, b \in \mathbb{Z}$, were principal, i.e., $3 = \alpha(a+b\sqrt{-5})$ and $2+\sqrt{-5} = \beta(a+b\sqrt{-5})$ for some $\alpha, \beta \in R$. Taking norms in the first equation gives $9 = N(\alpha)(a^2+5b^2)$ and since a^2+5b^2 is a positive integer it must be 1,3 or 9. If the value is 9 then $N(\alpha) = 1$ and $\alpha = \pm 1$, so $a+b\sqrt{-5} = \pm 3$, which is impossible by the second equation since the coefficients of $2+\sqrt{-5}$ are not divisible by 3. The value cannot be 3 since there are no integer solutions to $a^2+5b^2 = 3$. If the value is 1, then $a+b\sqrt{-5} = \pm 1$ and the ideal I would be the entire ring R. But then 1 would be an element of I, so $3\gamma + (2+\sqrt{-5})\delta = 1$ for some $\gamma, \delta \in R$. Multiplying both sides by $2-\sqrt{-5}$ would then imply that $2-\sqrt{-5}$ is a multiple of 3 in R, a contradiction. It follows that I is not a principal ideal and so R is not a Euclidean Domain (with respect to any norm).

One of the fundamental consequences of the Euclidean Algorithm in $\mathbb{Z}$ is that it produces a greatest common divisor of two nonzero elements. This is true in any Euclidean Domain. The notion of a greatest common divisor of two elements (if it exists) can be made precise in general rings.

Definition. Let R be a commutative ring and let $a, b \in R$ with $b \neq 0$.

(1) a is said to be a *multiple* of b if there exists an element $x \in R$ with $a = bx$. In this case b is said to *divide* a or be a *divisor* of a, written $b \mid a$.

(2) A *greatest common divisor* of a and b is a nonzero element d such that

(i) $d \mid a$ and $d \mid b$, and

(ii) if $d' \mid a$ and $d' \mid b$ then $d' \mid d$.

A greatest common divisor of a and b will be denoted by g.c.d.(a, b), or (abusing the notation) simply (a, b).

Note that $b \mid a$ in a ring R if and only if $a \in (b)$ if and only if $(a) \subseteq (b)$. In particular, if d is any divisor of both a and b then (d) must contain both a and b and hence must contain the ideal generated by a and b. The defining properties (i) and (ii) of a greatest common divisor of a and b translated into the language of ideals therefore become (respectively):

if I is the ideal of R generated by a and b, then d is a greatest common divisor of a and b if

(i) I is contained in the principal ideal (d), and

(ii) if (d') is any principal ideal containing I then $(d) \subseteq (d')$.

Thus a greatest common divisor of a and b (if such exists) is a generator for the unique smallest principal ideal containing a and b. There are rings in which greatest common divisors do not exist.

This discussion immediately gives the following *sufficient* condition for the existence of a greatest common divisor.

Proposition 2. If a and b are nonzero elements in the commutative ring R such that the ideal generated by a and b is a principal ideal (d), then d is a greatest common divisor of a and b.

This explains why the symbol (a, b) is often used to denote both the ideal generated by a and b and a greatest common divisor of a and b. An integral domain in which every ideal (a, b) generated by two elements is principal is called a *Bezout Domain*. The exercises in this and subsequent sections explore these rings and show that there are Bezout Domains containing nonprincipal (necessarily infinitely generated) ideals.

Note that the condition in Proposition 2 is *not* a *necessary* condition. For example, in the ring $R = \mathbb{Z}[x]$ the elements 2 and x generate a maximal, nonprincipal ideal (cf. the examples in Section 7.4). Thus $R = (1)$ is the unique principal ideal containing both 2 and x, so 1 is a greatest common divisor of 2 and x. We shall see other examples along these lines in Section 3.

Before returning to Euclidean Domains we examine the uniqueness of greatest common divisors.

Proposition 3. Let R be an integral domain. If two elements d and d' of R generate the same principal ideal, i.e., $(d) = (d')$, then $d' = ud$ for some unit u in R. In particular, if d and d' are both greatest common divisors of a and b, then $d' = ud$ for some unit u.

Proof: This is clear if either d or d' is zero so we may assume d and d' are nonzero. Since $d \in (d')$ there is some $x \in R$ such that $d = xd'$. Since $d' \in (d)$ there is some $y \in R$ such that $d' = yd$. Thus $d = xyd$ and so $d(1 - xy) = 0$. Since $d \neq 0$, $xy = 1$, that is, both x and y are units. This proves the first assertion. The second assertion follows from the first since any two greatest common divisors of a and b generate the same principal ideal (they divide each other).

One of the most important properties of Euclidean Domains is that greatest common divisors always exist and *can be computed algorithmically*.

Theorem 4. Let R be a Euclidean Domain and let a and b be nonzero elements of R. Let $d = r_n$ be the last nonzero remainder in the Euclidean Algorithm for a and b described at the beginning of this chapter. Then

(1) d is a greatest common divisor of a and b, and

(2) the principal ideal (d) is the ideal generated by a and b. In particular, d can be written as an *R-linear combination* of a and b, i.e., there are elements x and y in R such that

$$d = ax + by.$$

Proof: By Proposition 1, the ideal generated by a and b is principal so a, b do have a greatest common divisor, namely any element which generates the (principal) ideal (a, b). Both parts of the theorem will follow therefore once we show $d = r_n$ generates this ideal, i.e., once we show that

(i) $d \mid a$ and $d \mid b$ (so $(a, b) \subseteq (d)$)

(ii) d is an R-linear combination of a and b (so $(d) \subseteq (a, b)$).

To prove that d divides both a and b simply keep track of the divisibilities in the Euclidean Algorithm. Starting from the $(n+1)^{\text{st}}$ equation, $r_{n-1} = q_{n+1}r_n$, we see that $r_n \mid r_{n-1}$. Clearly $r_n \mid r_n$. By induction (proceeding from index n downwards to index 0) assume r_n divides r_{k+1} and r_k. By the $(k+1)^{\text{st}}$ equation, $r_{k-1} = q_{k+1}r_k + r_{k+1}$, and since r_n divides both terms on the right hand side we see that r_n also divides r_{k-1}. From the 1^{st} equation in the Euclidean Algorithm we obtain that r_n divides b and then from the 0^{th} equation we get that r_n divides a. Thus (i) holds.

To prove that r_n is in the ideal (a, b) generated by a and b proceed similarly by induction proceeding from equation (0) to equation (n). It follows from equation (0) that $r_0 \in (a, b)$ and by equation (1) that $r_1 = b - q_1r_0 \in (b, r_0) \subseteq (a, b)$. By induction assume $r_{k-1}, r_k \in (a, b)$. Then by the $(k+1)^{\text{st}}$ equation

$$r_{k+1} = r_{k-1} - q_{k+1}r_k \in (r_{k-1}, r_k) \subseteq (a, b).$$

This induction shows $r_n \in (a, b)$, which completes the proof.

Much of the material above may be familiar from elementary arithmetic in the case of the integers $\mathbb{Z}$, except possibly for the translation into the language of ideals. For example, if $a = 2210$ and $b = 1131$ then the smallest ideal of $\mathbb{Z}$ that contains both a and b (the ideal generated by a and b) is $13\mathbb{Z}$, since 13 is the greatest common divisor of 2210 and 1131. This fact follows quickly from the Euclidean Algorithm:

$$2210 = 1 \cdot 1131 + 1079$$

$$1131 = 1 \cdot 1079 + 52$$
$$1079 = 20 \cdot 52 + 39$$
$$52 = 1 \cdot 39 + 13$$
$$39 = 3 \cdot 13$$

so that $13 = (2210, 1131)$ is the last nonzero remainder. Using the procedure of Theorem 4 we can also write 13 as a linear combination of 2210 and 1131 by first solving the next to last equation above for $13 = 52 - 1 \cdot 39$, then using previous equations to solve for 39 and 52, etc., finally writing 13 entirely in terms of 2210 and 1131. The answer in this case is

$$13 = (-22) \cdot 2210 + 43 \cdot 1131.$$

The Euclidean Algorithm in the integers $\mathbb{Z}$ is extremely fast. It is a theorem that the number of steps required to determine the greatest common divisor of two integers a and b is at worst 5 times the number of digits of the smaller of the two numbers. Put another way, this algorithm is *logarithmic* in the size of the integers. To obtain an appreciation of the speed implied here, notice that for the example above we would have expected at worst $5 \cdot 4 = 20$ divisions (the example required far fewer). If we had started with integers on the order of 10^{100} (large numbers by physical standards), we would have expected at worst only 500 divisions.

There is no uniqueness statement for the integers x and y in $(a, b) = ax + by$. Indeed, $x' = x + b$ and $y' = y - a$ satisfy $(a, b) = ax' + by'$. This is essentially the only possibility — one can prove that if x_0 and y_0 are solutions to the equation $ax + by = N$, then any other solutions x and y to this equation are of the form

$$x = x_0 + m\frac{b}{(a, b)}$$
$$y = y_0 - m\frac{a}{(a, b)}$$

for some integer m (positive or negative).

This latter theorem (a proof of which is outlined in the exercises) provides a complete solution of the *First Order Diophantine Equation* $ax + by = N$ provided we know there is *at least one* solution to this equation. But the equation $ax + by = N$ is simply another way of stating that N is an element of the ideal generated by a and b. Since we know this ideal is just (d), the principal ideal generated by the greatest common divisor d of a and b, this is the same as saying $N \in (d)$, i.e., N is divisible by d. Hence, *the equation* $ax + by = N$ *is solvable in integers* x *and* y *if and only if* N *is divisible by the g.c.d. of* a *and* b (and then the result quoted above gives a full set of solutions of this equation).

We end this section with another criterion that can sometimes be used to prove that a given integral domain is not a Euclidean Domain.[1] For any integral domain let

[1]The material here and in some of the following section follows the exposition by J.C. Wilson in *A principal ideal ring that is not a Euclidean ring*, Math. Mag., 46(1973), pp. 34–38, of ideas of Th. Motzkin, and use a simplification by Kenneth S. Williams in *Note on non-Euclidean Principal Ideal Domains*, Math. Mag., 48(1975), pp. 176–177.

$\widetilde{R} = R^\times \cup \{0\}$ denote the collection of units of R together with 0. An element $u \in R - \widetilde{R}$ is called a *universal side divisor* if for every $x \in R$ there is some $z \in \widetilde{R}$ such that u divides $x - z$ in R, i.e., there is a type of "division algorithm" for u: every x may be written $x = qu + z$ where z is either zero or a unit. The existence of universal side divisors is a weakening of the Euclidean condition:

Proposition 5. Let R be an integral domain that is not a field. If R is a Euclidean Domain then there are universal side divisors in R.

Proof: Suppose R is Euclidean with respect to some norm N and let u be an element of $R - \widetilde{R}$ (which is nonempty since R is not a field) of minimal norm. For any $x \in R$, write $x = qu + r$ where r is either 0 or $N(r) < N(u)$. In either case the minimality of u implies $r \in \widetilde{R}$. Hence u is a universal side divisor in R.

Example

We can use Proposition 5 to prove that the quadratic integer ring $R = \mathbb{Z}[(1 + \sqrt{-19})/2]$ is not a Euclidean Domain with respect to any norm by showing that R contains no universal side divisors (we shall see in the next section that all of the ideals in R are principal, so the technique in the examples following Proposition 1 do not apply to this ring). We have already determined that ± 1 are the only units in R and so $\widetilde{R} = \{0, \pm 1\}$. Suppose $u \in R$ is a universal side divisor and let $N(a + b(1 + \sqrt{-19})/2) = a^2 + ab + 5b^2$ denote the field norm on R as in Section 7.1. Note that if $a, b \in \mathbb{Z}$ and $b \neq 0$ then $a^2 + ab + 5b^2 = (a + b/2)^2 + 19/4b^2 \geq 5$ and so the smallest nonzero values of N on R are 1 (for the units ± 1) and 4 (for ± 2). Taking $x = 2$ in the definition of a universal side divisor it follows that u must divide one of $2 - 0$ or 2 ± 1 in R, i.e., u is a nonunit divisor of 2 or 3 in R. If $2 = \alpha\beta$ then $4 = N(\alpha)N(\beta)$ and by the remark above it follows that one of α or β has norm 1, i.e., equals ± 1. Hence the only divisors of 2 in R are $\{\pm 1, \pm 2\}$. Similarly, the only divisors of 3 in R are $\{\pm 1, \pm 3\}$, so the only possible values for u are ± 2 or ± 3. Taking $x = (1 + \sqrt{-19})/2$ it is easy to check that none of x, $x \pm 1$ are divisible by ± 2 or ± 3 in R, so none of these is a universal side divisor.

EXERCISES

1. For each of the following five pairs of integers a and b, determine their greatest common divisor d and write d as a linear combination $ax + by$ of a and b.
   - **(a)** $a = 20, b = 13$.
   - **(b)** $a = 69, b = 372$.
   - **(c)** $a = 11391, b = 5673$.
   - **(d)** $a = 507885, b = 60808$.
   - **(e)** $a = 91442056588823$, $b = 779086434385541$ (the Euclidean Algorithm requires only 7 steps for these integers).
2. For each of the following pairs of integers a and n, show that a is relatively prime to n and determine the inverse of a mod n (cf. Section 3 of the Preliminaries chapter).
   - **(a)** $a = 13, n = 20$.
   - **(b)** $a = 69, n = 89$.
   - **(c)** $a = 1891, n = 3797$.

(d) $a = 6003722857$, $n = 77695236973$ (the Euclidean Algorithm requires only 3 steps for these integers).

3. Let R be a Euclidean Domain. Let m be the minimum integer in the set of norms of nonzero elements of R. Prove that every nonzero element of R of norm m is a unit. Deduce that a nonzero element of norm zero (if such an element exists) is a unit.

4. Let R be a Euclidean Domain.

(a) Prove that if $(a, b) = 1$ and a divides bc, then a divides c. More generally, show that if a divides bc with nonzero a, b then $\frac{a}{(a, b)}$ divides c.

(b) Consider the Diophantine Equation $ax + by = N$ where a, b and N are integers and a, b are nonzero. Suppose x_0, y_0 is a solution: $ax_0 + by_0 = N$. Prove that the full set of solutions to this equation is given by

$$x = x_0 + m\frac{b}{(a, b)}, \qquad y = y_0 - m\frac{a}{(a, b)}$$

as m ranges over the integers. [If x, y is a solution to $ax + by = N$, show that $a(x - x_0) = b(y_0 - y)$ and use (a).]

5. Determine all integer solutions of the following equations:

(a) $2x + 4y = 5$

(b) $17x + 29y = 31$

(c) $85x + 145y = 505$.

6. (*The Postage Stamp Problem*) Let a and b be two relatively prime positive integers. Prove that every sufficiently large positive integer N can be written as a linear combination $ax + by$ of a and b where x and y are both *nonnegative*, i.e., there is an integer N_0 such that for all $N \geq N_0$ the equation $ax + by = N$ can be solved with both x and y nonnegative integers. Prove in fact that the integer $ab - a - b$ cannot be written as a positive linear combination of a and b but that every integer greater than $ab - a - b$ is a positive linear combination of a and b (so every "postage" greater than $ab - a - b$ can be obtained using only stamps in denominations a and b).

7. Find a generator for the ideal $(85, 1+13i)$ in $\mathbb{Z}[i]$, i.e., a greatest common divisor for 85 and $1+13i$, by the Euclidean Algorithm. Do the same for the ideal $(47 - 13i, 53 + 56i)$.

It is known (but not so easy to prove) that $D = -1, -2, -3, -7, -11, -19, -43, -67$, and -163 are the only negative values of D for which every ideal in $\mathcal{O}$ is principal (i.e., $\mathcal{O}$ is a P.I.D. in the terminology of the next section). The results of the next exercise determine precisely which quadratic integer rings with $D < 0$ are Euclidean.

8. Let $F = \mathbb{Q}(\sqrt{D})$ be a quadratic field with associated quadratic integer ring $\mathcal{O}$ and field norm N as in Section 7.1.

(a) Suppose D is $-1, -2, -3, -7$ or -11. Prove that $\mathcal{O}$ is a Euclidean Domain with respect to N. [Modify the proof for $\mathbb{Z}[i]$ ($D = -1$) in the text. For $D = -3, -7, -11$ prove that every element of F differs from an element in $\mathcal{O}$ by an element whose norm is at most $(1 + |D|)^2/(16|D|)$, which is less than 1 for these values of D. Plotting the points of $\mathcal{O}$ in $\mathbb{C}$ may be helpful.]

(b) Suppose that $D = -43, -67$, or -163. Prove that $\mathcal{O}$ is not a Euclidean Domain with respect to any norm. [Apply the same proof as for $D = -19$ in the text.]

9. Prove that the ring of integers $\mathcal{O}$ in the quadratic integer ring $\mathbb{Q}(\sqrt{2})$ is a Euclidean Domain with respect to the norm given by the absolute value of the field norm N in Section 7.1.

10. Prove that the quotient ring $\mathbb{Z}[i]/I$ is finite for any nonzero ideal I of $\mathbb{Z}[i]$. [Use the fact

that $I = (\alpha)$ for some nonzero α and then use the Division Algorithm in this Euclidean Domain to see that every coset of I is represented by an element of norm less than $N(\alpha)$.]

11. Let R be a commutative ring with 1 and let a and b be nonzero elements of R. A *least common multiple* of a and b is an element e of R such that
(i) $a \mid e$ and $b \mid e$, and
(ii) if $a \mid e'$ and $b \mid e'$ then $e \mid e'$.

(a) Prove that a least common multiple of a and b (if such exists) is a generator for the unique largest principal ideal contained in $(a) \cap (b)$.

(b) Deduce that any two nonzero elements in a Euclidean Domain have a least common multiple which is unique up to multiplication by a unit.

(c) Prove that in a Euclidean Domain the least common multiple of a and b is $\dfrac{ab}{(a, b)}$, where (a, b) is the greatest common divisor of a and b.

12. (*A Public Key Code*) Let N be a positive integer. Let M be an integer relatively prime to N and let d be an integer relatively prime to $\varphi(N)$, where φ denotes Euler's φ-function. Prove that if $M_1 \equiv M^d \pmod{N}$ then $M \equiv M_1^{d'} \pmod{N}$ where d' is the inverse of d mod $\varphi(N)$: $dd' \equiv 1 \pmod{\varphi(N)}$.

Remark: This result is the basis for a standard *Public Key Code*. Suppose $N = pq$ is the product of two distinct large primes (each on the order of 100 digits, for example). If M is a message, then $M_1 \equiv M^d \pmod{N}$ is a scrambled (*encoded*) version of M, which can be unscrambled (*decoded*) by computing $M_1^{d'} \pmod{N}$ (these powers can be computed quite easily even for large values of M and N by successive squarings). The values of N and d (but not p and q) are made publicly known (hence the name) and then anyone with a message M can send their encoded message $M^d \pmod{N}$. To decode the message it seems necessary to determine d', which requires the determination of the value $\varphi(N) = \varphi(pq) = (p-1)(q-1)$ (no one has as yet *proved* that there is no other decoding scheme, however). The success of this method as a code rests on the necessity of determining the *factorization* of N into primes, for which no sufficiently efficient algorithm exists (for example, the most naive method of checking all factors up to $\sqrt{N}$ would here require on the order of 10^{100} computations, or approximately 300 years even at 10 billion computations per second, and of course one can always increase the size of p and q).

8.2 PRINCIPAL IDEAL DOMAINS (P.I.D.s)

Definition. A *Principal Ideal Domain* (P.I.D.) is an integral domain in which every ideal is principal.

Proposition 1 proved that *every Euclidean Domain is a Principal Ideal Domain* so that every result about Principal Ideal Domains automatically holds for Euclidean Domains.

Examples

(1) As mentioned after Proposition 1, the integers $\mathbb{Z}$ are a P.I.D. We saw in Section 7.4 that the polynomial ring $\mathbb{Z}[x]$ contains nonprincipal ideals, hence is not a P.I.D.

(2) Example 2 following Proposition 1 showed that the quadratic integer ring $\mathbb{Z}[\sqrt{-5}]$ is not a P.I.D., in fact the ideal $(3, 1+\sqrt{-5})$ is a nonprincipal ideal. It is possible

for the product IJ of two nonprincipal ideals I and J to be principal, for example the ideals $(3, 1+\sqrt{-5})$ and $(3, 1-\sqrt{-5})$ are both nonprincipal and their product is the principal ideal generated by 3, i.e., $(3, 1+\sqrt{-5})(3, 1-\sqrt{-5}) = (3)$ (cf. Exercise 5 and the example preceding Proposition 12 below).

It is not true that every Principal Ideal Domain is a Euclidean Domain. We shall prove below that the quadratic integer ring $\mathbb{Z}[(1+\sqrt{-19})/2]$, which was shown not to be a Euclidean Domain in the previous section, nevertheless is a P.I.D.

From an ideal-theoretic point of view Principal Ideal Domains are a natural class of rings to study beyond rings which are fields (where the ideals are just the trivial ones: (0) and (1)). Many of the properties enjoyed by Euclidean Domains are also satisfied by Principal Ideal Domains. A significant advantage of Euclidean Domains over Principal Ideal Domains, however, is that although greatest common divisors exist in both settings, in Euclidean Domains one has an *algorithm* for computing them. Thus (as we shall see in Chapter 12 in particular) results which depend on the existence of greatest common divisors may often be proved in the larger class of Principal Ideal Domains although computation of examples (i.e., concrete applications of these results) are more effectively carried out using a Euclidean Algorithm (if one is available).

We collect some facts about greatest common divisors proved in the preceding section.

Proposition 6. Let R be a Principal Ideal Domain and let a and b be nonzero elements of R. Let d be a generator for the principal ideal generated by a and b. Then

(1) d is a greatest common divisor of a and b

(2) d can be written as an *R-linear combination* of a and b, i.e., there are elements x and y in R with

$$d = ax + by$$

(3) d is unique up to multiplication by a unit of R.

Proof: This is just Propositions 2 and 3.

Recall that maximal ideals are always prime ideals but the converse is not true in general. We observed in Section 7.4, however, that every nonzero prime ideal of $\mathbb{Z}$ is a maximal ideal. This useful fact is true in an arbitrary Principal Ideal Domain, as the following proposition shows.

Proposition 7. Every nonzero prime ideal in a Principal Ideal Domain is a maximal ideal.

Proof: Let (p) be a nonzero prime ideal in the Principal Ideal Domain R and let $I = (m)$ be any ideal containing (p). We must show that $I = (p)$ or $I = R$. Now $p \in (m)$ so $p = rm$ for some $r \in R$. Since (p) is a prime ideal and $rm \in (p)$, either r or m must lie in (p). If $m \in (p)$ then $(p) = (m) = I$. If, on the other hand, $r \in (p)$ write $r = ps$. In this case $p = rm = psm$, so $sm = 1$ (recall that R is an integral domain) and m is a unit so $I = R$.

As we have already mentioned, if F is a field, then the polynomial ring $F[x]$ is a Euclidean Domain, hence also a Principal Ideal Domain (this will be proved in the next chapter). The converse to this is also true. Intuitively, if I is an ideal in R (such as the ideal (2) in $\mathbb{Z}$) then the ideal (I, x) in $R[x]$ (such as the ideal $(2, x)$ in $\mathbb{Z}[x]$) requires one more generator than does I, hence in general is not principal.

Corollary 8. If R is any commutative ring such that the polynomial ring $R[x]$ is a Principal Ideal Domain (or a Euclidean Domain), then R is necessarily a field.

Proof: Assume $R[x]$ is a Principal Ideal Domain. Since R is a subring of $R[x]$ then R must be an integral domain (recall that $R[x]$ has an identity if and only if R does). The ideal (x) is a nonzero prime ideal in $R[x]$ because $R[x]/(x)$ is isomorphic to the integral domain R. By Proposition 7, (x) is a maximal ideal, hence the quotient R is a field by Proposition 12 in Section 7.4.

The last result in this section will be used to prove that not every P.I.D. is a Euclidean Domain and relates the principal ideal property with another weakening of the Euclidean condition.

Definition. Define N to be a *Dedekind–Hasse norm* if N is a positive norm and for every nonzero $a, b \in R$ either a is an element of the ideal (b) or there is a nonzero element in the ideal (a, b) of norm strictly smaller than the norm of b (i.e., either b divides a in R or there exist $s, t \in R$ with $0 < N(sa - tb) < N(b)$).

Note that R is Euclidean with respect to a positive norm N if it is always possible to satisfy the Dedekind–Hasse condition with $s = 1$, so this is indeed a weakening of the Euclidean condition.

Proposition 9. The integral domain R is a P.I.D. if and only if R has a Dedekind–Hasse norm.[2]

Proof: Let I be any nonzero ideal in R and let b be a nonzero element of I with $N(b)$ minimal. Suppose a is any nonzero element in I, so that the ideal (a, b) is contained in I. Then the Dedekind–Hasse condition on N and the minimality of b implies that $a \in (b)$, so $I = (b)$ is principal. The converse will be proved in the next section (Corollary 16).

---

[2]That a Dedekind–Hasse norm on R implies that R is a P.I.D. (and is equivalent when R is a ring of algebraic integers) is the classical *Criterion of Dedekind and Hasse*, cf. *Über eindeutige Zerlegung in Primelemente oder in Primhauptideale in Integritätsbereichen*, Jour. für die Reine und Angew. Math., 159(1928), pp. 3–12. The observation that the converse holds generally is more recent and due to John Greene, *Principal Ideal Domains are almost Euclidean*, Amer. Math. Monthly, 104(1997), pp. 154–156.

Example

Let $R = \mathbb{Z}[(1+\sqrt{-19})/2]$ be the quadratic integer ring considered at the end of the previous section. We show that the positive field norm $N(a + b(1 + \sqrt{-19})/2) = a^2 + ab + 5b^2$ defined on R is a Dedekind–Hasse norm, which by Proposition 9 and the results of the previous section will prove that R is a P.I.D. but not a Euclidean Domain.

Suppose α, β are nonzero elements of R and $\alpha/\beta \notin R$. We must show that there are elements $s, t \in R$ with $0 < N(s\alpha - t\beta) < N(\beta)$, which by the multiplicativity of the field norm is equivalent to

$$0 < N(\frac{\alpha}{\beta}s - t) < 1. \qquad (*)$$

Write $\frac{\alpha}{\beta} = \frac{a + b\sqrt{-19}}{c} \in \mathbb{Q}[\sqrt{-19}]$ with integers a, b, c having no common divisor and with $c > 1$ (since β is assumed not to divide α). Since a, b, c have no common divisor there are integers x, y, z with $ax + by + cz = 1$. Write $ay - 19bx = cq + r$ for some quotient q and remainder r with $|r| \leq c/2$ and let $s = y + x\sqrt{-19}$ and $t = q - z\sqrt{-19}$. Then a quick computation shows that

$$0 < N(\frac{\alpha}{\beta}s - t) = \frac{(ay - 19bx - cq)^2 + 19(ax + by + cz)^2}{c^2} \leq \frac{1}{4} + \frac{19}{c^2}$$

and so $(*)$ is satisfied with this s and t provided $c \geq 5$.

Suppose that $c = 2$. Then one of a, b is even and the other is odd (otherwise $\alpha/\beta \in R$), and then a quick check shows that $s = 1$ and $t = \frac{(a-1) + b\sqrt{-19}}{2}$ are elements of R satisfying $(*)$.

Suppose that $c = 3$. The integer $a^2 + 19b^2$ is not divisible by 3 (modulo 3 this is $a^2 + b^2$ which is easily seen to be 0 modulo 3 if and only if a and b are both 0 modulo 3; but then a, b, c have a common factor). Write $a^2 + 19b^2 = 3q + r$ with $r = 1$ or 2. Then again a quick check shows that $s = a - b\sqrt{-19}$, $t = q$ are elements of R satisfying $(*)$.

Finally, suppose that $c = 4$, so a and b are not both even. If one of a, b is even and the other odd, then $a^2 + 19b^2$ is odd, so we can write $a^2 + 19b^2 = 4q + r$ for some $q, r \in \mathbb{Z}$ and $0 < r < 4$. Then $s = a - b\sqrt{-19}$ and $t = q$ satisfy $(*)$. If a and b are both odd, then $a^2 + 19b^2 \equiv 1 + 3 \bmod 8$, so we can write $a^2 + 19b^2 = 8q + 4$ for some $q \in \mathbb{Z}$. Then $s = \frac{a - b\sqrt{-19}}{2}$ and $t = q$ are elements of R that satisfy $(*)$.

EXERCISES

1. Prove that in a Principal Ideal Domain two ideals (a) and (b) are comaximal (cf. Section 7.6) if and only if a greatest common divisor of a and b is 1 (in which case a and b are said to be *coprime* or *relatively prime*).
2. Prove that any two nonzero elements of a P.I.D. have a least common multiple (cf. Exercise 11, Section 1).
3. Prove that a quotient of a P.I.D. by a prime ideal is again a P.I.D.
4. Let R be an integral domain. Prove that if the following two conditions hold then R is a Principal Ideal Domain:
   (i) any two nonzero elements a and b in R have a greatest common divisor which can be written in the form $ra + sb$ for some $r, s \in R$, and

(ii) if $a_1, a_2, a_3, \ldots$ are nonzero elements of R such that $a_{i+1} \mid a_i$ for all i, then there is a positive integer N such that a_n is a unit times a_N for all $n \geq N$.

5. Let R be the quadratic integer ring $\mathbb{Z}[\sqrt{-5}]$. Define the ideals $I_2 = (2, 1+\sqrt{-5})$, $I_3 = (3, 2+\sqrt{-5})$, and $I_3' = (3, 2-\sqrt{-5})$.
   (a) Prove that I_2, I_3, and I_3' are nonprincipal ideals in R. [Note that Example 2 following Proposition 1 proves this for I_3.]
   (b) Prove that the product of two nonprincipal ideals can be principal by showing that I_2^2 is the principal ideal generated by 2, i.e., $I_2^2 = (2)$.
   (c) Prove similarly that $I_2 I_3 = (1-\sqrt{-5})$ and $I_2 I_3' = (1+\sqrt{-5})$ areprincipal. Conclude that the principal ideal (6) is the product of 4 ideals: $(6) = I_2^2 I_3 I_3'$.

6. Let R be an integral domain and suppose that every *prime* ideal in R is principal. This exercise proves that every ideal of R is principal, i.e., R is a P.I.D.
   (a) Assume that the set of ideals of R that are not principal is nonempty and prove that this set has a maximal element under inclusion (which, by hypothesis, is not prime). [Use Zorn's Lemma.]
   (b) Let I be an ideal which is maximal with respect to being nonprincipal, and let $a, b \in R$ with $ab \in I$ but $a \notin I$ and $b \notin I$. Let $I_a = (I, a)$ be the ideal generated by I and a, let $I_b = (I, b)$ be the ideal generated by I and b, and define $J = \{r \in R \mid rI_a \subseteq I\}$. Prove that $I_a = (\alpha)$ and $J = (\beta)$ are principal ideals in R with $I \subsetneq I_b \subseteq J$ and $I_a J = (\alpha\beta) \subseteq I$.
   (c) If $x \in I$ show that $x = s\alpha$ for some $s \in J$. Deduce that $I = I_a J$ is principal, a contradiction, and conclude that R is a P.I.D.

7. An integral domain R in which every ideal generated by two elements is principal (i.e., for every $a, b \in R$, $(a, b) = (d)$ for some $d \in R$) is called a *Bezout Domain.* [cf. also Exercise 11 in Section 3.]
   (a) Prove that the integral domain R is a Bezout Domain if and only if every pair of elements a, b of R has a g.c.d. d in R that can be written as an R-linear combination of a and b, i.e., $d = ax + by$ for some $x, y \in R$.
   (b) Prove that every finitely generated ideal of a Bezout Domain is principal. [cf. the exercises in Sections 9.2 and 9.3 for Bezout Domains in which not every ideal is principal.]
   (c) Let F be the fraction field of the Bezout Domain R. Prove that every element of F can be written in the form a/b with $a, b \in R$ and a and b relatively prime (cf. Exercise 1).

8. Prove that if R is a Principal Ideal Domain and D is a multiplicatively closed subset of R, then $D^{-1}R$ is also a P.I.D. (cf. Section 7.5).

8.3 UNIQUE FACTORIZATION DOMAINS (U.F.D.s)

In the case of the integers $\mathbb{Z}$, there is another method for determining the greatest common divisor of two elements a and b familiar from elementary arithmetic, namely the notion of "factorization into primes" for a and b, from which the greatest common divisor can easily be determined. This can also be extended to a larger class of rings called Unique Factorization Domains (U.F.D.s) — these will be defined shortly. We shall then prove that

every Principal Ideal Domain is a Unique Factorization Domain

so that every result about Unique Factorization Domains will automatically hold for both Euclidean Domains and Principal Ideal Domains.

We first introduce some terminology.

Definition. Let R be an integral domain.

(1) Suppose $r \in R$ is nonzero and is not a unit. Then r is called *irreducible* in R if whenever $r = ab$ with $a, b \in R$, at least one of a or b must be a unit in R. Otherwise r is said to be *reducible*.

(2) The nonzero element $p \in R$ is called *prime* in R if the ideal (p) generated by p is a prime ideal. In other words, a nonzero element p is a prime if it is not a unit and whenever $p \mid ab$ for any $a, b \in R$, then either $p \mid a$ or $p \mid b$.

(3) Two elements a and b of R differing by a unit are said to be *associate* in R (i.e., $a = ub$ for some unit u in R).

Proposition 10. In an integral domain a prime element is always irreducible.

Proof: Suppose (p) is a nonzero prime ideal and $p = ab$. Then $ab = p \in (p)$, so by definition of prime ideal one of a or b, say a, is in (p). Thus $a = pr$ for some r. This implies $p = ab = prb$ so $rb = 1$ and b is a unit. This shows that p is irreducible.

It is not true in general that an irreducible element is necessarily prime. For example, consider the element 3 in the quadratic integer ring $R = \mathbb{Z}[\sqrt{-5}]$. The computations in Section 1 show that 3 is irreducible in R, but 3 is not a prime since $(2+\sqrt{-5})(2-\sqrt{-5}) = 3^2$ is divisible by 3, but neither $2+\sqrt{-5}$ nor $2-\sqrt{-5}$ is divisible by 3 in R.

If R is a Principal Ideal Domain however, the notions of prime and irreducible elements are the same. In particular these notions coincide in $\mathbb{Z}$ and in $F[x]$ (where F is a field).

Proposition 11. In a Principal Ideal Domain a nonzero element is a prime if and only if it is irreducible.

Proof: We have shown above that prime implies irreducible. We must show conversely that if p is irreducible, then p is a prime, i.e., the ideal (p) is a prime ideal. If M is any ideal containing (p) then by hypothesis $M = (m)$ is a principal ideal. Since $p \in (m)$, $p = rm$ for some r. But p is irreducible so by definition either r or m is a unit. This means either $(p) = (m)$ or $(m) = (1)$, respectively. Thus the only ideals containing (p) are (p) or (1), i.e., (p) is a maximal ideal. Since maximal ideals are prime ideals, the proof is complete.

Example

Proposition 11 gives another proof that the quadratic integer ring $\mathbb{Z}[\sqrt{-5}]$ is not a P.I.D. since 3 is irreducible but not prime in this ring.

The irreducible elements in the integers $\mathbb{Z}$ are the prime numbers (and their negatives) familiar from elementary arithmetic, and two integers a and b are associates of one another if and only if $a = \pm b$.

In the integers $\mathbb{Z}$ any integer n can be written as a product of primes (not necessarily distinct), as follows. If n is not itself a prime then by definition it is possible to write $n = n_1 n_2$ for two other integers n_1 and n_2 neither of which is a unit, i.e., neither of which is ± 1. Both n_1 and n_2 must be smaller in absolute value than n itself. If they are both primes, we have already written n as a product of primes. If one of n_1 or n_2 is not prime, then it in turn can be factored into two (smaller) integers. Since integers cannot decrease in absolute value indefinitely, we must at some point be left only with prime integer factors, and so we have written n as a product of primes.

For example, if $n = 2210$, the algorithm above proceeds as follows: n is not itself prime, since we can write $n = 2 \cdot 1105$. The integer 2 is a prime, but 1105 is not: $1105 = 5 \cdot 221$. The integer 5 is prime, but 221 is not: $221 = 13 \cdot 17$. Here the algorithm terminates, since both 13 and 17 are primes. This gives the *prime factorization* of 2210 : $2210 = 2 \cdot 5 \cdot 13 \cdot 17$. Similarly, we find $1131 = 3 \cdot 13 \cdot 29$. In these examples each prime occurs only to the first power, but of course this need not be the case generally.

In the ring $\mathbb{Z}$ not only is it true that every integer n can be written as a product of primes, but in fact this decomposition is *unique* in the sense that any two prime factorizations of the same positive integer n differ only in the order in which the positive prime factors are written. The restriction to positive integers is to avoid considering the factorizations (3)(5) and $(-3)(-5)$ of 15 as essentially distinct. This *unique factorization* property of $\mathbb{Z}$ (which we shall prove very shortly) is extremely useful for the arithmetic of the integers. General rings with the analogous property are given a name.

Definition. A *Unique Factorization Domain* (U.F.D.) is an integral domain R in which every nonzero element $r \in R$ which is not a unit has the following two properties:

- **(i)** r can be written as a finite product of irreducibles p_i of R (not necessarily distinct): $r = p_1 p_2 \cdots p_n$ and
- **(ii)** the decomposition in (i) is *unique up to associates*: namely, if $r = q_1 q_2 \cdots q_m$ is another factorization of r into irreducibles, then $m = n$ and there is some renumbering of the factors so that p_i is associate to q_i for $i = 1, 2, \ldots, n$.

Examples

- **(1)** A field F is trivially a Unique Factorization Domain since every nonzero element is a unit, so there are no elements for which properties (i) and (ii) must be verified.
- **(2)** As indicated above, we shall prove shortly that every Principal Ideal Domain is a Unique Factorization Domain (so, in particular, $\mathbb{Z}$ and $F[x]$ where F is a field are both Unique Factorization Domains).
- **(3)** We shall also prove in the next chapter that the ring $R[x]$ of polynomials is a Unique Factorization Domain whenever R itself is a Unique Factorization Domain (in contrast to the properties of being a Principal Ideal Domain or being a Euclidean Domain, which do not carry over from a ring R to the polynomial ring $R[x]$). This result together with the preceding example will show that $\mathbb{Z}[x]$ is a Unique Factorization Domain.
- **(4)** The subring of the Gaussian integers $R = \mathbb{Z}[2i] = \{a + 2bi \mid a, b \in \mathbb{Z}\}$, where $i^2 = -1$, is an integral domain but not a Unique Factorization Domain (rings of this nature were introduced in Exercise 23 of Section 7.1). The elements 2 and $2i$ are

irreducibles which are not associates in R since $i \notin R$, and $4 = 2 \cdot 2 = (-2i) \cdot (2i)$ has two distinct factorizations in R. One may also check directly that $2i$ is irreducible but not prime in R (since $R/(2i) \cong \mathbb{Z}/4\mathbb{Z}$). In the larger ring of Gaussian integers, $\mathbb{Z}[i]$, (which is a Unique Factorization Domain) 2 and $2i$ *are* associates since i is a unit in this larger ring. We shall give a slightly different proof that $\mathbb{Z}[2i]$ is not a Unique Factorization Domain at the end of Section 9.3 (one in which we do not have to check that 2 and $2i$ are irreducibles).

(5) The quadratic integer ring $\mathbb{Z}[\sqrt{-5}]$ is another example of an integral domain that is not a Unique Factorization Domain, since $6 = 2 \cdot 3 = (1 + \sqrt{-5})(1 - \sqrt{-5})$ gives two distinct factorizations of 6 into irreducibles. The principal ideal (6) in $\mathbb{Z}[\sqrt{-5}]$ can be written as a product of 4 nonprincipal prime ideals: $(6) = P_2^2 P_3 P_3'$ and the two distinct factorizations of the element 6 in $\mathbb{Z}[\sqrt{-5}]$ can be interpreted as arising from two rearrangements of this product of ideals into products of principal ideals: the product of $P_2^2 = (2)$ with $P_3 P_3' = (3)$, and the product of $P_2 P_3 = (1 + \sqrt{-5})$ with $P_2 P_3' = (1 - \sqrt{-5})$ (cf. Exercise 8).

While the *elements* of the quadratic integer ring $\mathcal{O}$ need not have unique factorization, it is a theorem (Corollary 16.16) that every *ideal* in $\mathcal{O}$ can be written uniquely as a product of prime *ideals*. The unique factorization of ideals into the product of prime ideals holds in general for rings of integers of algebraic number fields (examples of which are the quadratic integer rings) and leads to the notion of a Dedekind Domain considered in Chapter 16. It was the failure to have unique factorization into irreducibles for elements in algebraic integer rings in number theory that originally led to the definition of an ideal. The resulting uniqueness of the decomposition into prime ideals in these rings gave the elements of the ideals an "ideal" (in the sense of "perfect" or "desirable") behavior that is the basis for the choice of terminology for these (now fundamental) algebraic objects.

The first property of irreducible elements in a Unique Factorization Domain is that they are also primes. One might think that we could deduce Proposition 11 from this proposition together with the previously mentioned theorem (that we shall prove shortly) that every Principal Ideal Domain is a Unique Factorization Domain, however Proposition 11 will be used in the proof of the latter theorem.

Proposition 12. In a Unique Factorization Domain a nonzero element is a prime if and only if it is irreducible.

Proof: Let R be a Unique Factorization Domain. Since by Proposition 10, primes of R are irreducible it remains to prove that each irreducible element is a prime. Let p be an irreducible in R and assume $p \mid ab$ for some $a, b \in R$; we must show that p divides either a or b. To say that p divides ab is to say $ab = pc$ for some c in R. Writing a and b as a product of irreducibles, we see from this last equation and from the *uniqueness* of the decomposition into irreducibles of ab that the irreducible element p must be *associate* to one of the irreducibles occurring either in the factorization of a or in the factorization of b. We may assume that p is associate to one of the irreducibles in the factorization of a, i.e., that a can be written as a product $a = (up)p_2 \cdots p_n$ for u a unit and some (possibly empty set of) irreducibles $p_2, \ldots, p_n$. But then p divides a, since $a = pd$ with $d = up_2 \cdots p_n$, completing the proof.

In a Unique Factorization Domain we shall now use the terms "prime" and "irreducible" interchangeably although we shall usually refer to the "primes" in $\mathbb{Z}$ and the "irreducibles" in $F[x]$.

We shall use the preceding proposition to show that in a Unique Factorization Domain any two nonzero elements a and b have a greatest common divisor:

Proposition 13. Let a and b be two nonzero elements of the Unique Factorization Domain R and suppose

$$a = up_1^{e_1}p_2^{e_2}\cdots p_n^{e_n} \qquad \text{and} \qquad b = vp_1^{f_1}p_2^{f_2}\cdots p_n^{f_n}$$

are prime factorizations for a and b, where u and v are units, the primes $p_1, p_2, \ldots, p_n$ are *distinct* and the exponents e_i and f_i are ≥ 0. Then the element

$$d = p_1^{\min(e_1, f_1)}p_2^{\min(e_2, f_2)}\cdots p_n^{\min(e_n, f_n)}$$

(where $d = 1$ if all the exponents are 0) is a greatest common divisor of a and b.

Proof: Since the exponents of each of the primes occurring in d are no larger than the exponents occurring in the factorizations of both a and b, d divides both a and b. To show that d is a greatest common divisor, let c be any common divisor of a and b and let $c = q_1^{g_1}q_2^{g_2}\cdots q_m^{g_m}$ be the prime factorization of c. Since each q_i divides c, hence divides a and b, we see from the preceding proposition that q_i must divide one of the primes p_j. In particular, up to associates (so up to multiplication by a unit) the primes occurring in c must be a subset of the primes occurring in a and b : $\{q_1, q_2, \ldots, q_m\} \subseteq \{p_1, p_2, \ldots, p_n\}$. Similarly, the exponents for the primes occurring in c must be no larger than those occurring in d. This implies that c divides d, completing the proof.

Example

In the example above, where $a = 2210$ and $b = 1131$, we find immediately from their prime factorizations that $(a, b) = 13$. Note that if the prime factorizations for a and b are known, the proposition above gives their greatest common divisor instantly, but that finding these prime factorizations is extremely time-consuming computationally. The Euclidean Algorithm is the fastest method for determining the g.c.d. of two integers but unfortunately it gives almost no information on the prime factorizations of the integers.

We now come to one of the principal results relating some of the rings introduced in this chapter.

Theorem 14. Every Principal Ideal Domain is a Unique Factorization Domain. In particular, every Euclidean Domain is a Unique Factorization Domain.

Proof: Note that the second assertion follows from the first since Euclidean Domains are Principal Ideal Domains. To prove the first assertion let R be a Principal Ideal Domain and let r be a nonzero element of R which is not a unit. We must show first that r can be written as a finite product of irreducible elements of R and then we must verify that this decomposition is unique up to units.

The method of proof of the first part is precisely analogous to the determination of the prime factor decomposition of an integer. Assume r is nonzero and is not a unit. If r is itself irreducible, then we are done. If not, then by definition r can be written as a product $r = r_1 r_2$ where neither r_1 nor r_2 is a unit. If both these elements are irreducibles, then again we are done, having written r as a product of irreducible elements. Otherwise, at least one of the two elements, say r_1 is reducible, hence can be written as a product of two nonunit elements $r_1 = r_{11} r_{12}$, and so forth. What we must verify is that this process *terminates*, i.e., that we must necessarily reach a point where all of the elements obtained as factors of r are irreducible. Suppose this is not the case. From the factorization $r = r_1 r_2$ we obtain a *proper* inclusion of ideals: $(r) \subset (r_1) \subset R$. The first inclusion is proper since r_2 is not a unit, and the last inclusion is proper since r_1 is not a unit. From the factorization of r_1 we similarly obtain $(r) \subset (r_1) \subset (r_{11}) \subset R$. If this process of factorization did not terminate after a finite number of steps, then we would obtain an *infinite ascending chain* of ideals:

$$(r) \subset (r_1) \subset (r_{11}) \subset \cdots \subset R$$

where all containments are proper, and the Axiom of Choice ensures that an infinite chain exists (cf. Appendix I).

We now show that any ascending chain $I_1 \subseteq I_2 \subseteq \cdots \subseteq R$ of ideals in a Principal Ideal Domain eventually becomes stationary, i.e., there is some positive integer n such that $I_k = I_n$ for all $k \geq n$.[3] In particular, it is not possible to have an infinite ascending chain of ideals where all containments are proper. Let $I = \cup_{i=1}^{\infty} I_i$. It follows easily (as in the proof of Proposition 11 in Section 7.4) that I is an ideal. Since R is a Principal Ideal Domain it is principally generated, say $I = (a)$. Since I is the union of the ideals above, a must be an element of one of the ideals in the chain, say $a \in I_n$. But then we have $I_n \subseteq I = (a) \subseteq I_n$ and so $I = I_n$ and the chain becomes stationary at I_n. This proves that every nonzero element of R which is not a unit has some factorization into irreducibles in R.

It remains to prove that the above decomposition is essentially unique. We proceed by induction on the number, n, of irreducible factors in some factorization of the element r. If $n = 0$, then r is a unit. If we had $r = qc$ (some other factorization) for some irreducible q, then q would divide a unit, hence would itself be a unit, a contradiction. Suppose now that n is at least 1 and that we have two products

$$r = p_1 p_2 \cdots p_n = q_1 q_2 \cdots q_m \qquad m \geq n$$

for r where the p_i and q_j are (not necessarily distinct) irreducibles. Since then p_1 divides the product on the right, we see by Proposition 11 that p_1 must divide one of the factors. Renumbering if necessary, we may assume p_1 divides q_1. But then $q_1 = p_1 u$ for some element u of R which must in fact be a unit since q_1 is irreducible. Thus p_1 and q_1 are associates. Cancelling p_1 (recall we are in an integral domain, so this is legitimate), we obtain the equation

$$p_2 \cdots p_n = u q_2 q_3 \cdots q_m = q_2' q_3 \cdots q_m \qquad m \geq n.$$

[3]This same argument can be used to prove the more general statement: an ascending chain of ideals becomes stationary in any commutative ring where all the ideals are *finitely generated*. This result will be needed in Chapter 12 where the details will be repeated.

where $q_2' = uq_2$ is again an irreducible (associate to q_2). By induction on n, we conclude that each of the factors on the left matches bijectively (up to associates) with the factors on the far right, hence with the factors in the middle (which are the same, up to associates). Since p_1 and q_1 (after the initial renumbering) have already been shown to be associate, this completes the induction step and the proof of the theorem.

Corollary 15. *(Fundamental Theorem of Arithmetic)* The integers $\mathbb{Z}$ are a Unique Factorization Domain.

Proof: The integers $\mathbb{Z}$ are a Euclidean Domain, hence are a Unique Factorization Domain by the theorem.

We can now complete the equivalence (Proposition 9) between the existence of a Dedekind–Hasse norm on the integral domain R and whether R is a P.I.D.

Corollary 16. Let R be a P.I.D. Then there exists a multiplicative Dedekind–Hasse norm on R.

Proof: If R is a P.I.D. then R is a U.F.D. Define the norm N by setting $N(0) = 0$, $N(u) = 1$ if u is a unit, and $N(a) = 2^n$ if $a = p_1 p_2 \cdots p_n$ where the p_i are irreducibles in R (well defined since the number of irreducible factors of a is unique). Clearly $N(ab) = N(a)N(b)$ so N is positive and multiplicative. To show that N is a Dedekind–Hasse norm, suppose that a, b are nonzero elements of R. Then the ideal generated by a and b is principal by assumption, say $(a, b) = (r)$. If a is not contained in the ideal (b) then also r is not contained in (b), i.e., r is not divisible by b. Since $b = xr$ for some $x \in R$, it follows that x is not a unit in R and so $N(b) = N(x)N(r) > N(r)$. Hence (a, b) contains a nonzero element with norm strictly smaller than the norm of b, completing the proof.

Factorization in the Gaussian Integers

We end our discussion of Unique Factorization Domains by describing the irreducible elements in the Gaussian integers $\mathbb{Z}[i]$ and the corresponding application to a famous theorem of Fermat in elementary number theory. This is particularly appropriate since the classical study of $\mathbb{Z}[i]$ initiated the algebraic study of rings.

In general, let $\mathcal{O}$ be a quadratic integer ring and let N be the associated field norm introduced in Section 7.1. Suppose $\alpha \in \mathcal{O}$ is an element whose norm is a prime p in $\mathbb{Z}$. If $\alpha = \beta\gamma$ for some $\beta, \gamma \in \mathcal{O}$ then $p = N(\alpha) = N(\beta)N(\gamma)$ so that one of $N(\beta)$ or $N(\gamma)$ is ± 1 and the other is $\pm p$. Since we have seen that an element of $\mathcal{O}$ has norm ± 1 if and only if it is a unit in $\mathcal{O}$, one of the factors of α is a unit. It follows that

if $N(\alpha)$ is $\pm$ a prime (in $\mathbb{Z}$), then α is irreducible in $\mathcal{O}$.

Suppose that π is a prime element in $\mathcal{O}$ and let (π) be the ideal generated by π in $\mathcal{O}$. Since (π) is a prime ideal in $\mathcal{O}$ it is easy to check that $(\pi) \cap \mathbb{Z}$ is a prime ideal in $\mathbb{Z}$ (if a and b are integers with $ab \in (\pi)$ then either a or b is an element of (π), so a or b is in $(\pi) \cap \mathbb{Z}$). Since $N(\pi)$ is a nonzero integer in (π) we have $(\pi) \cap \mathbb{Z} = p\mathbb{Z}$ for some integer prime p. It follows from $p \in (\pi)$ that π is a divisor in $\mathcal{O}$ of the

integer prime p, and so the prime elements in $\mathcal{O}$ can be found by determining how the primes in $\mathbb{Z}$ factor in the larger ring $\mathcal{O}$. Suppose π divides the prime p in $\mathcal{O}$, say $p = \pi\pi'$. Then $N(\pi)N(\pi') = N(p) = p^2$, so since π is not a unit there are only two possibilities: either $N(\pi) = \pm p^2$ or $N(\pi) = \pm p$. In the former case $N(\pi') = \pm 1$, hence π' is a unit and $p = \pi$ (up to associates) is irreducible in $\mathbb{Z}[i]$. In the latter case $N(\pi) = N(\pi') = \pm p$, hence π' is also irreducible and $p = \pi\pi'$ is the product of precisely two irreducibles.

Consider now the special case $D = -1$ of the Gaussian integers $\mathbb{Z}[i]$. We have seen that the units in $\mathbb{Z}[i]$ are the elements ± 1 and $\pm i$. We proved in Section 1 that $\mathbb{Z}[i]$ is a Euclidean Domain, hence is also a Principal Ideal Domain and a Unique Factorization Domain, so the irreducible elements are the same as the prime elements, and can be determined by seeing how the primes in $\mathbb{Z}$ factor in the larger ring $\mathbb{Z}[i]$.

In this case $\alpha = a + bi$ has $N(\alpha) = \alpha\overline{\alpha} = a^2 + b^2$, where $\overline{\alpha} = a - bi$ is the complex conjugate of α. It follows by what we just saw that *p factors in $\mathbb{Z}[i]$ into precisely two irreducibles if and only if $p = a^2 + b^2$ is the sum of two integer squares (otherwise p remains irreducible in $\mathbb{Z}[i]$).* If $p = a^2 + b^2$ then the corresponding irreducible elements in $\mathbb{Z}[i]$ are $a \pm bi$.

Clearly $2 = 1^2 + 1^2$ is the sum of two squares, giving the factorization $2 = (1+i)(1-i) = -i(1+i)^2$. The irreducibles $1+i$ and $1-i = -i(1+i)$ are associates and it is easy to check that this is the only situation in which conjugate irreducibles $a + bi$ and $a - bi$ can be associates.

Since the square of any integer is congruent to either 0 or 1 modulo 4, an odd prime in $\mathbb{Z}$ that is the sum of two squares must be congruent to 1 modulo 4. Thus if p is a prime of $\mathbb{Z}$ with $p \equiv 3 \bmod 4$ then p is not the sum of two squares and p remains irreducible in $\mathbb{Z}[i]$.

Suppose now that p is a prime of $\mathbb{Z}$ with $p \equiv 1 \bmod 4$. We shall prove that p cannot be irreducible in $\mathbb{Z}[i]$ which will show that $p = (a + bi)(a - bi)$ factors as the product of two distinct irreducibles in $\mathbb{Z}[i]$ or, equivalently, that $p = a^2 + b^2$ is the sum of two squares. We first prove the following result from elementary number theory:

Lemma 17. The prime number $p \in \mathbb{Z}$ divides an integer of the form $n^2 + 1$ if and only if p is either 2 or is an odd prime congruent to 1 modulo 4.

Proof: The statement for $p = 2$ is trivial since $2 \mid 1^2 + 1$. If p is an odd prime, note that $p \mid n^2 + 1$ is equivalent to $n^2 = -1$ in $\mathbb{Z}/p\mathbb{Z}$. This in turn is equivalent to saying the residue class of n is of order 4 in the multiplicative group $(\mathbb{Z}/p\mathbb{Z})^\times$. Thus p divides an integer of the form $n^2 + 1$ if and only if $(\mathbb{Z}/p\mathbb{Z})^\times$ contains an element of order 4. By Lagrange's Theorem, if $(\mathbb{Z}/p\mathbb{Z})^\times$ contains an element of order 4 then $|(\mathbb{Z}/p\mathbb{Z})^\times| = p - 1$ is divisible by 4, i.e., p is congruent to 1 modulo 4.

Conversely, suppose $p - 1$ is divisible by 4. We first argue that $(\mathbb{Z}/p\mathbb{Z})^\times$ contains a unique element of order 2. If $m^2 \equiv 1 \bmod p$ then p divides $m^2 - 1 = (m-1)(m+1)$. Thus p divides either $m - 1$ (i.e., $m \equiv 1 \bmod p$) or $m + 1$ (i.e., $m \equiv -1 \bmod p$), so -1 is the unique residue class of order 2 in $(\mathbb{Z}/p\mathbb{Z})^\times$. Now the abelian group $(\mathbb{Z}/p\mathbb{Z})^\times$ contains a subgroup H of order 4 (for example, the quotient by the subgroup $\{\pm 1\}$ contains a subgroup of order 2 whose preimage is a subgroup of order 4 in $(\mathbb{Z}/p\mathbb{Z})^\times$).

Since the Klein 4-group has three elements of order 2 whereas $(\mathbb{Z}/p\mathbb{Z})^\times$ — hence also H — has a unique element of order 2, H must be the cyclic group of order 4. Thus $(\mathbb{Z}/p\mathbb{Z})^\times$ contains an element of order 4, namely a generator for H.

Remark: We shall prove later (Corollary 19 in Section 9.5) that $(\mathbb{Z}/p\mathbb{Z})^\times$ is a cyclic group, from which it is immediate that there is an element of order 4 if and only if $p-1$ is divisible by 4.

By Lemma 17, if $p \equiv 1 \bmod 4$ is a prime then p divides n^2+1 in $\mathbb{Z}$ for some $n \in \mathbb{Z}$, so certainly p divides $n^2+1 = (n+i)(n-i)$ in $\mathbb{Z}[i]$. If p were irreducible in $\mathbb{Z}[i]$ then p would divide either $n+i$ or $n-i$ in $\mathbb{Z}[i]$. In this situation, since p is a real number, it would follow that p divides both $n+i$ and its complex conjugate $n-i$; hence p would divide their difference, $2i$. This is clearly not the case. We have proved the following result:

Proposition 18.

(1) *(Fermat's Theorem on sums of squares)* The prime p is the sum of two integer squares, $p = a^2+b^2$, $a, b \in \mathbb{Z}$, if and only if $p = 2$ or $p \equiv 1 \bmod 4$. Except for interchanging a and b or changing the signs of a and b, the representation of p as a sum of two squares is unique.

(2) The irreducible elements in the Gaussian integers $\mathbb{Z}[i]$ are as follows:

- **(a)** $1+i$ (which has norm 2),
- **(b)** the primes $p \in \mathbb{Z}$ with $p \equiv 3 \bmod 4$ (which have norm p^2), and
- **(c)** $a+bi$, $a-bi$, the distinct irreducible factors of $p = a^2+b^2 = (a+bi)(a-bi)$ for the primes $p \in \mathbb{Z}$ with $p \equiv 1 \bmod 4$ (both of which have norm p).

The first part of Proposition 18 is a famous theorem of Fermat in elementary number theory, for which a number of alternate proofs can be given.

More generally, the question of whether the integer $n \in \mathbb{Z}$ can be written as a sum of two integer squares, $n = A^2+B^2$, is equivalent to the question of whether n is the norm of an element $A+Bi$ in the Gaussian integers, i.e., $n = A^2+B^2 = N(A+Bi)$. Writing $A+Bi = \pi_1\pi_2\cdots\pi_k$ as a product of irreducibles (uniquely up to units) it follows from the explicit description of the irreducibles in $\mathbb{Z}[i]$ in Proposition 18 that n is a norm if and only if the prime divisors of n that are congruent to 3 mod 4 occur to even exponents. Further, if this condition on n is satisfied, then the uniqueness of the factorization of $A+Bi$ in $\mathbb{Z}[i]$ allows us to count the number of representations of n as a sum of two squares, as in the following corollary.

Corollary 19. Let n be a positive integer and write

$$n = 2^k p_1^{a_1}\cdots p_r^{a_r} q_1^{b_1}\cdots q_s^{b_s}$$

where $p_1, \ldots, p_r$ are distinct primes congruent to 1 modulo 4 and $q_1, \ldots, q_s$ are distinct primes congruent to 3 modulo 4. Then n can be written as a sum of two squares in $\mathbb{Z}$, i.e., $n = A^2+B^2$ with $A, B \in \mathbb{Z}$, if and only if each b_i is even. Further, if this condition on n is satisfied, then the number of representations of n as a sum of two squares is $4(a_1+1)(a_2+1)\cdots(a_r+1)$.

Proof: The first statement in the corollary was proved above. Assume now that $b_1, \ldots, b_s$ are all even. For each prime p_i congruent to 1 modulo 4 write $p_i = \pi_i \overline{\pi_i}$ for $i = 1, 2, \ldots, r$, where π_i and $\overline{\pi_i}$ are irreducibles as in (2)(c) of Proposition 18. If $N(A + Bi) = n$ then examining norms we see that, up to units, the factorization of $A + Bi$ into irreducibles in $\mathbb{Z}[i]$ is given by

$$A + Bi = (1 + i)^k (\pi_1^{a_{1,1}} \overline{\pi_1}^{a_{1,2}}) \ldots (\pi_r^{a_{r,1}} \overline{\pi_r}^{a_{r,2}}) q_1^{b_1/2} \ldots q_s^{b_s/2}$$

with nonnegative integers $a_{i,1}, a_{i,2}$ satisfying $a_{i,1} + a_{i,2} = a_i$ for $i = 1, 2, \ldots, r$. Since $a_{i,1}$ can have the values $0, 1, \ldots, a_i$ (and then $a_{i,2}$ is determined), there are a total of $(a_1 + 1)(a_2 + 1) \cdots (a_r + 1)$ distinct elements $A + Bi$ in $\mathbb{Z}[i]$ of norm n, up to units. Finally, since there are four units in $\mathbb{Z}[i]$, the second statement in the corollary follows.

Example

Since $493 = 17 \cdot 29$ and both primes are congruent to 1 modulo 4, $493 = A^2 + B^2$ is the sum of two integer squares. Since $17 = (4 + i)(4 - i)$ and $29 = (5 + 2i)(5 - 2i)$ the possible factorizations of $A + Bi$ in $\mathbb{Z}[i]$ up to units are $(4 + i)(5 + 2i) = 18 + 13i$, $(4 + i)(5 - 2i) = 22 - 3i$, $(4 - i)(5 - 2i) = 22 + 3i$, and $(4 - i)(5 - 2i) = 18 - 13i$. Multiplying by -1 reverses both signs and multiplication by i interchanges the A and B and introduces one sign change. Then $493 = (\pm 18)^2 + (\pm 13)^2 = (\pm 22)^2 + (\pm 3)^2$ with all possible choices of signs give 8 of the 16 possible representations of 493 as the sum of two squares; the remaining 8 are obtained by interchanging the two summands.

Similarly, the integer $58000957 = 7^6 \cdot 17 \cdot 29$ can be written as a sum of two squares in precisely 16 ways, obtained by multiplying each of the integers A, B in $493 = A^2 + B^2$ above by 7^3.

Summary

In summary, we have the following inclusions among classes of commutative rings with identity:

$$\textit{fields} \subset \textit{Euclidean Domains} \subset \textit{P.I.D.s} \subset \textit{U.F.D.s} \subset \textit{integral domains}$$

with all containments being proper. Recall that $\mathbb{Z}$ is a Euclidean Domain that is not a field, the quadratic integer ring $\mathbb{Z}[(1 + \sqrt{-19})/2]$ is a Principal Ideal Domain that is not a Euclidean Domain, $\mathbb{Z}[x]$ is a Unique Factorization Domain (Theorem 7 in Chapter 9) that is not a Principal Ideal Domain and $\mathbb{Z}[\sqrt{-5}]$ is an integral domain that is not a Unique Factorization Domain.

EXERCISES

1. Let $G = \mathbb{Q}^\times$ be the multiplicative group of nonzero rational numbers. If $\alpha = p/q \in G$, where p and q are relatively prime integers, let $\varphi : G \to G$ be the map which interchanges the primes 2 and 3 in the prime power factorizations of p and q (so, for example, $\varphi(2^4 3^{11} 5^1 13^2) = 3^4 2^{11} 5^1 13^2$, $\varphi(3/16) = \varphi(3/2^4) = 2/3^4 = 2/81$, and φ is the identity on all rational numbers with numerators and denominators relatively prime to 2 and to 3).
(a) Prove that φ is a group isomorphism.
(b) Prove that there are infinitely many isomorphisms of the group G to itself.

(c) Prove that none of the isomorphisms above can be extended to an isomorphism of the *ring* $\mathbb{Q}$ to itself. In fact prove that the identity map is the only ring isomorphism of $\mathbb{Q}$.

2. Let a and b be nonzero elements of the Unique Factorization Domain R. Prove that a and b have a least common multiple (cf. Exercise 11 of Section 1) and describe it in terms of the prime factorizations of a and b in the same fashion that Proposition 13 describes their greatest common divisor.

3. Determine all the representations of the integer $2130797 = 17^2 \cdot 73 \cdot 101$ as a sum of two squares.

4. Prove that if an integer is the sum of two rational squares, then it is the sum of two integer squares (for example, $13 = (1/5)^2 + (18/5)^2 = 2^2 + 3^2$).

5. Let $R = \mathbb{Z}[\sqrt{-n}]$ where n is a squarefree integer greater than 3.

(a) Prove that 2, $\sqrt{-n}$ and $1 + \sqrt{-n}$ are irreducibles in R.

(b) Prove that R is not a U.F.D. Conclude that the quadratic integer ring $\mathcal{O}$ is not a U.F.D. for $D \equiv 2, 3 \bmod 4$, $D < -3$ (so also not Euclidean and not a P.I.D.). [Show that either $\sqrt{-n}$ or $1 + \sqrt{-n}$ is not prime.]

(c) Give an explicit ideal in R that is not principal. [Using (b) consider a maximal ideal containing the nonprime ideal $(\sqrt{-n})$ or $(1 + \sqrt{-n})$.]

6. (a) Prove that the quotient ring $\mathbb{Z}[i]/(1 + i)$ is a field of order 2.

(b) Let $q \in \mathbb{Z}$ be a prime with $q \equiv 3 \bmod 4$. Prove that the quotient ring $\mathbb{Z}[i]/(q)$ is a field with q^2 elements.

(c) Let $p \in \mathbb{Z}$ be a prime with $p \equiv 1 \bmod 4$ and write $p = \pi\bar{\pi}$ as in Proposition 18. Show that the hypotheses for the Chinese Remainder Theorem (Theorem 17 in Section 7.6) are satisfied and that $\mathbb{Z}[i]/(p) \cong \mathbb{Z}[i]/(\pi) \times \mathbb{Z}[i]/(\bar{\pi})$ as rings. Show that the quotient ring $\mathbb{Z}[i]/(p)$ has order p^2 and conclude that $\mathbb{Z}[i]/(\pi)$ and $\mathbb{Z}[i]/(\bar{\pi})$ are both fields of order p.

7. Let π be an irreducible element in $\mathbb{Z}[i]$.

(a) For any integer $n \geq 0$, prove that $(\pi^{n+1}) = \pi^{n+1}\mathbb{Z}[i]$ is an ideal in $(\pi^n) = \pi^n\mathbb{Z}[i]$ and that multiplication by π^n induces an isomorphism $\mathbb{Z}[i]/(\pi) \cong (\pi^n)/(\pi^{n+1})$ as additive abelian groups.

(b) Prove that $|\mathbb{Z}[i]/(\pi^n)| = |\mathbb{Z}[i]/(\pi)|^n$.

(c) Prove for any nonzero α in $\mathbb{Z}[i]$ that the quotient ring $\mathbb{Z}[i]/(\alpha)$ has order equal to $N(\alpha)$. [Use (b) together with the Chinese Remainder Theorem and the results of the previous exercise.]

8. Let R be the quadratic integer ring $\mathbb{Z}[\sqrt{-5}]$ and define the ideals $I_2 = (2, 1 + \sqrt{-5})$, $I_3 = (3, 2 + \sqrt{-5})$, and $I_3' = (3, 2 - \sqrt{-5})$.

(a) Prove that 2, 3, $1 + \sqrt{-5}$ and $1 - \sqrt{-5}$ are irreducibles in R, no two of which are associate in R, and that $6 = 2 \cdot 3 = (1 + \sqrt{-5}) \cdot (1 - \sqrt{-5})$ are two distinct factorizations of 6 into irreducibles in R.

(b) Prove that I_2, I_3, and I_3' are prime ideals in R. [One approach: for I_3, observe that $R/I_3 \cong (R/(3))/(I_3/(3))$ by the Third Isomorphism Theorem for Rings. Show that $R/(3)$ has 9 elements, $(I_3/(3))$ has 3 elements, and that $R/I_3 \cong \mathbb{Z}/3\mathbb{Z}$ as an additive abelian group. Conclude that I_3 is a maximal (hence prime) ideal and that $R/I_3 \cong \mathbb{Z}/3\mathbb{Z}$ as rings.]

(c) Show that the factorizations in (a) imply the equality of ideals $(6) = (2)(3)$ and $(6) = (1 + \sqrt{-5})(1 - \sqrt{-5})$. Show that these two ideal factorizations give the same factorization of the ideal (6) as the product of prime ideals (cf. Exercise 5 in Section 2).

9. Suppose that the quadratic integer ring $\mathcal{O}$ is a P.I.D. Prove that the absolute value of the field norm N on $\mathcal{O}$ (cf. Section 7.1) is a Dedekind–Hasse norm on $\mathcal{O}$. Conclude that if the quadratic integer ring $\mathcal{O}$ possesses *any* Dedekind–Hasse norm, then in fact the absolute value of the field norm on $\mathcal{O}$ already provides a Dedekind–Hasse norm on $\mathcal{O}$. [If $\alpha, \beta \in \mathcal{O}$ then $(\alpha, \beta) = (\gamma)$ for some $\gamma \in \mathcal{O}$. Show that if β does not divide α then $0 < |N(\gamma)| < |N(\beta)|$ — use the fact that the units in $\mathcal{O}$ are precisely the elements whose norm is ± 1.]

Remark: If $\mathcal{O}$ is a Euclidean Domain with respect to some norm it is not necessarily true that it is a Euclidean Domain with respect to the absolute value of the field norm (although this is true for $D < 0$, cf. Exercise 8 in Section 1). An example is $D = 69$ (cf. D. Clark, *A quadratic field which is Euclidean but not norm-Euclidean*, Manuscripta Math., 83(1994), pp. 327–330).

10. (*k-stage Euclidean Domains*) Let R be an integral domain and let $N : R \to \mathbb{Z}^+ \cup \{0\}$ be a norm on R. The ring R is Euclidean with respect to N if for any $a, b \in R$ with $b \neq 0$, there exist elements q and r in R with

$$a = qb + r \qquad \text{with } r = 0 \text{ or } N(r) < N(b).$$

Suppose now that this condition is weakened, namely that for any $a, b \in R$ with $b \neq 0$, there exist elements q, q' and r, r' in R with

$$a = qb + r, \quad b = q'r + r' \qquad \text{with } r' = 0 \text{ or } N(r') < N(b),$$

i.e., the remainder after two divisions is smaller. Call such a domain a *2-stage Euclidean Domain.*

(a) Prove that iterating the divisions in a 2-stage Euclidean Domain produces a greatest common divisor of a and b which is a linear combination of a and b. Conclude that every *finitely generated* ideal of a 2-stage Euclidean Domain is principal. (There are 2-stage Euclidean Domains that are *not* P.I.D.s, however.) [Imitate the proof of Theorem 4.]

(b) Prove that a 2-stage Euclidean Domain in which every nonzero nonunit can be factored into a finite number of irreducibles is a Unique Factorization Domain. [Prove first that irreducible elements are prime, as follows. If p is irreducible and $p \mid ab$ with p not dividing a, use part (a) to write $px + ay = 1$ for some x, y. Multiply through by b to conclude that $p \mid b$, so p is prime. Now follow the proof of uniqueness in Theorem 14.]

(c) Make the obvious generalization to define the notion of a k-stage Euclidean Domain for any integer $k \geq 1$. Prove that statements (a) and (b) remain valid if "2-stage Euclidean" is replaced by "k-stage Euclidean."

Remarks: There are examples of rings which are 2-stage Euclidean but are not Euclidean. There are also examples of rings which are not Euclidean with respect to a given norm but which are k-stage Euclidean with respect to the norm (for example, the ring $\mathbb{Z}[\sqrt{14}]$ is not Euclidean with respect to the usual norm $N(a+b\sqrt{14}) = |a^2 - 14b^2|$, but is 2-stage Euclidean with respect to this norm). The k-stage Euclidean condition is also related to the question of whether the group $GL_n(R)$ of invertible $n \times n$ matrices with entries from R is generated by elementary matrices (matrices with 1's along the main diagonal, a single 1 somewhere off the main diagonal, and 0's elsewhere).

11. (*Characterization of P.I.D.s*) Prove that R is a P.I.D. if and only if R is a U.F.D. that is also a Bezout Domain (cf. Exercise 7 in Section 2). [One direction is given by Theorem 14. For the converse, let a be a nonzero element of the ideal I with a minimal number of irreducible factors. Prove that $I = (a)$ by showing that if there is an element $b \in I$ that is not in (a) then $(a, b) = (d)$ leads to a contradiction.]

CHAPTER 9

Polynomial Rings

We begin this chapter on polynomial rings with a summary of facts from the preceding two chapters (with references where needed). The basic definitions were given in slightly greater detail in Section 7.2. For convenience, the ring R will always be a commutative ring with identity $1 \neq 0$.

9.1 DEFINITIONS AND BASIC PROPERTIES

The polynomial ring $R[x]$ in the indeterminate x with coefficients from R is the set of all formal sums $a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$ with $n \geq 0$ and each $a_i \in R$. If $a_n \neq 0$ then the polynomial is of degree n, $a_n x^n$ is the leading term, and a_n is the leading coefficient (where the leading coefficient of the zero polynomial is defined to be 0). The polynomial is monic if $a_n = 1$. Addition of polynomials is "componentwise":

$$\sum_{i=0}^{n} a_i x^i + \sum_{i=0}^{n} b_i x^i = \sum_{i=0}^{n} (a_i + b_i) x^i$$

(here a_n or b_n may be zero in order for addition of polynomials of different degrees to be defined). Multiplication is performed by first defining $(ax^i)(bx^j) = abx^{i+j}$ and then extending to all polynomials by the distributive laws so that in general

$$\left(\sum_{i=0}^{n} a_i x^i\right) \times \left(\sum_{i=0}^{m} b_i x^i\right) = \sum_{k=0}^{n+m} \left(\sum_{i=0}^{k} a_i b_{k-i}\right) x^k.$$

In this way $R[x]$ is a commutative ring with identity (the identity 1 from R) in which we identify R with the subring of constant polynomials.

We have already noted that if R is an integral domain then the leading term of a product of polynomials is the product of the leading terms of the factors. The following is Proposition 4 of Section 7.2 which we record here for completeness.

Proposition 1. Let R be an integral domain. Then

(1) degree $p(x)q(x)$ = degree $p(x)$ + degree $q(x)$ if $p(x), q(x)$ are nonzero

(2) the units of $R[x]$ are just the units of R

(3) $R[x]$ is an integral domain.

Recall also that if R is an integral domain, the quotient field of $R[x]$ consists of all

quotients $\frac{p(x)}{q(x)}$ where $q(x)$ is not the zero polynomial (and is called the field of rational functions in x with coefficients in R).

The next result describes a relation between the ideals of R and those of $R[x]$.

Proposition 2. Let I be an ideal of the ring R and let $(I) = I[x]$ denote the ideal of $R[x]$ generated by I (the set of polynomials with coefficients in I). Then

$$R[x]/(I) \cong (R/I)[x].$$

In particular, if I is a prime ideal of R then (I) is a prime ideal of $R[x]$.

Proof: There is a natural map $\varphi : R[x] \to (R/I)[x]$ given by reducing each of the coefficients of a polynomial modulo I. The definition of addition and multiplication in these two rings shows that φ is a ring homomorphism. The kernel is precisely the set of polynomials each of whose coefficients is an element of I, which is to say that $\ker \varphi = I[x] = (I)$, proving the first part of the proposition. The last statement follows from Proposition 1, since if I is a prime ideal in R, then R/I is an integral domain, hence also $(R/I)[x]$ is an integral domain. This shows if I is a prime ideal of R, then (I) is a prime ideal of $R[x]$.

Note that it is not true that if I is a maximal ideal of R then (I) is a maximal ideal of $R[x]$. However, if I is maximal in R then the ideal of $R[x]$ generated by I and x is maximal in $R[x]$.

We now give an example of the "reduction homomorphism" of Proposition 2 which will be useful on a number of occasions later ("reduction homomorphisms" were also discussed at the end of Section 7.3 with reference to reducing the integers mod n) .

Example

Let $R = \mathbb{Z}$ and consider the ideal $n\mathbb{Z}$ of $\mathbb{Z}$. Then the isomorphism above can be written

$$\mathbb{Z}[x]/n\mathbb{Z}[x] \cong \mathbb{Z}/n\mathbb{Z}[x]$$

and the natural projection map of $\mathbb{Z}[x]$ to $\mathbb{Z}/n\mathbb{Z}[x]$ by reducing the coefficients modulo n is a ring homomorphism. If n is composite, then the quotient ring is not an integral domain. If, however, n is a prime p, then $\mathbb{Z}/p\mathbb{Z}$ is a field and so $\mathbb{Z}/p\mathbb{Z}[x]$ is an integral domain (in fact, a Euclidean Domain, as we shall see shortly). We also see that the set of polynomials whose coefficients are divisible by p is a prime ideal in $\mathbb{Z}[x]$.

We close this section with a description of the natural extension to polynomial rings in *several* variables.

Definition. The *polynomial ring in the variables* $x_1, x_2, \ldots, x_n$ *with coefficients in* R, denoted $R[x_1, x_2, \ldots, x_n]$, is defined inductively by

$$R[x_1, x_2, \ldots, x_n] = R[x_1, x_2, \ldots, x_{n-1}][x_n]$$

This definition means that we can consider polynomials in n variables with coefficients in R simply as polynomials in *one* variable (say x_n) but now with coefficients that

are themselves *polynomials in $n-1$ variables*. In a slightly more concrete formulation, a nonzero polynomial in $x_1, x_2, \ldots, x_n$ with coefficients in R is a finite sum of nonzero *monomial terms*, i.e., a finite sum of elements of the form

$$ax_1^{d_1}x_2^{d_2}\ldots x_n^{d_n}$$

where $a \in R$ (the *coefficient* of the term) and the d_i are nonnegative integers. A monic term $x_1^{d_1}x_2^{d_2}\ldots x_n^{d_n}$ is called simply a *monomial* and is the *monomial part* of the term $ax_1^{d_1}x_2^{d_2}\ldots x_n^{d_n}$. The exponent d_i is called the *degree in* x_i of the term and the sum

$$d = d_1 + d_2 + \cdots + d_n$$

is called the *degree* of the term. The ordered n-tuple $(d_1, d_2, \ldots, d_n)$ is the *multidegree* of the term. The *degree* of a nonzero polynomial is the largest degree of any of its monomial terms. A polynomial is called *homogeneous* or a *form* if all its terms have the same degree. If f is a nonzero polynomial in n variables, the sum of all the monomial terms in f of degree k is called the *homogeneous component of f of degree k*. If f has degree d then f may be written uniquely as the sum $f_0 + f_1 + \cdots + f_d$ where f_k is the homogeneous component of f of degree k, for $0 \leq k \leq d$ (where some f_k may be zero).

Finally, to define a polynomial ring in an *arbitrary* number of variables with coefficients in R we take finite sums of monomial terms of the type above (but where the variables are not restricted to just $x_1, \ldots, x_n$), with the natural addition and multiplication. Alternatively, we could define this ring as the *union* of *all* the polynomial rings in a *finite* number of the variables being considered.

Example

The polynomial ring $\mathbb{Z}[x, y]$ in two variables x and y with integer coefficients consists of all finite sums of monomial terms of the form ax^iy^j (of degree $i + j$). For example,

$$p(x, y) = 2x^3 + xy - y^2$$

and

$$q(x, y) = -3xy + 2y^2 + x^2y^3$$

are both elements of $\mathbb{Z}[x, y]$, of degrees 3 and 5, respectively. We have

$$p(x, y) + q(x, y) = 2x^3 - 2xy + y^2 + x^2y^3$$

and

$$p(x, y)q(x, y) = -6x^4y + 4x^3y^2 + 2x^5y^3 - 3x^2y^2 + 5xy^3 + x^3y^4 - 2y^4 - x^2y^5,$$

a polynomial of degree 8. To view this last polynomial, say, as a polynomial in y with coefficients in $\mathbb{Z}[x]$ as in the definition of several variable polynomial rings above, we would write the polynomial in the form

$$(-6x^4)y + (4x^3 - 3x^2)y^2 + (2x^5 + 5x)y^3 + (x^3 - 2)y^4 - (x^2)y^5.$$

The nonzero homogeneous components of $f = f(x, y) = p(x, y)q(x, y)$ are the polynomials $f_4 = -3x^2y^2 + 5xy^3 - 2y^4$ (degree 4), $f_5 = -6x^4y + 4x^3y^2$ (degree 5), $f_7 = x^3y^4 - x^2y^5$ (degree 7), and $f_8 = 2x^5y^3$ (degree 8).

Each of the statements in Proposition 1 is true for polynomial rings with an arbitrary number of variables. This follows by induction for finitely many variables and from the definition in terms of unions in the case of polynomial rings in arbitrarily many variables.

EXERCISES

1. Let $p(x, y, z) = 2x^2y - 3xy^3z + 4y^2z^5$ and $q(x, y, z) = 7x^2 + 5x^2y^3z^4 - 3x^2z^3$ be polynomials in $\mathbb{Z}[x, y, z]$.
   (a) Write each of p and q as a polynomial in x with coefficients in $\mathbb{Z}[y, z]$.
   (b) Find the degree of each of p and q.
   (c) Find the degree of p and q in each of the three variables x, y and z.
   (d) Compute pq and find the degree of pq in each of the three variables x, y and z.
   (e) Write pq as a polynomial in the variable z with coefficients in $\mathbb{Z}[x, y]$.
2. Repeat the preceding exercise under the assumption that the coefficients of p and q are in $\mathbb{Z}/3\mathbb{Z}$.
3. If R is a commutative ring and $x_1, x_2, \ldots, x_n$ are independent variables over R, prove that $R[x_{\pi(1)}, x_{\pi(2)}, \ldots, x_{\pi(n)}]$ is isomorphic to $R[x_1, x_2, \ldots, x_n]$ for any permutation π of $\{1, 2, \ldots, n\}$.
4. Prove that the ideals (x) and (x, y) are prime ideals in $\mathbb{Q}[x, y]$ but only the latter ideal is a maximal ideal.
5. Prove that (x, y) and $(2, x, y)$ are prime ideals in $\mathbb{Z}[x, y]$ but only the latter ideal is a maximal ideal.
6. Prove that (x, y) is not a principal ideal in $\mathbb{Q}[x, y]$.
7. Let R be a commutative ring with 1. Prove that a polynomial ring in more than one variable over R is not a Principal Ideal Domain.
8. Let F be a field and let $R = F[x, x^2y, x^3y^2, \ldots, x^ny^{n-1}, \ldots]$ be a subring of the polynomial ring $F[x, y]$.
   (a) Prove that the fields of fractions of R and $F[x, y]$ are the same.
   (b) Prove that R contains an ideal that is not finitely generated.
9. Prove that a polynomial ring in infinitely many variables with coefficients in any commutative ring contains ideals that are not finitely generated.
10. Prove that the ring $\mathbb{Z}[x_1, x_2, x_3, \ldots]/(x_1x_2, x_3x_4, x_5x_6, \ldots)$ contains infinitely many minimal prime ideals (cf. Exercise 36 of Section 7.4).
11. Show that the radical of the ideal $I = (x, y^2)$ in $\mathbb{Q}[x, y]$ is (x, y) (cf. Exercise 30, Section 7.4). Deduce that I is a primary ideal that is not a power of a prime ideal (cf. Exercise 41, Section 7.4).
12. Let $R = \mathbb{Q}[x, y, z]$ and let bars denote passage to $\mathbb{Q}[x, y, z]/(xy - z^2)$. Prove that $\overline{P} = (\overline{x}, \overline{z})$ is a prime ideal. Show that $\overline{xy} \in \overline{P}^2$ but that no power of $\overline{y}$ lies in $\overline{P}^2$. (This shows $\overline{P}$ is a prime ideal whose square is *not* a primary ideal — cf. Exercise 41, Section 7.4).
13. Prove that the rings $F[x, y]/(y^2 - x)$ and $F[x, y]/(y^2 - x^2)$ are not isomorphic for any field F.
14. Let R be an integral domain and let i, j be relatively prime integers. Prove that the ideal $(x^i - y^j)$ is a prime ideal in $R[x, y]$. [Consider the ring homomorphism φ from $R[x, y]$ to $R[t]$ defined by mapping x to t^j and mapping y to t^i. Show that an element of $R[x, y]$

differs from an element in $(x^i - y^j)$ by a polynomial $f(x)$ of degree at most $j-1$ in y and observe that the exponents of $\varphi(x^r y^s)$ are distinct for $0 \leq s < j$.]

15. Let $p(x_1, x_2, \ldots, x_n)$ be a homogeneous polynomial of degree k in $R[x_1, \ldots, x_n]$. Prove that for all $\lambda \in R$ we have $p(\lambda x_1, \lambda x_2, \ldots, \lambda x_n) = \lambda^k p(x_1, x_2, \ldots, x_n)$.

16. Prove that the product of two homogeneous polynomials is again homogeneous.

17. An ideal I in $R[x_1, \ldots, x_n]$ is called a *homogeneous ideal* if whenever $p \in I$ then each homogeneous component of p is also in I. Prove that an ideal is a homogeneous ideal if and only if it may be generated by homogeneous polynomials. [Use induction on degrees to show the "if" implication.]

The following exercise shows that some care must be taken when working with polynomials over noncommutative rings R (the ring operations in $R[x]$ are defined in the same way as for commutative rings R), in particular when considering polynomials as functions.

18. Let R be an arbitrary ring and let Func(R) be the ring of all functions from R to itself. If $p(x) \in R[x]$ is a polynomial, let $f_p \in$ Func(R) be the function on R defined by $f_p(r) = p(r)$ (the usual way of viewing a polynomial in $R[x]$ as defining a function on R by "evaluating at r").

(a) For fixed $a \in R$, prove that "evaluation at a" is a ring homomorphism from Func(R) to R (cf. Example 4 following Theorem 7 in Section 7.3).

(b) Prove that the map $\varphi : R[x] \to$ Func(R) defined by $\varphi(p(x)) = f_p$ is not a ring homomorphism in general. Deduce that polynomial identities need not give corresponding identities when the polynomials are viewed as functions. [If $R = \mathbb{H}$ is the ring of real Hamilton Quaternions show that $p(x) = x^2 + 1$ factors as $(x+i)(x-i)$, but that $p(j) = 0$ while $(j+i)(j-i) \neq 0$.]

(c) For fixed $a \in R$, prove that the composite "evaluation at a" of the maps in (a) and (b) mapping $R[x]$ to R is a ring homomorphism if and only if a is in the center of R.

9.2 POLYNOMIAL RINGS OVER FIELDS I

We now consider more carefully the situation where the coefficient ring is a *field* F. We can define a *norm* on $F[x]$ by defining $N(p(x)) =$ degree of $p(x)$ (where we set $N(0) = 0$). From elementary algebra we know that we can divide one polynomial with, say, rational coefficients by another (nonzero) polynomial with rational coefficients to obtain a quotient and remainder. The same is true over any field.

Theorem 3. Let F be a field. The polynomial ring $F[x]$ is a Euclidean Domain. Specifically, if $a(x)$ and $b(x)$ are two polynomials in $F[x]$ with $b(x)$ nonzero, then there are *unique* $q(x)$ and $r(x)$ in $F[x]$ such that

$$a(x) = q(x)b(x) + r(x) \qquad \text{with } r(x) = 0 \text{ or degree } r(x) < \text{degree } b(x) .$$

Proof: If $a(x)$ is the zero polynomial then take $q(x) = r(x) = 0$. We may therefore assume $a(x) \neq 0$ and prove the existence of $q(x)$ and $r(x)$ by induction on $n =$ degree $a(x)$. Let $b(x)$ have degree m. If $n < m$ take $q(x) = 0$ and $r(x) = a(x)$. Otherwise $n \geq m$. Write

$$a(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0$$

and

$$b(x) = b_m x^m + b_{m-1}x^{m-1} + \cdots + b_1 x + b_0.$$

Then the polynomial $a'(x) = a(x) - \dfrac{a_n}{b_m}x^{n-m}b(x)$ is of degree less than n (we have arranged to subtract the leading term from $a(x)$). Note that this polynomial is well defined because the coefficients are taken from a *field* and $b_m \neq 0$. By induction then, there exist polynomials $q'(x)$ and $r(x)$ with

$$a'(x) = q'(x)b(x) + r(x) \qquad \text{with } r(x) = 0 \text{ or degree } r(x) < \text{degree } b(x).$$

Then, letting $q(x) = q'(x) + \dfrac{a_n}{b_m}x^{n-m}$ we have

$$a(x) = q(x)b(x) + r(x) \qquad \text{with } r(x) = 0 \text{ or degree } r(x) < \text{degree } b(x)$$

completing the induction step.

As for the uniqueness, suppose $q_1(x)$ and $r_1(x)$ also satisfied the conditions of the theorem. Then both $a(x) - q(x)b(x)$ and $a(x) - q_1(x)b(x)$ are of degree less than m = degree $b(x)$. The difference of these two polynomials, i.e., $b(x)(q(x) - q_1(x))$ is also of degree less than m. But the degree of the product of two nonzero polynomials is the sum of their degrees (since F is an integral domain), hence $q(x) - q_1(x)$ must be 0, that is, $q(x) = q_1(x)$. This implies $r(x) = r_1(x)$, completing the proof.

Corollary 4. If F is a field, then $F[x]$ is a Principal Ideal Domain and a Unique Factorization Domain.

Proof: This is immediate from the results of the last chapter.

Recall also from Corollary 8 in Section 8.2 that if R is any commutative ring such that $R[x]$ is a Principal Ideal Domain (or Euclidean Domain) then R must be a field. We shall see in the next section, however, that $R[x]$ is a Unique Factorization Domain whenever R itself is a Unique Factorization Domain.

Examples

(1) By the above remarks the ring $\mathbb{Z}[x]$ is not a Principal Ideal Domain. As we have already seen (Example 3 beginning of Section 7.4) the ideal $(2, x)$ is not principal in this ring.

(2) $\mathbb{Q}[x]$ is a Principal Ideal Domain since the coefficients lie in the field $\mathbb{Q}$. The ideal generated in $\mathbb{Z}[x]$ by 2 and x is not principal in the subring $\mathbb{Z}[x]$ of $\mathbb{Q}[x]$. However, the ideal generated in $\mathbb{Q}[x]$ is principal; in fact it is the entire ring (so has 1 as a generator) since 2 is a unit in $\mathbb{Q}[x]$.

(3) If p is a prime, the ring $\mathbb{Z}/p\mathbb{Z}[x]$ obtained by reducing $\mathbb{Z}[x]$ modulo the prime ideal (p) is a Principal Ideal Domain, since the coefficients lie in the field $\mathbb{Z}/p\mathbb{Z}$. This example shows that the quotient of a ring which is not a Principal Ideal Domain *may* be a Principal Ideal Domain. To follow the ideal $(2, x)$ above in this example, note that if $p = 2$, then the ideal $(2, x)$ reduces to the ideal (x) in the quotient $\mathbb{Z}/2\mathbb{Z}[x]$, which is a proper (maximal) ideal. If $p \neq 2$, then 2 is a unit in the quotient, so the ideal $(2, x)$ reduces to the entire ring $\mathbb{Z}/p\mathbb{Z}[x]$.

(4) $\mathbb{Q}[x, y]$, the ring of polynomials in two variables with rational coefficients, is *not* a Principal Ideal Domain since this ring is $\mathbb{Q}[x][y]$ and $\mathbb{Q}[x]$ is not a field (any element

of positive degree is not invertible). It is an exercise to see that the ideal (x, y) is not a principal ideal in this ring. We shall see shortly that $\mathbb{Q}[x, y]$ *is* a Unique Factorization Domain.

We note that the quotient and remainder in the Division Algorithm applied to $a(x), b(x) \in F[x]$ are *independent of field extensions* in the following sense. Suppose the field F is contained in the field E and $a(x) = Q(x)b(x) + R(x)$ for some $Q(x)$, $R(x)$ satisfying the conditions of Theorem 3 in $E[x]$. Write $a(x) = q(x)b(x)+r(x)$ for some $q(x), r(x) \in F[x]$ and apply the uniqueness condition of Theorem 3 in the ring $E[x]$ to deduce that $Q(x) = q(x)$ and $R(x) = r(x)$. In particular, $b(x)$ divides $a(x)$ in the ring $E[x]$ if and only if $b(x)$ divides $a(x)$ in $F[x]$. Also, the greatest common divisor of $a(x)$ and $b(x)$ (which can be obtained from the Euclidean Algorithm) is the same, once we make it unique by specifying it to be monic, whether these elements are viewed in $F[x]$ or in $E[x]$.

EXERCISES

Let F be a field and let x be an indeterminate over F.

1. Let $f(x) \in F[x]$ be a polynomial of degree $n \geq 1$ and let bars denote passage to the quotient $F[x]/(f(x))$. Prove that for each $\overline{g(x)}$ there is a unique polynomial $g_0(x)$ of degree $\leq n - 1$ such that $\overline{g(x)} = \overline{g_0(x)}$ (equivalently, the elements $\overline{1}, \overline{x}, \ldots, \overline{x^{n-1}}$ are a *basis* of the vector space $F[x]/(f(x))$ over F — in particular, the dimension of this space is n). [Use the Division Algorithm.]
2. Let F be a finite field of order q and let $f(x)$ be a polynomial in $F[x]$ of degree $n \geq 1$. Prove that $F[x]/(f(x))$ has q^n elements. [Use the preceding exercise.]
3. Let $f(x)$ be a polynomial in $F[x]$. Prove that $F[x]/(f(x))$ is a field if and only if $f(x)$ is irreducible. [Use Proposition 7, Section 8.2.]
4. Let F be a finite field. Prove that $F[x]$ contains infinitely many primes. (Note that over an infinite field the polynomials of degree 1 are an infinite set of primes in the ring of polynomials).
5. Exhibit *all* the ideals in the ring $F[x]/(p(x))$, where F is a field and $p(x)$ is a polynomial in $F[x]$ (describe them in terms of the factorization of $p(x)$).
6. Describe (briefly) the ring structure of the following rings:
   (a) $\mathbb{Z}[x]/(2)$, **(b)** $\mathbb{Z}[x]/(x)$, **(c)** $\mathbb{Z}[x]/(x^2)$, **(d)** $\mathbb{Z}[x, y]/(x^2, y^2, 2)$.
   Show that $\alpha^2 = 0$ or 1 for every α in the last ring and determine those elements with $\alpha^2 = 0$. Determine the characteristics of each of these rings (cf. Exercise 26, Section 7.3).
7. Determine all the ideals of the ring $\mathbb{Z}[x]/(2, x^3 + 1)$.
8. Determine the greatest common divisor of $a(x) = x^3 - 2$ and $b(x) = x + 1$ in $\mathbb{Q}[x]$ and write it as a linear combination (in $\mathbb{Q}[x]$) of $a(x)$ and $b(x)$.
9. Determine the greatest common divisor of $a(x) = x^5+2x^3+x^2+x+1$ and the polynomial $b(x) = x^5 + x^4 + 2x^3 + 2x^2 + 2x + 1$ in $\mathbb{Q}[x]$ and write it as a linear combination (in $\mathbb{Q}[x]$) of $a(x)$ and $b(x)$.
10. Determine the greatest common divisor of $a(x) = x^3+4x^2+x-6$ and $b(x) = x^5-6x+5$ in $\mathbb{Q}[x]$ and write it as a linear combination (in $\mathbb{Q}[x]$) of $a(x)$ and $b(x)$.
11. Suppose $f(x)$ and $g(x)$ are two nonzero polynomials in $\mathbb{Q}[x]$ with greatest common divisor $d(x)$.

(a) Given $h(x) \in \mathbb{Q}[x]$, show that there are polynomials $a(x), b(x) \in \mathbb{Q}[x]$ satisfying the equation $a(x)f(x) + b(x)g(x) = h(x)$ if and only if $h(x)$ is divisible by $d(x)$.

(b) If $a_0(x), b_0(x) \in \mathbb{Q}[x]$ are particular solutions to the equation in (a), show that the full set of solutions to this equation is given by

$$a(x) = a_0(x) + m(x)\frac{g(x)}{(x)d}$$
$$b(x) = b_0(x) - m(x)\frac{f(x)}{d(x)}$$

as $m(x)$ ranges over the polynomials in $\mathbb{Q}[x]$. [cf. Exercise 4 in Section 8.1]

12. Let $F[x, y_1, y_2, \ldots]$ be the polynomial ring in the infinite set of variables $x, y_1, y_2, \ldots$ over the field F, and let I be the ideal $(x - y_1^2, y_1 - y_2^2, \ldots, y_i - y_{i+1}^2, \ldots)$ in this ring. Define R to be the ring $F[x, y_1, y_2, \ldots]/I$, so that in R the square of each y_{i+1} is y_i and $y_1^2 = x$ modulo I, i.e., x has a 2^ith root, for every i. Denote the image of y_i in R as $x^{1/2^i}$. Let R_n be the subring of R generated by F and $x^{1/2^n}$.

(a) Prove that $R_1 \subseteq R_2 \subseteq \cdots$ and that R is the union of all R_n, i.e., $R = \cup_{n=1}^{\infty} R_n$.

(b) Prove that R_n is isomorphic to a polynomial ring in one variable over F, so that R_n is a P.I.D. Deduce that R is a Bezout Domain (cf. Exercise 7 in Section 8.2). [First show that the ring $S_n = F[x, y_1, \ldots, y_n]/(x - y_1^2, y_1 - y_2^2, \ldots, y_{n-1} - y_n^2)$ is isomorphic to the polynomial ring $F[y_n]$. Then show any polynomial relation y_n satisfies in R_n gives a corresponding relation in S_N for some $N \geq n$.]

(c) Prove that the ideal generated by $x, x^{1/2}, x^{1/4}, \ldots$ in R is not finitely generated (so R is not a P.I.D.).

13. This exercise introduces a noncommutative ring which is a "right" Euclidean Domain (and a "left" Principal Ideal Domain) but is not a "left" Euclidean Domain (and not a "right" Principal Ideal Domain). Let F be a field of characteristic p in which not every element is a p^{th} power: $F \neq F^p$ (for example the field $F = \mathbb{F}_p(t)$ of rational functions in the variable t with coefficients in $\mathbb{F}_p$ is such a field). Let $R = F\{x\}$ be the "twisted" polynomial ring of polynomials $\sum_{i=0}^{n} a_i x^i$ in x with coefficients in F with the usual (termwise) addition

$$\sum_{i=0}^{n} a_i x^i + \sum_{i=0}^{n} b_i x^i = \sum_{i=0}^{n} (a_i + b_i) x^i$$

but with a noncommutative multiplication defined by

$$\left(\sum_{i=0}^{n} a_i x^i\right)\left(\sum_{j=0}^{m} b_j x^j\right) = \sum_{k=0}^{n+m}\left(\sum_{i+j=k} a_i b_j^{p^i}\right) x^k .$$

This multiplication arises from defining $xa = a^p x$ for every $a \in F$ (so the powers of x do not commute with the coefficients) and extending in a natural way. Let N be the norm defined by taking the degree of a polynomial in R: $N(f) = \deg(f)$.

(a) Show that $x^k a = a^{p^k} x^k$ for every $a \in F$ and every integer $k \geq 0$ and that R is a ring with this definition of multiplication. [Use the fact that $(a + b)^p = a^p + b^p$ for every $a, b \in F$ since F has characteristic p, so also $(a + b)^{p^k} = a^{p^k} + b^{p^k}$ for every $a, b \in F$.]

(b) Prove that the degree of a product of two elements of R is the sum of the degrees of the elements. Prove that R has no zero divisors.

(c) Prove that R is "right Euclidean" with respect to N, i.e., for any polynomials $f, g \in R$ with $g \neq 0$, there exist polynomials q and r in R with

$$f = qg + r \qquad \text{with } r = 0 \text{ or } \deg(r) < \deg(g).$$

Use this to prove that every *left* ideal of R is principal.

(d) Let $f = \theta x$ for some $\theta \in F$, $\theta \notin F^p$ and let $g = x$. Prove that there are no polynomials q and r in R with

$$f = gq + r \qquad \text{with } r = 0 \text{ or } \deg(r) < \deg(g),$$

so in particular R is not "left Euclidean" with respect to N. Prove that the right ideal of R generated by x and θx is not principal. Conclude that R is not "left Euclidean" with respect to *any* norm.

9.3 POLYNOMIAL RINGS THAT ARE UNIQUE FACTORIZATION DOMAINS

We have seen in Proposition 1 that if R is an integral domain then $R[x]$ is also an integral domain. Also, such an R can be embedded in its field of fractions F (Theorem 15, Section 7.5), so that $R[x] \subseteq F[x]$ is a subring, and $F[x]$ is a Euclidean Domain (hence a Principal Ideal Domain and a Unique Factorization Domain). Many computations for $R[x]$ may be accomplished in $F[x]$ at the expense of allowing fractional coefficients. This raises the immediate question of how computations (such as factorizations of polynomials) in $F[x]$ can be used to give information in $R[x]$.

For instance, suppose $p(x)$ is a polynomial in $R[x]$. Since $F[x]$ is a Unique Factorization Domain we can factor $p(x)$ uniquely into a product of irreducibles in $F[x]$. It is natural to ask whether we can do the same in $R[x]$, i.e., is $R[x]$ a Unique Factorization Domain? In general the answer is no because if $R[x]$ were a Unique Factorization Domain, the constant polynomials would have to be uniquely factored into irreducible elements of $R[x]$, necessarily of degree 0 since the degrees of products add, that is, R would itself have to be a Unique Factorization Domain. Thus if R is an integral domain which is not a Unique Factorization Domain, $R[x]$ cannot be a Unique Factorization Domain. On the other hand, it turns out that if R is a Unique Factorization Domain, then $R[x]$ is also a Unique Factorization Domain. The method of proving this is to first factor uniquely in $F[x]$ and then "clear denominators" to obtain a unique factorization in $R[x]$. The first step in making this precise is to compare the factorization of a polynomial in $F[x]$ to a factorization in $R[x]$.

Proposition 5. *(Gauss' Lemma)* Let R be a Unique Factorization Domain with field of fractions F and let $p(x) \in R[x]$. If $p(x)$ is reducible in $F[x]$ then $p(x)$ is reducible in $R[x]$. More precisely, if $p(x) = A(x)B(x)$ for some nonconstant polynomials $A(x), B(x) \in F[x]$, then there are nonzero elements $r, s \in F$ such that $rA(x) = a(x)$ and $sB(x) = b(x)$ both lie in $R[x]$ and $p(x) = a(x)b(x)$ is a factorization in $R[x]$.

Proof: The coefficients of the polynomials on the right hand side of the equation $p(x) = A(x)B(x)$ are elements in the field F, hence are quotients of elements from the Unique Factorization Domain R. Multiplying through by a common denominator

for all these coefficients, we obtain an equation $dp(x) = a'(x)b'(x)$ where now $a'(x)$ and $b'(x)$ are elements of $R[x]$ and d is a nonzero element of R. If d is a unit in R, the proposition is true with $a(x) = d^{-1}a'(x)$ and $b(x) = b'(x)$. Assume d is not a unit and write d as a product of irreducibles in R, say $d = p_1 \cdots p_n$. Since p_1 is irreducible in R, the ideal (p_1) is prime (cf. Proposition 12, Section 8.3), so by Proposition 2 above, the ideal $p_1R[x]$ is prime in $R[x]$ and $(R/p_1R)[x]$ is an integral domain. Reducing the equation $dp(x) = a'(x)b'(x)$ modulo p_1, we obtain the equation $0 = \overline{a'(x)}\,\overline{b'(x)}$ in this integral domain (the bars denote the images of these polynomials in the quotient ring), hence one of the two factors, say $\overline{a'(x)}$ must be 0. But this means all the coefficients of $a'(x)$ are divisible by p_1, so that $\frac{1}{p_1}a'(x)$ also has coefficients in R. In other words, in the equation $dp(x) = a'(x)b'(x)$ we can cancel a factor of p_1 from d (on the left) and from either $a'(x)$ or $b'(x)$ (on the right) and still have an equation in $R[x]$. But now the factor d on the left hand side has one fewer irreducible factors. Proceeding in the same fashion with each of the remaining factors of d, we can cancel all of the factors of d into the two polynomials on the right hand side, leaving an equation $p(x) = a(x)b(x)$ with $a(x), b(x) \in R[x]$ and with $a(x), b(x)$ being F-multiples of $A(x)$, $B(x)$, respectively. This completes the proof.

Note that we cannot prove that $a(x)$ and $b(x)$ are necessarily R-multiples of $A(x)$, $B(x)$, respectively, because, for example, we could factor x^2 in $\mathbb{Q}[x]$ with $A(x) = 2x$ and $B(x) = \frac{1}{2}x$ but no *integer* multiples of $A(x)$ and $B(x)$ give a factorization of x^2 in $\mathbb{Z}[x]$.

The elements of the ring R become *units* in the Unique Factorization Domain $F[x]$ (the units in $F[x]$ being the nonzero elements of F). For example, $7x$ factors in $\mathbb{Z}[x]$ into a product of two irreducibles: 7 and x (so $7x$ is not irreducible in $\mathbb{Z}[x]$), whereas $7x$ is the unit 7 times the irreducible x in $\mathbb{Q}[x]$ (so $7x$ is irreducible in $\mathbb{Q}[x]$). The following corollary shows that this is essentially the *only* difference between the irreducible elements in $R[x]$ and those in $F[x]$.

Corollary 6. Let R be a Unique Factorization Domain, let F be its field of fractions and let $p(x) \in R[x]$. Suppose the greatest common divisor of the coefficients of $p(x)$ is 1. Then $p(x)$ is irreducible in $R[x]$ if and only if it is irreducible in $F[x]$. In particular, if $p(x)$ is a monic polynomial that is irreducible in $R[x]$, then $p(x)$ is irreducible in $F[x]$.

Proof: By Gauss' Lemma above, if $p(x)$ is reducible in $F[x]$, then it is reducible in $R[x]$. Conversely, the assumption on the greatest common divisor of the coefficients of $p(x)$ implies that if it is reducible in $R[x]$, then $p(x) = a(x)b(x)$ where neither $a(x)$ nor $b(x)$ are constant polynomials in $R[x]$. This same factorization shows that $p(x)$ is reducible in $F[x]$, completing the proof.

Theorem 7. R is a Unique Factorization Domain if and only if $R[x]$ is a Unique Factorization Domain.

Proof: We have indicated above that $R[x]$ a Unique Factorization Domain forces R to be a Unique Factorization Domain. Suppose conversely that R is a Unique Factorization Domain, F is its field of fractions and $p(x)$ is a nonzero element of $R[x]$. Let d be

the greatest common divisor of the coefficients of $p(x)$, so that $p(x) = dp'(x)$, where the g.c.d. of the coefficients of $p'(x)$ is 1. Such a factorization of $p(x)$ is unique up to a change in d (so up to a unit in R), and since d can be factored uniquely into irreducibles in R (and these are also irreducibles in the larger ring $R[x]$), it suffices to prove that $p'(x)$ can be factored uniquely into irreducibles in $R[x]$. Thus we may assume that the greatest common divisor of the coefficients of $p(x)$ is 1. We may further assume $p(x)$ is not a unit in $R[x]$, i.e., degree $p(x) > 0$.

Since $F[x]$ is a Unique Factorization Domain, $p(x)$ can be factored uniquely into irreducibles in $F[x]$. By Gauss' Lemma, such a factorization implies there is a factorization of $p(x)$ in $R[x]$ whose factors are F-multiples of the factors in $F[x]$. Since the greatest common divisor of the coefficients of $p(x)$ is 1, the g.c.d. of the coefficients in each of these factors in $R[x]$ must be 1. By Corollary 6, each of these factors is an irreducible in $R[x]$. This shows that $p(x)$ can be written as a finite product of irreducibles in $R[x]$.

The uniqueness of the factorization of $p(x)$ follows from the uniqueness in $F[x]$. Suppose

$$p(x) = q_1(x) \cdots q_r(x) = q_1'(x) \cdots q_s'(x)$$

are two factorizations of $p(x)$ into irreducibles in $R[x]$. Since the g.c.d. of the coefficients of $p(x)$ is 1, the same is true for each of the irreducible factors above — in particular, each has positive degree. By Corollary 6, each $q_i(x)$ and $q_j'(x)$ is an irreducible in $F[x]$. By unique factorization in $F[x]$, $r = s$ and, possibly after rearrangement, $q_i(x)$ and $q_i'(x)$ are associates in $F[x]$ for all $i \in \{1, \ldots, r\}$. It remains to show they are associates in $R[x]$. Since the units of $F[x]$ are precisely the elements of $F^\times$ we need to consider when $q(x) = \frac{a}{b}q'(x)$ for some $q(x), q'(x) \in R[x]$ and nonzero elements a, b of R, where the greatest common divisor of the coefficients of each of $q(x)$ and $q'(x)$ is 1. In this case $bq(x) = aq'(x)$; the g.c.d. of the coefficients on the left hand side is b and on the right hand side is a. Since in a Unique Factorization Domain the g.c.d. of the coefficients of a nonzero polynomial is unique up to units, $a = ub$ for some unit u in R. Thus $q(x) = uq'(x)$ and so $q(x)$ and $q'(x)$ are associates in R as well. This completes the proof.

Corollary 8. If R is a Unique Factorization Domain, then a polynomial ring in an arbitrary number of variables with coefficients in R is also a Unique Factorization Domain.

Proof: For finitely many variables, this follows by induction from Theorem 7, since a polynomial ring in n variables can be considered as a polynomial ring in one variable with coefficients in a polynomial ring in $n-1$ variables. The general case follows from the definition of a polynomial ring in an arbitrary number of variables as the union of polynomial rings in finitely many variables.

Examples

(1) $\mathbb{Z}[x]$, $\mathbb{Z}[x, y]$, etc. are Unique Factorization Domains. The ring $\mathbb{Z}[x]$ gives an example of a Unique Factorization Domain that is not a Principal Ideal Domain.

(2) Similarly, $\mathbb{Q}[x]$, $\mathbb{Q}[x, y]$, etc. are Unique Factorization Domains.

We saw earlier that if R is a Unique Factorization Domain with field of fractions F and $p(x) \in R[x]$, then we can factor out the greatest common divisor d of the coefficients of $p(x)$ to obtain $p(x) = dp'(x)$, where $p'(x)$ is irreducible in both $R[x]$ and $F[x]$. Suppose now that R is an *arbitrary* integral domain with field of fractions F. In R the notion of greatest common divisor may not make sense, however one might still ask if, say, a *monic* polynomial which is irreducible in $R[x]$ is still irreducible in $F[x]$ (i.e., whether the last statement in Corollary 6 is true).

Note first that if a monic polynomial $p(x)$ is reducible, it must have a factorization $p(x) = a(x)b(x)$ in $R[x]$ with both $a(x)$ and $b(x)$ *monic, nonconstant* polynomials (recall that the leading term of $p(x)$ is the product of the leading terms of the factors, so the leading coefficients of both $a(x)$ and $b(x)$ are units — we can thus arrange these to be 1). In other words, a nonconstant *monic* polynomial $p(x)$ is irreducible if and only if it cannot be factored as a product of two *monic* polynomials of smaller degree.

We now see that it is not true that if R is an arbitrary integral domain and $p(x)$ is a monic irreducible polynomial in $R[x]$, then $p(x)$ is irreducible in $F[x]$. For example, let $R = \mathbb{Z}[2i] = \{a + 2bi \mid a, b \in \mathbb{Z}\}$ (a subring of the complex numbers) and let $p(x) = x^2+1$. Then the fraction field of R is $F = \{a+bi \mid a, b \in \mathbb{Q}\}$. The polynomial $p(x)$ factors uniquely into a product of two linear factors in $F[x]$: $x^2+1 = (x-i)(x+i)$ so in particular, $p(x)$ *is reducible in* $F[x]$. Neither of these factors lies in $R[x]$ (because $i \notin R$) so $p(x)$ *is irreducible in* $R[x]$. In particular, by Corollary 6, $\mathbb{Z}[2i]$ *is not a Unique Factorization Domain.*

EXERCISES

1. Let R be an integral domain with quotient field F and let $p(x)$ be a monic polynomial in $R[x]$. Assume that $p(x) = a(x)b(x)$ where $a(x)$ and $b(x)$ are monic polynomials in $F[x]$ of smaller degree than $p(x)$. Prove that if $a(x) \notin R[x]$ then R is not a Unique Factorization Domain. Deduce that $\mathbb{Z}[2\sqrt{2}]$ is not a U.F.D.
2. Prove that if $f(x)$ and $g(x)$ are polynomials with rational coefficients whose product $f(x)g(x)$ has integer coefficients, then the product of any coefficient of $g(x)$ with any coefficient of $f(x)$ is an integer.
3. Let F be a field. Prove that the set R of polynomials in $F[x]$ whose coefficient of x is equal to 0 is a subring of $F[x]$ and that R is not a U.F.D. [Show that $x^6 = (x^2)^3 = (x^3)^2$ gives two distinct factorizations of x^6 into irreducibles.]
4. Let $R = \mathbb{Z} + x\mathbb{Q}[x] \subset \mathbb{Q}[x]$ be the set of polynomials in x with rational coefficients whose constant term is an integer.
   (a) Prove that R is an integral domain and its units are ± 1.
   (b) Show that the irreducibles in R are $\pm p$ where p is a prime in $\mathbb{Z}$ and the polynomials $f(x)$ that are irreducible in $\mathbb{Q}[x]$ and have constant term ± 1. Prove that these irreducibles are prime in R.
   (c) Show that x cannot be written as the product of irreducibles in R (in particular, x is not irreducible) and conclude that R is not a U.F.D.
   (d) Show that x is not a prime in R and describe the quotient ring $R/(x)$.
5. Let $R = \mathbb{Z} + x\mathbb{Q}[x] \subset \mathbb{Q}[x]$ be the ring considered in the previous exercise.
   (a) Suppose that $f(x), g(x) \in \mathbb{Q}[x]$ are two nonzero polynomials with rational coefficients and that x^r is the largest power of x dividing both $f(x)$ and $g(x)$ in $\mathbb{Q}[x]$, (i.e., r is the degree of the lowest order term appearing in either $f(x)$ or $g(x)$). Let f_r and

g_r be the coefficients of x^r in $f(x)$ and $g(x)$, respectively (one of which is nonzero by definition of r). Then $\mathbb{Z}f_r + \mathbb{Z}g_r = \mathbb{Z}d_r$ for some nonzero $d_r \in \mathbb{Q}$ (cf. Exercise 14 in Section 2.4). Prove that there is a polynomial $d(x) \in \mathbb{Q}[x]$ that is a g.c.d. of $f(x)$ and $g(x)$ in $\mathbb{Q}[x]$ and whose term of minimal degree is $d_r x^r$.

(b) Prove that $f(x) = d(x)q_1(x)$ and $g(x) = d(x)q_2(x)$ where $q_1(x)$ and $q_2(x)$ are elements of the subring R of $\mathbb{Q}[x]$.

(c) Prove that $d(x) = a(x)f(x) + b(x)g(x)$ for polynomials $a(x), b(x)$ in R. [The existence of $a(x), b(x)$ in the Euclidean Domain $\mathbb{Q}[x]$ is immediate. Use Exercise 11 in Section 2 to show that $a(x)$ and $b(x)$ can be chosen to lie in R.]

(d) Conclude from (a) and (b) that $Rf(x) + Rg(x) = Rd(x)$ in $\mathbb{Q}[x]$ and use this to prove that R is a Bezout Domain (cf. Exercise 7 in Section 8.2).

(e) Show that (d), the results of the previous exercise, and Exercise 11 of Section 8.3 imply that R must contain ideals that are not principal (hence not finitely generated). Prove that in fact $I = x\mathbb{Q}[x]$ is an ideal of R that is not finitely generated.

9.4 IRREDUCIBILITY CRITERIA

If R is a Unique Factorization Domain, then by Corollary 8 a polynomial ring in any number of variables with coefficients in R is also a Unique Factorization Domain. It is of interest then to determine the irreducible elements in such a polynomial ring, particularly in the ring $R[x]$. In the one-variable case, a nonconstant monic polynomial is irreducible in $R[x]$ if it cannot be factored as the product of two other polynomials of smaller degrees. Determining whether a polynomial has factors is frequently difficult to check, particularly for polynomials of large degree in several variables. The purpose of irreducibility criteria is to give an easier mechanism for determining when some types of polynomials are irreducible.

For the most part we restrict attention to polynomials in one variable where the coefficient ring is a Unique Factorization Domain. By Gauss' Lemma it suffices to consider factorizations in $F[x]$ where F is the field of fractions of R (although we shall occasionally consider questions of irreducibility when the coefficient ring is just an integral domain). The next proposition considers when there is a factor of degree one (a *linear* factor).

Proposition 9. Let F be a field and let $p(x) \in F[x]$. Then $p(x)$ has a factor of degree one if and only if $p(x)$ has a root in F, i.e., there is an $\alpha \in F$ with $p(\alpha) = 0$.

Proof: If $p(x)$ has a factor of degree one, then since F is a field, we may assume the factor is monic, i.e., is of the form $(x - \alpha)$ for some $\alpha \in F$. But then $p(\alpha) = 0$. Conversely, suppose $p(\alpha) = 0$. By the Division Algorithm in $F[x]$ we may write

$$p(x) = q(x)(x - \alpha) + r$$

where r is a constant. Since $p(\alpha) = 0$, r must be 0, hence $p(x)$ has $(x - \alpha)$ as a factor.

Proposition 9 gives a criterion for irreducibility for polynomials of small degree:

Proposition 10. A polynomial of degree two or three over a field F is reducible if and only if it has a root in F.

Proof: This follows immediately from the previous proposition, since a polynomial of degree two or three is reducible if and only if it has at least one linear factor.

The next result limits the possibilities for roots of polynomials with integer coefficients (it is stated for $\mathbb{Z}[x]$ for convenience although it clearly generalizes to $R[x]$, where R is any Unique Factorization Domain).

Proposition 11. Let $p(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_0$ be a polynomial of degree n with integer coefficients. If $r/s \in \mathbb{Q}$ is in lowest terms (i.e., r and s are relatively prime integers) and r/s is a root of $p(x)$, then r divides the constant term and s divides the leading coefficient of $p(x)$: $r \mid a_0$ and $s \mid a_n$. In particular, if $p(x)$ is a *monic* polynomial with integer coefficients and $p(d) \neq 0$ for all integers d dividing the constant term of $p(x)$, then $p(x)$ has no roots in $\mathbb{Q}$.

Proof: By hypothesis, $p(r/s) = 0 = a_n(r/s)^n + a_{n-1}(r/s)^{n-1} + \cdots + a_0$. Multiplying through by s^n gives

$$0 = a_n r^n + a_{n-1}r^{n-1}s + \cdots + a_0 s^n.$$

Thus $a_n r^n = s(-a_{n-1}r^{n-1} - \cdots - a_0 s^{n-1})$, so s divides $a_n r^n$. By assumption, s is relatively prime to r and it follows that $s \mid a_n$. Similarly, solving the equation for $a_0 s^n$ shows that $r \mid a_0$. The last assertion of the proposition follows from the previous ones.

Examples

(1) The polynomial $x^3 - 3x - 1$ is irreducible in $\mathbb{Z}[x]$. To prove this, by Gauss' Lemma and Proposition 10 it suffices to show it has no rational roots. By Proposition 11 the only candidates for rational roots are integers which divide the constant term 1, namely ± 1. Substituting both 1 and -1 into the polynomial shows that these are not roots.

(2) For p any prime the polynomials $x^2 - p$ and $x^3 - p$ are irreducible in $\mathbb{Q}[x]$. This is because they have degrees ≤ 3 so it suffices to show they have no rational roots. By Proposition 11 the only candidates for roots are ± 1 and $\pm p$, but none of these give 0 when they are substituted into the polynomial.

(3) The polynomial $x^2 + 1$ is reducible in $\mathbb{Z}/2\mathbb{Z}[x]$ since it has 1 as a root, and it factors as $(x + 1)^2$.

(4) The polynomial $x^2 + x + 1$ is irreducible in $\mathbb{Z}/2\mathbb{Z}[x]$ since it does not have a root in $\mathbb{Z}/2\mathbb{Z}$: $0^2 + 0 + 1 = 1$ and $1^2 + 1 + 1 = 1$.

(5) Similarly, the polynomial $x^3 + x + 1$ is irreducible in $\mathbb{Z}/2\mathbb{Z}[x]$.

This technique is limited to polynomials of low degree because it relies on the presence of a factor of degree one. A polynomial of degree 4, for example, may be the product of two irreducible quadratics, hence be reducible but have no linear factor. One fairly general technique for checking irreducibility uses Proposition 2 above and consists of reducing the coefficients modulo some ideal.

Proposition 12. Let I be a proper ideal in the integral domain R and let $p(x)$ be a nonconstant monic polynomial in $R[x]$. If the image of $p(x)$ in $(R/I)[x]$ cannot be factored in $(R/I)[x]$ into two polynomials of smaller degree, then $p(x)$ is irreducible in $R[x]$.

Proof: Suppose $p(x)$ cannot be factored in $(R/I)[x]$ but that $p(x)$ is reducible in $R[x]$. As noted at the end of the preceding section this means there are monic, nonconstant polynomials $a(x)$ and $b(x)$ in $R[x]$ such that $p(x) = a(x)b(x)$. By Proposition 2, reducing the coefficients modulo I gives a factorization in $(R/I)[x]$ with nonconstant factors, a contradiction.

This proposition indicates that if it is possible to find a proper ideal I such that the *reduced* polynomial cannot be factored, then the polynomial is itself irreducible. Unfortunately, there are examples of polynomials even in $\mathbb{Z}[x]$ which are irreducible but whose reductions modulo every ideal are reducible (so their irreducibility is not detectable by this technique). For example, the polynomial $x^4 + 1$ is irreducible in $\mathbb{Z}[x]$ but is reducible modulo every prime (we shall verify this in Chapter 14) and the polynomial $x^4 - 72x^2 + 4$ is irreducible in $\mathbb{Z}[x]$ but is reducible modulo every integer.

Examples

(1) Consider the polynomial $p(x) = x^2 + x + 1$ in $\mathbb{Z}[x]$. Reducing modulo 2, we see from Example 4 above that $p(x)$ is irreducible in $\mathbb{Z}[x]$. Similarly, $x^3 + x + 1$ is irreducible in $\mathbb{Z}[x]$ because it is irreducible in $\mathbb{Z}/2\mathbb{Z}[x]$.

(2) The polynomial $x^2 + 1$ is irreducible in $\mathbb{Z}[x]$ since it is irreducible in $\mathbb{Z}/3\mathbb{Z}[x]$ (no root in $\mathbb{Z}/3\mathbb{Z}$), but is reducible mod 2. This shows that the converse to Proposition 12 does not hold.

(3) The idea of reducing modulo an ideal to determine irreducibility can be used also in several variables, but some care must be exercised. For example, the polynomial $x^2 + xy + 1$ in $\mathbb{Z}[x, y]$ is irreducible since modulo the ideal (y) it is $x^2 + 1$ in $\mathbb{Z}[x]$, which is irreducible and of the same degree. In this sort of argument it is necessary to be careful about "collapsing." For example, the polynomial $xy + x + y + 1$ (which is $(x + 1)(y + 1)$) is reducible, but appears irreducible modulo both (x) and (y). The reason for this is that nonunit polynomials in $\mathbb{Z}[x, y]$ can reduce to units in the quotient. To take account of this it is necessary to determine which elements in the original ring become units in the quotient. The elements in $\mathbb{Z}[x, y]$ which are units modulo (y), for example, are the polynomials in $\mathbb{Z}[x, y]$ with constant term ± 1 and all nonconstant terms divisible by y. The fact that $x^2 + xy + 1$ and its reduction mod (y) have the same degree therefore eliminates the possibility of a factor which is a unit modulo (y), but not a unit in $\mathbb{Z}[x, y]$ and gives the irreducibility of this polynomial.

A special case of reducing modulo an ideal to test for irreducibility which is frequently useful is known as *Eisenstein's Criterion* (although originally proved earlier by Schönemann, so more properly known as the *Eisenstein-Schönemann Criterion*):

Proposition 13. *(Eisenstein's Criterion)* Let P be a prime ideal of the integral domain R and let $f(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0$ be a polynomial in $R[x]$ (here $n \geq 1$). Suppose $a_{n-1}, \ldots, a_1, a_0$ are all elements of P and suppose a_0 is not an element of P^2. Then $f(x)$ is irreducible in $R[x]$.

Proof: Suppose $f(x)$ were reducible, say $f(x) = a(x)b(x)$ in $R[x]$, where $a(x)$ and $b(x)$ are nonconstant polynomials. Reducing this equation modulo P and using the assumptions on the coefficients of $f(x)$ we obtain the equation $x^n = \overline{a(x)}\,\overline{b(x)}$ in $(R/P)[x]$, where the bar denotes the polynomials with coefficients reduced mod P. Since P is a prime ideal, R/P is an integral domain, and it follows that both $\overline{a(x)}$ and $\overline{b(x)}$ have 0 constant term, i.e., the constant terms of both $a(x)$ and $b(x)$ are elements of P. But then the constant term a_0 of $f(x)$ as the product of these two would be an element of P^2, a contradiction.

Eisenstein's Criterion is most frequently applied to $\mathbb{Z}[x]$ so we state the result explicitly for this case:

Corollary 14. *(Eisenstein's Criterion for $\mathbb{Z}[x]$)* Let p be a prime in $\mathbb{Z}$ and let $f(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0 \in \mathbb{Z}[x]$, $n \geq 1$. Suppose p divides a_i for all $i \in \{0, 1, \ldots, n-1\}$ but that p^2 does not divide a_0. Then $f(x)$ is irreducible in both $\mathbb{Z}[x]$ and $\mathbb{Q}[x]$.

Proof: This is simply a restatement of Proposition 13 in the case of the prime ideal (p) in $\mathbb{Z}$ together with Corollary 6.

Examples

(1) The polynomial $x^4 + 10x + 5$ in $\mathbb{Z}[x]$ is irreducible by Eisenstein's Criterion applied for the prime 5.

(2) If a is any integer which is divisible by some prime p but not divisible by p^2, then $x^n - a$ is irreducible in $\mathbb{Z}[x]$ by Eisenstein's Criterion. In particular, $x^n - p$ is irreducible for all positive integers n and so for $n \geq 2$ the n^{th} roots of p are not rational numbers (i.e., this polynomial has no root in $\mathbb{Q}$).

(3) Consider the polynomial $f(x) = x^4 + 1$ mentioned previously. Eisenstein's Criterion does not apply directly to $f(x)$. The polynomial $g(x) = f(x+1)$ is $(x+1)^4 + 1$, i.e., $x^4 + 4x^3 + 6x^2 + 4x + 2$, and Eisenstein's Criterion for the prime 2 shows that this polynomial is irreducible. It follows then that $f(x)$ must also be irreducible, since any factorization for $f(x)$ would provide a factorization for $g(x)$ (just replace x by $x + 1$ in each of the factors). This example shows that Eisenstein's Criterion can sometimes be used to verify the irreducibility of a polynomial to which it does not immediately apply.

(4) As another example of this, let p be a prime and consider the polynomial

$$\Phi_p(x) = \frac{x^p - 1}{x - 1} = x^{p-1} + x^{p-2} + \cdots + x + 1,$$

an example of a *cyclotomic polynomial* which we shall consider more thoroughly in Part IV. Again, Eisenstein's Criterion does not immediately apply, but it does apply for the prime p to the polynomial

$$\Phi_p(x+1) = \frac{(x+1)^p - 1}{x} = x^{p-1} + px^{p-2} + \cdots + \frac{p(p-1)}{2}x + p \in \mathbb{Z}[x]$$

since all the coefficients except the first are divisible by p by the Binomial Theorem. As before, this shows $\Phi_p(x)$ is irreducible in $\mathbb{Z}[x]$.

(5) As an example of the use of the more general Eisenstein's Criterion in Proposition 13 we mimic Example 2 above. Let $R = \mathbb{Q}[x]$ and let n be any positive integer. Consider

the polynomial $X^n - x$ in the ring $R[X]$. The ideal (x) is prime in the coefficient ring R since $R/(x) = \mathbb{Q}[x]/(x)$ is the integral domain $\mathbb{Q}$. Eisenstein's Criterion for the ideal (x) of R applies directly to show that $X^n - x$ is irreducible in $R[X]$. Note that this construction works with $\mathbb{Q}$ replaced by any field or, indeed, by any integral domain.

There are now efficient algorithms for factoring polynomials over certain fields. For polynomials with integer coefficients these algorithms have been implemented in a number of computer packages. An efficient algorithm for factoring polynomials over $\mathbb{F}_p$, called the Berlekamp Algorithm, is described in detail in the exercises at the end of Section 14.3.

EXERCISES

1. Determine whether the following polynomials are irreducible in the rings indicated. For those that are reducible, determine their factorization into irreducibles. The notation $\mathbb{F}_p$ denotes the finite field $\mathbb{Z}/p\mathbb{Z}$, p a prime.
   (a) $x^2 + x + 1$ in $\mathbb{F}_2[x]$.
   (b) $x^3 + x + 1$ in $\mathbb{F}_3[x]$.
   (c) $x^4 + 1$ in $\mathbb{F}_5[x]$.
   (d) $x^4 + 10x^2 + 1$ in $\mathbb{Z}[x]$.
2. Prove that the following polynomials are irreducible in $\mathbb{Z}[x]$:
   (a) $x^4 - 4x^3 + 6$
   (b) $x^6 + 30x^5 - 15x^3 + 6x - 120$
   (c) $x^4 + 4x^3 + 6x^2 + 2x + 1$ [Substitute $x - 1$ for x.]
   (d) $\dfrac{(x+2)^p - 2^p}{x}$, where p is an odd prime.
3. Show that the polynomial $(x-1)(x-2)\cdots(x-n) - 1$ is irreducible over $\mathbb{Z}$ for all $n \geq 1$. [If the polynomial factors consider the values of the factors at $x = 1, 2, \ldots, n$.]
4. Show that the polynomial $(x-1)(x-2)\cdots(x-n)+1$ is irreducible over $\mathbb{Z}$ for all $n \geq 1$, $n \neq 4$.
5. Find all the monic irreducible polynomials of degree ≤ 3 in $\mathbb{F}_2[x]$, and the same in $\mathbb{F}_3[x]$.
6. Construct fields of each of the following orders: **(a)** 9, **(b)** 49, **(c)** 8, **(d)** 81 (you may exhibit these as $F[x]/(f(x))$ for some F and f). [Use Exercises 2 and 3 in Section 2.]
7. Prove that $\mathbb{R}[x]/(x^2 + 1)$ is a field which is isomorphic to the complex numbers.
8. Prove that $K_1 = \mathbb{F}_{11}[x]/(x^2+1)$ and $K_2 = \mathbb{F}_{11}[y]/(y^2+2y+2)$ are both fields with 121 elements. Prove that the map which sends the element $p(\bar{x})$ of K_1 to the element $p(\bar{y}+1)$ of K_2 (where p is any polynomial with coefficients in $\mathbb{F}_{11}$) is well defined and gives a ring (hence field) isomorphism from K_1 to K_2.
9. Prove that the polynomial $x^2 - \sqrt{2}$ is irreducible over $\mathbb{Z}[\sqrt{2}]$ (you may use the fact that $\mathbb{Z}[\sqrt{2}]$ is a U.F.D. — cf. Exercise 9 of Section 8.1).
10. Prove that the polynomial $p(x) = x^4 - 4x^2 + 8x + 2$ is irreducible over the quadratic field $F = \mathbb{Q}(\sqrt{-2}) = \{a + b\sqrt{-2} \mid a, b \in \mathbb{Q}\}$. [First use the method of Proposition 11 for the Unique Factorization Domain $\mathbb{Z}[\sqrt{-2}]$ (cf. Exercise 8, Section 8.1) to show that if $\alpha \in \mathbb{Z}[\sqrt{-2}]$ is a root of $p(x)$ then α is a divisor of 2 in $\mathbb{Z}[\sqrt{-2}]$. Conclude that α must be ± 1, $\pm\sqrt{-2}$ or ± 2, and hence show $p(x)$ has no linear factor over F. Show similarly that $p(x)$ is not the product of two quadratics with coefficients in F.]

11. Prove that $x^2 + y^2 - 1$ is irreducible in $\mathbb{Q}[x, y]$.

12. Prove that $x^{n-1} + x^{n-2} + \cdots + x + 1$ is irreducible over $\mathbb{Z}$ if and only if n is a prime.

13. Prove that $x^3 + nx + 2$ is irreducible over $\mathbb{Z}$ for all integers $n \neq 1, -3, -5$.

14. Factor each of the two polynomials: $x^8 - 1$ and $x^6 - 1$ into irreducibles over each of the following rings: **(a)** $\mathbb{Z}$, **(b)** $\mathbb{Z}/2\mathbb{Z}$, **(c)** $\mathbb{Z}/3\mathbb{Z}$.

15. Prove that if F is a field then the polynomial $X^n - x$ which has coefficients in the ring $F[[x]]$ of formal power series (cf. Exercise 3 of Section 7.2) is irreducible over $F[[x]]$. [Recall that $F[[x]]$ is a Euclidean Domain — cf. Exercise 5, Section 7.2 and Example 4, Section 8.1.]

16. Let F be a field and let $f(x)$ be a polynomial of degree n in $F[x]$. The polynomial $g(x) = x^n f(1/x)$ is called the *reverse* of $f(x)$.

(a) Describe the coefficients of g in terms of the coefficients of f.

(b) Prove that f is irreducible if and only if g is irreducible.

17. Prove the following variant of Eisenstein's Criterion: let P be a prime ideal in the Unique Factorization Domain R and let $f(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$ be a polynomial in $R[x]$, $n \geq 1$. Suppose $a_n \notin P$, $a_{n-1}, \ldots, a_0 \in P$ and $a_0 \notin P^2$. Prove that $f(x)$ is irreducible in $F[x]$, where F is the quotient field of R.

18. Show that $6x^5 + 14x^3 - 21x + 35$ and $18x^5 - 30x^2 + 120x + 360$ are irreducible in $\mathbb{Q}[x]$.

19. Let F be a field and let $f(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_0 \in F[x]$. The *derivative*, $D_x(f(x))$, of $f(x)$ is defined by

$$D_x(f(x)) = na_n x^{n-1} + (n-1)a_{n-1}x^{n-2} + \cdots + a_1$$

where, as usual, $na = a + a + \cdots + a$ (n times). Note that $D_x(f(x))$ is again a polynomial with coefficients in F.

The polynomial $f(x)$ is said to have a *multiple root* if there is some field E containing F and some $\alpha \in E$ such that $(x - \alpha)^2$ divides $f(x)$ in $E[x]$. For example, the polynomial $f(x) = (x - 1)^2(x - 2) \in \mathbb{Q}[x]$ has $\alpha = 1$ as a multiple root and the polynomial $f(x) = x^4 + 2x^2 + 1 = (x^2 + 1)^2 \in \mathbb{R}[x]$ has $\alpha = \pm i \in \mathbb{C}$ as multiple roots. We shall prove in Section 13.5 that a nonconstant polynomial $f(x)$ has a multiple root if and only if $f(x)$ is not relatively prime to its derivative (which can be detected by the Euclidean Algorithm in $F[x]$). Use this criterion to determine whether the following polynomials have multiple roots:

(a) $x^3 - 3x - 2 \in \mathbb{Q}[x]$

(b) $x^3 + 3x + 2 \in \mathbb{Q}[x]$

(c) $x^6 - 4x^4 + 6x^3 + 4x^2 - 12x + 9 \in \mathbb{Q}[x]$

(d) Show for any prime p and any $a \in \mathbb{F}_p$ that the polynomial $x^p - a$ has a multiple root.

20. Show that the polynomial $f(x) = x$ in $\mathbb{Z}/6\mathbb{Z}[x]$ factors as $(3x + 4)(4x + 3)$, hence is not an irreducible polynomial.

(a) Show that the reduction of $f(x)$ modulo both of the nontrivial ideals (2) and (3) of $\mathbb{Z}/6\mathbb{Z}$ is an irreducible polynomial, showing that the condition that R be an integral domain in Proposition 12 is necessary.

(b) Show that in any factorization $f(x) = g(x)h(x)$ in $\mathbb{Z}/6\mathbb{Z}[x]$ the reduction of $g(x)$ modulo (2) is either 1 or x and the reduction of $h(x)$ modulo (2) is then either x or 1, and similarly for the reductions modulo (3). Determine all the factorizations of $f(x)$ in $\mathbb{Z}/6\mathbb{Z}[x]$. [Use the Chinese Remainder Theorem.]

(c) Show that the ideal $(3, x)$ is a principal ideal in $\mathbb{Z}/6\mathbb{Z}[x]$.

(d) Show that over the ring $\mathbb{Z}/30\mathbb{Z}[x]$ the polynomial $f(x) = x$ has the factorization

$f(x) = (10x+21)(15x+16)(6x+25)$. Prove that the product of any of these factors is again of the same degree. Prove that the reduction of $f(x)$ modulo any prime in $\mathbb{Z}/30\mathbb{Z}$ is an irreducible polynomial. Determine all the factorizations of $f(x)$ in $\mathbb{Z}/30\mathbb{Z}[x]$. [Consider the reductions modulo (2), (3) and (5) and use the Chinese Remainder Theorem.]

(e) Generalize part (d) to $\mathbb{Z}/n\mathbb{Z}[x]$ where n is the product of k distinct primes.

9.5 POLYNOMIAL RINGS OVER FIELDS II

Let F be a field. We prove here some additional results for the one-variable polynomial ring $F[x]$. The first is a restatement of results obtained earlier.

Proposition 15. The maximal ideals in $F[x]$ are the ideals $(f(x))$ generated by irreducible polynomials $f(x)$. In particular, $F[x]/(f(x))$ is a field if and only if $f(x)$ is irreducible.

Proof: This follows from Proposition 7 of Section 8.2 applied to the Principal Ideal Domain $F[x]$.

Proposition 16. Let $g(x)$ be a nonconstant element of $F[x]$ and let

$$g(x) = f_1(x)^{n_1} f_2(x)^{n_2} \cdots f_k(x)^{n_k}$$

be its factorization into irreducibles, where the $f_i(x)$ are distinct. Then we have the following isomorphism of rings:

$$F[x]/(g(x)) \cong F[x]/(f_1(x)^{n_1}) \times F[x]/(f_2(x)^{n_2}) \times \cdots \times F[x]/(f_k(x)^{n_k}).$$

Proof: This follows from the Chinese Remainder Theorem (Theorem 7.17), since the ideals $(f_i(x)^{n_i})$ and $(f_j(x)^{n_j})$ are comaximal if $f_i(x)$ and $f_j(x)$ are distinct (they are relatively prime in the Euclidean Domain $F[x]$, hence the ideal generated by them is $F[x]$).

The next result concerns the number of roots of a polynomial over a field F. By Proposition 9, a root α corresponds to a linear factor $(x-\alpha)$ of $f(x)$. If $f(x)$ is divisible by $(x-\alpha)^m$ but not by $(x-\alpha)^{m+1}$, then α is said to be a root of *multiplicity* m.

Proposition 17. If the polynomial $f(x)$ has roots $\alpha_1, \alpha_2, \ldots, \alpha_k$ in F (not necessarily distinct), then $f(x)$ has $(x-\alpha_1)\cdots(x-\alpha_k)$ as a factor. In particular, a polynomial of degree n in one variable over a field F has at most n roots in F, even counted with multiplicity.

Proof: The first statement follows easily by induction from Proposition 9. Since linear factors are irreducible, the second statement follows since $F[x]$ is a Unique Factorization Domain.

This last result has the following interesting consequence.

Proposition 18. A finite subgroup of the multiplicative group of a field is cyclic. In particular, if F is a finite field, then the multiplicative group $F^\times$ of nonzero elements of F is a cyclic group.

Proof: We give a proof of this result using the Fundamental Theorem of Finitely Generated Abelian Groups (Theorem 3 in Section 5.2). A more number-theoretic proof is outlined in the exercises, or Proposition 5 in Section 6.1 may be used in place of the Fundamental Theorem. By the Fundamental Theorem, the finite subgroup can be written as the direct product of cyclic groups

$$\mathbb{Z}/n_1\mathbb{Z} \times \mathbb{Z}/n_2\mathbb{Z} \times \cdots \times \mathbb{Z}/n_k\mathbb{Z}$$

where $n_k \mid n_{k-1} \mid \cdots \mid n_2 \mid n_1$. In general, if G is a cyclic group and $d \mid |G|$ then G contains precisely d elements of order dividing d. Since n_k divides the order of each of the cyclic groups in the direct product, it follows that each direct factor contains n_k elements of order dividing n_k. If k were greater than 1, there would therefore be a total of more than n_k such elements. But then there would be more than n_k roots of the polynomial $x^{n_k} - 1$ in the field F, contradicting Proposition 17. Hence $k = 1$ and the group is cyclic.

Corollary 19. Let p be a prime. The multiplicative group $(\mathbb{Z}/p\mathbb{Z})^\times$ of nonzero residue classes mod p is cyclic.

Proof: This is the multiplicative group of the finite field $\mathbb{Z}/p\mathbb{Z}$.

Corollary 20. Let $n \geq 2$ be an integer with factorization $n = p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_r^{\alpha_r}$ in $\mathbb{Z}$, where $p_1, \ldots, p_r$ are distinct primes. We have the following isomorphisms of (multiplicative) groups:

(1) $(\mathbb{Z}/n\mathbb{Z})^\times \cong (\mathbb{Z}/p_1^{\alpha_1}\mathbb{Z})^\times \times (\mathbb{Z}/p_2^{\alpha_2}\mathbb{Z})^\times \times \cdots \times (\mathbb{Z}/p_r^{\alpha_r}\mathbb{Z})^\times$

(2) $(\mathbb{Z}/2^\alpha\mathbb{Z})^\times$ is the direct product of a cyclic group of order 2 and a cyclic group of order $2^{\alpha-2}$, for all $\alpha \geq 2$

(3) $(\mathbb{Z}/p^\alpha\mathbb{Z})^\times$ is a cyclic group of order $p^{\alpha-1}(p-1)$, for all odd primes p.

Remark: These isomorphisms describe the group-theoretic structure of the automorphism group of the cyclic group, Z_n, of order n since $\mathrm{Aut}(Z_n) \cong (\mathbb{Z}/n\mathbb{Z})^\times$ (cf. Proposition 16 in Section 4.4). In particular, for p a prime the automorphism group of the cyclic group of order p is cyclic of order $p-1$.

Proof: This is mainly a matter of collecting previous results. The isomorphism in (1) follows from the Chinese Remainder Theorem (see Corollary 18, Section 7.6). The isomorphism in (2) follows directly from Exercises 22 and 23 of Section 2.3.

For p an odd prime, $(\mathbb{Z}/p^\alpha\mathbb{Z})^\times$ is an abelian group of order $p^{\alpha-1}(p-1)$. By Exercise 21 of Section 2.3 the Sylow p-subgroup of this group is cyclic. The map

$$\mathbb{Z}/p^\alpha\mathbb{Z} \to \mathbb{Z}/p\mathbb{Z} \qquad \text{defined by} \qquad a + (p^\alpha) \mapsto a + (p)$$

is a ring homomorphism (reduction mod p) which gives a surjective group homomorphism from $(\mathbb{Z}/p^\alpha\mathbb{Z})^\times$ onto $(\mathbb{Z}/p\mathbb{Z})^\times$. The latter group is cyclic of order $p-1$

(Corollary 19). The kernel of this map is of order $p^{\alpha-1}$, hence for all primes $q \neq p$, the Sylow q-subgroup of $(\mathbb{Z}/p^\alpha\mathbb{Z})^\times$ maps isomorphically into the cyclic group $(\mathbb{Z}/p\mathbb{Z})^\times$. All Sylow subgroups of $(\mathbb{Z}/p^\alpha\mathbb{Z})^\times$ are therefore cyclic, so (3) holds, completing the proof.

EXERCISES

1. Let F be a field and let $f(x)$ be a nonconstant polynomial in $F[x]$. Describe the nilradical of $F[x]/(f(x))$ in terms of the factorization of $f(x)$ (cf. Exercise 29, Section 7.3).
2. For each of the fields constructed in Exercise 6 of Section 4 exhibit a generator for the (cyclic) multiplicative group of nonzero elements.
3. Let p be an odd prime in $\mathbb{Z}$ and let n be a positive integer. Prove that $x^n - p$ is irreducible over $\mathbb{Z}[i]$. [Use Proposition 18 in Chapter 8 and Eisenstein's Criterion.]
4. Prove that $x^3 + 12x^2 + 18x + 6$ is irreducible over $\mathbb{Z}[i]$. [Use Proposition 8.18 and Eisenstein's Criterion.]
5. Let φ denote Euler's φ-function. Prove the identity $\sum_{d|n} \varphi(d) = n$, where the sum is extended over all the divisors d of n. [First observe that the identity is valid when $n = p^m$ is the power of a prime p since the sum telescopes. Write $n = p^m n'$ where p does not divide n'. Prove that $\sum_{d|n} \varphi(d) = \sum_{d''|p^m} \varphi(d'') \sum_{d'|n'} \varphi(d')$ by multiplying out the right hand side and using the multiplicativity $\varphi(ab) = \varphi(a)\varphi(b)$ when a and b are relatively prime. Use induction to complete the proof. This problem may be done alternatively by letting Z be the cyclic group of order n and showing that since Z contains a unique subgroup of order d for each d dividing n, the number of elements of Z of order d is $\varphi(d)$. Then $|Z|$ is the sum of $\varphi(d)$ as d runs over all divisors of n.]
6. Let G be a finite subgroup of order n of the multiplicative group $F^\times$ of nonzero elements of the field F. Let φ denote Euler's φ-function and let $\psi(d)$ denote the number of elements of G of order d. Prove that $\psi(d) = \varphi(d)$ for every divisor d of n. In particular conclude that $\psi(n) \geq 1$, so that G is a cyclic group. [Observe that for any integer $N \geq 1$ the polynomial $x^N - 1$ has at most N roots in F. Conclude that for any integer N we have $\sum_{d|N} \psi(d) \leq N$. Since $\sum_{d|N} \varphi(d) = N$ by the previous exercise, show by induction that $\psi(d) \leq \varphi(d)$ for every divisor d of n. Since $\sum_{d|n} \psi(d) = n = \sum_{d|n} \varphi(d)$ show that this implies $\psi(d) = \varphi(d)$ for every divisor d of n.]
7. Prove that the additive and multiplicative groups of a field are never isomorphic. [Consider three cases: when $|F|$ is finite, when $-1 \neq 1$ in F, and when $-1 = 1$ in F.]

9.6 POLYNOMIALS IN SEVERAL VARIABLES OVER A FIELD AND GRÖBNER BASES

In this section we consider polynomials in many variables, present some basic computational tools, and indicate some applications. The results of this section are not required in Chapters 10 through 14. Additional applications will be given in Chapter 15.

We proved in Section 2 that a polynomial ring $F[x]$ in a variable x over a field F is a Euclidean Domain, and Corollary 8 showed that the polynomial ring $F[x_1, \ldots, x_n]$ is a U.F.D. However it follows from Corollary 8 in Section 8.2 that the latter ring is not a P.I.D. unless $n = 1$. Our first result below shows that ideals in such polynomial rings, although not necessarily principal, are always finitely generated. General rings with this property are given a special name:

Definition. A commutative ring R with 1 is called *Noetherian* if every ideal of R is finitely generated.

Noetherian rings will be studied in greater detail in Chapters 15 and 16. In this section we develop some of the basic theory and resulting algorithms for working with (finitely generated) ideals in $F[x_1, \ldots, x_n]$.

As we saw in Section 1, a polynomial ring in n variables can be considered as a polynomial ring in one variable with coefficients in a polynomial ring in $n-1$ variables. By following this inductive approach—as we did in Theorem 7 and Corollary 8—we can deduce that $F[x_1, x_2, \ldots, x_n]$ is Noetherian from the following more general result.

Theorem 21. *(Hilbert's Basis Theorem)* If R is a Noetherian ring then so is the polynomial ring $R[x]$.

Proof: Let I be an ideal in $R[x]$ and let L be the set of all leading coefficients of the elements in I. We first show that L is an ideal of R, as follows. Since I contains the zero polynomial, $0 \in L$. Let $f = ax^d + \cdots$ and $g = bx^e + \cdots$ be polynomials in I of degrees d, e and leading coefficients $a, b \in R$. Then for any $r \in R$ either $ra - b$ is zero or it is the leading coefficient of the polynomial $rx^e f - x^d g$. Since the latter polynomial is in I we have $ra - b \in L$, which shows L is an ideal of R. Since R is assumed Noetherian, the ideal L in R is finitely generated, say by $a_1, a_2, \ldots, a_n \in R$. For each $i = 1, \ldots, n$ let f_i be an element of I whose leading coefficient is a_i. Let e_i denote the degree of f_i, and let N be the maximum of $e_1, e_2, \ldots, e_n$.

For each $d \in \{0, 1, \ldots, N-1\}$, let L_d be the set of all leading coefficients of polynomials in I of degree d together with 0. A similar argument as that for L shows each L_d is also an ideal of R, again finitely generated since R is Noetherian. For each nonzero ideal L_d let $b_{d,1}, b_{d,2}, \ldots, b_{d,n_d} \in R$ be a set of generators for L_d, and let $f_{d,i}$ be a polynomial in I of degree d with leading coefficient $b_{d,i}$.

We show that the polynomials $f_1, \ldots, f_n$ together with all the polynomials $f_{d,i}$ for all the nonzero ideals L_d are a set of generators for I, i.e., that

$$I = (\{f_1, \ldots, f_n\} \cup \{f_{d,i} \mid 0 \le d < N,\ 1 \le i \le n_d\}).$$

By construction, the ideal I' on the right above is contained in I since all the generators were chosen in I. If $I' \neq I$, there exists a nonzero polynomial $f \in I$ of minimum degree with $f \notin I'$. Let $d = \deg f$ and let a be the leading coefficient of f.

Suppose first that $d \ge N$. Since $a \in L$ we may write a as an R-linear combination of the generators of L: $a = r_1 a_1 + \cdots + r_n a_n$. Then $g = r_1 x^{d-e_1} f_1 + \cdots + r_n x^{d-e_n} f_n$ is an element of I' with the same degree d and the same leading coefficient a as f. Then $f - g \in I$ is a polynomial in I of smaller degree than f. By the minimality of f, we must have $f - g = 0$, so $f = g \in I'$, a contradiction.

Suppose next that $d < N$. In this case $a \in L_d$ for some $d < N$, and so we may write $a = r_1 b_{d,1} + \cdots + r_{n_d} b_{n_d}$ for some $r_i \in R$. Then $g = r_1 f_{d,1} + \cdots + r_{n_d} f_{n_d}$ is a polynomial in I' with the same degree d and the same leading coefficient a as f, and we have a contradiction as before.

It follows that $I = I'$ is finitely generated, and since I was arbitrary, this completes the proof that $R[x]$ is Noetherian.

Since a field is clearly Noetherian, Hilbert's Basis Theorem and induction immediately give:

Corollary 22. Every ideal in the polynomial ring $F[x_1, x_2, \ldots, x_n]$ with coefficients from a field F is finitely generated.

If I is an ideal in $F[x_1, \ldots, x_n]$ generated by a (possibly infinite) set $\mathcal{S}$ of polynomials, Corollary 22 shows that I is finitely generated, and in fact I is generated by a finite number of the polynomials from the set $\mathcal{S}$ (cf. Exercise 1).

As the proof of Hilbert's Basis Theorem shows, the collection of leading coefficients of the polynomials in an ideal I in $R[x]$ forms an extremely useful ideal in R that can be used to understand I. This suggests studying "leading terms" in $F[x_1, x_2, \ldots, x_n]$ more generally (and somewhat more intrinsically). To do this we need to specify a total ordering on the monomials, since without some sort of ordering we cannot in general tell which is the "leading" term of a polynomial. We implicitly chose such an ordering in the inductive proof of Corollary 22—we first viewed a polynomial f as a polynomial in x_1 with coefficients in $R = F[x_2, \ldots, x_n]$, say, then viewed its "leading coefficient" in $F[x_2, \ldots, x_n]$ as a polynomial in x_2 with coefficients in $F[x_3, \ldots, x_n]$, etc. This is an example of a *lexicographic* monomial ordering on the polynomial ring $F[x_1, \ldots, x_n]$ which is defined by first declaring an ordering of the variables, for example $x_1 > x_2 > \cdots > x_n$ and then declaring that the monomial term $Ax_1^{a_1}x_2^{a_2}\cdots x_n^{a_n}$ with exponents $(a_1, a_2, \ldots, a_n)$ has higher order than the monomial term $Bx_1^{b_1}x_2^{b_2}\cdots x_n^{b_n}$ with exponents $(b_1, b_2, \ldots, b_n)$ if the first component where the n-tuples differ has $a_i > b_i$. This is analogous to the ordering used in a dictionary (hence the name), where the letter "a" comes before "b" which in turn comes before "c", etc., and then "aardvark" comes before "abacus" (although the 'word' $a^2 = aa$ comes before a in the lexicographical order). Note that the ordering is only defined up to multiplication by units (elements of $F^\times$) and that multiplying two monomials by the same nonzero monomial does not change their ordering. This can be formalized in general.

Definition. A *monomial ordering* is a well ordering "$\geq$" on the set of monomials that satisfies $mm_1 \geq mm_2$ whenever $m_1 \geq m_2$ for monomials m, m_1, m_2. Equivalently, a monomial ordering may be specified by defining a well ordering on the n-tuples $\alpha = (a_1, \ldots, a_n) \in \mathbb{Z}^n$ of multidegrees of monomials $Ax_1^{a_1} \cdots x_n^{a_n}$ that satisfies $\alpha + \gamma \geq \beta + \gamma$ if $\alpha \geq \beta$.

It is easy to show for any monomial ordering that $m \geq 1$ for every monomial m (cf. Exercise 2). It is not difficult to show, using Hilbert's Basis Theorem, that any total ordering on monomials which for every monomial m satisfies $m \geq 1$ and $mm_1 \geq mm_2$ whenever $m_1 \geq m_2$, is necessarily a well ordering (hence a monomial ordering)—this equivalent set of axioms for a monomial ordering may be easier to verify. For simplicity we shall limit the examples to the particularly easy and intuitive lexicographic ordering, but it is important to note that there are useful computational advantages to using other monomial orderings in practice. Some additional commonly used monomial orderings are introduced in the exercises.

As mentioned, once we have a monomial ordering we can define the leading term of a polynomial:

Definition. Fix a monomial ordering on the polynomial ring $F[x_1, x_2, \ldots, x_n]$.

(1) The *leading term* of a nonzero polynomial f in $F[x_1, x_2, \ldots, x_n]$, denoted $LT(f)$, is the monomial term of maximal order in f and the leading term of $f = 0$ is 0. Define the *multidegree of* f, denoted $\partial(f)$, to be the multidegree of the leading term of f.

(2) If I is an ideal in $F[x_1, x_2, \ldots, x_n]$, the *ideal of leading terms*, denoted $LT(I)$, is the ideal generated by the leading terms of all the elements in the ideal, i.e., $LT(I) = (LT(f) \mid f \in I)$.

The leading term and the multidegree of a polynomial clearly depend on the choice of the ordering. For example $LT(2xy + y^3) = 2xy$ with multidegree $(1, 1)$ if $x > y$, but $LT(2xy + y^3) = y^3$ with multidegree $(0, 3)$ if $y > x$. In particular, the leading term of a polynomial need not be the term of largest total degree. Similarly, the ideal of leading terms $LT(I)$ of an ideal I in general depends on the ordering used. Note also that the multidegree of a polynomial satisfies $\partial(fg) = \partial f + \partial g$ when f and g are nonzero, and that in this case $LT(fg) = LT(f) + LT(g)$ (cf. Exercise 2).

The ideal $LT(I)$ is by definition generated by monomials. Such ideals are called *monomial ideals* and are typically much easier to work with than generic ideals. For example, a polynomial is contained in a monomial ideal if and only if each of its monomial terms is a multiple of one of the generators for the ideal (cf. Exercise 10).

It was important in the proof of Hilbert's Basis Theorem to have *all* of the leading terms of the ideal I. If $I = (f_1, \ldots, f_m)$, then $LT(I)$ contains the leading terms $LT(f_1), \ldots, LT(f_m)$ of the generators for I by definition. Since $LT(I)$ is an ideal, it contains the ideal generated by these leading terms:

$$(LT(f_1), \ldots, LT(f_m)) \subseteq LT(I).$$

The first of the following examples shows that the ideal $LT(I)$ of leading terms can in general be strictly larger than the ideal generated just by the leading terms of some generators for I.

Examples

(1) Choose the lexicographic ordering $x > y$ on $F[x, y]$. The leading terms of the polynomials $f_1 = x^3y - xy^2 + 1$ and $f_2 = x^2y^2 - y^3 - 1$ are $LT(f_1) = x^3y$ (so the multidegree of f_1 is $\partial(f_1) = (3, 1)$) and $LT(f_2) = x^2y^2$ (so $\partial(f_2) = (2, 2)$). If $I = (f_1, f_2)$ is the ideal generated by f_1 and f_2 then the leading term ideal $LT(I)$ contains $LT(f_1) = x^3y$ and $LT(f_2) = x^2y^2$, so $(x^3y, x^2y^2) \subseteq LT(I)$. Since

$$yf_1 - xf_2 = y(x^3y - xy^2 + 1) - x(x^2y^2 - y^3 - 1) = x + y$$

we see that $g = x + y$ is an element of I and so the ideal $LT(I)$ also contains the leading term $LT(g) = x$. This shows that $LT(I)$ is strictly larger than $(LT(f_1), LT(f_2))$, since every element in $(LT(f_1), LT(f_2)) = (x^3y, x^2y^2)$ has total degree at least 4. We shall see later that in this case $LT(I) = (x, y^4)$.

(2) With respect to the lexicographic ordering $y > x$, the leading terms of f_1 and f_2 in the previous example are $LT(f_1) = -xy^2$ (which one could write as $-y^2x$ to emphasize the chosen ordering) and $LT(f_2) = -y^3$. We shall see later that in this ordering $LT(I) = (x^4, y)$, which is a different ideal than the ideal $LT(I)$ obtained in the previous example using the ordering $x > y$, and is again strictly larger than $(LT(f_1), LT(f_2))$.

(3) Choose any ordering on $F[x, y]$ and let $f = f(x, y)$ be any nonzero polynomial. The leading term of every element of the principal ideal $I = (f)$ is then a multiple of the leading term of f, so in this case $LT(I) = (LT(f))$.

In the case of one variable, leading terms are used in the Division Algorithm to reduce one polynomial g modulo another polynomial f to get a unique remainder r, and this remainder is 0 if and only if g is contained in the ideal (f). Since $F[x_1, x_2, \ldots, x_n]$ is not a Euclidean Domain if $n \geq 2$ (since it is not a P.I.D.), the situation is more complicated for polynomials in more than one variable. In the first example above, neither f_1 nor f_2 divides g in $F[x, y]$ (by degree considerations, for example), so attempting to first divide g by one of f_1 or f_2 and then by the other to try to reduce g modulo the ideal I would produce a (nonzero) "remainder" of g itself. In particular, this would suggest that $g = yf_1 - xf_2$ is not an element of the ideal I even though it is. The reason the polynomial g of degree 1 can be a linear combination of the two polynomials f_1 and f_2 of degree 4 is that the leading terms in yf_1 and xf_2 cancel in the difference, and this is reflected in the fact that $LT(f_1)$ and $LT(f_2)$ are not sufficient to generate $LT(I)$. A set of generators for an ideal I in $F[x_1, \ldots, x_n]$ whose leading terms generate the leading terms of *all* the elements in I is given a special name.

Definition. A *Gröbner basis* for an ideal I in the polynomial ring $F[x_1, \ldots, x_n]$ is a finite set of generators $\{g_1, \ldots, g_m\}$ for I whose leading terms generate the ideal of all leading terms in I, i.e.,

$$I = (g_1, \ldots, g_m) \quad \text{and} \quad LT(I) = (LT(g_1), \ldots, LT(g_m)).$$

Remark: Note that a Gröbner "basis" is in fact a set of *generators* for I (that depends on the choice of ordering), i.e., every element in I is a linear combination of the generators, and not a basis in the sense of vector spaces (where the linear combination would be *unique*, cf. Sections 10.3 and 11.1). Although potentially misleading, the terminology "Gröbner basis" has been so widely adopted that it would be hazardous to introduce a different nomenclature.

One of the most important properties of a Gröbner basis (proved in Theorem 23 following) is that every polynomial g can be written *uniquely* as the sum of an element in I and a remainder r obtained by a general polynomial division. In particular, we shall see that g is an element of I if and only if this remainder r is 0. While there is a similar decomposition in general, we shall see that if we do not use a Gröbner basis the uniqueness is lost (and we cannot detect membership in I by checking whether the remainder is 0) because there are leading terms not accounted for by the leading terms of the generators.

We first use the leading terms of polynomials defined by a monomial ordering on $F[x_1, \ldots, x_n]$ to extend the one variable Division Algorithm to a noncanonical polynomial division in several variables. Recall that for polynomials in one variable, the usual Division Algorithm determines the quotient $q(x)$ and remainder $r(x)$ in the equation $f(x) = q(x)g(x) + r(x)$ by successively testing whether the leading term of the dividend $f(x)$ is divisible by the leading term of $g(x)$: if $LT(f) = a(x)LT(g)$, the monomial term $a(x)$ is added to the quotient and the process is iterated with $f(x)$ replaced by the dividend $f(x) - a(x)g(x)$, which is of smaller degree since the leading terms cancel (by the choice of $a(x)$). The process terminates when the leading term of the divisor $g(x)$ no longer divides the leading term of the dividend, leaving the remainder $r(x)$. We can extend this to division by a finite number of polynomials in several variables simply by allowing successive divisions, resulting in a remainder and several quotients, as follows.

General Polynomial Division

Fix a monomial ordering on $F[x_1, \ldots, x_n]$, and suppose $g_1, \ldots, g_m$ is a set of nonzero polynomials in $F[x_1, \ldots, x_n]$. If f is any polynomial in $F[x_1, \ldots, x_n]$, start with a set of quotients $q_1, \ldots, q_m$ and a remainder r initially all equal to 0 and successively test whether the leading term of the dividend f is divisible by the leading terms of the divisors $g_1, \ldots, g_m$, in that order. Then

i. If $LT(f)$ is divisible by $LT(g_i)$, say, $LT(f) = a_i LT(g_i)$, add a_i to the quotient q_i, replace f by the dividend $f - a_i g_i$ (a polynomial with lower order leading term), and reiterate the entire process.

ii. If the leading term of the dividend f is not divisible by any of the leading terms $LT(g_1), \ldots, LT(g_m)$, add the leading term of f to the remainder r, replace f by the dividend $f - LT(f)$ (i.e., remove the leading term of f), and reiterate the entire process.

The process terminates (cf. Exercise 3) when the dividend is 0 and results in a set of quotients $q_1, \ldots, q_m$ and a remainder r with

$$f = q_1 g_1 + \cdots + q_m g_m + r.$$

Each $q_i g_i$ has multidegree less than or equal to the multidegree of f and the remainder r has the property that no nonzero term in r is divisible by any of the leading terms $LT(g_1), \ldots, LT(g_m)$ (since only terms with this property are added to r in (ii)).

Examples

Fix the lexicographic ordering $x > y$ on $F[x, y]$.

(1) Suppose $f = x^3y^3 + 3x^2y^4$ and $g = xy^4$. The leading term of f is x^3y^3, which is not divisible by (the leading term of) g, so x^3y^3 is added to the remainder r (so now $r = x^3y^3$) and f is replaced by $f - LT(f) = 3x^2y^4$ and we start over. Since $3x^2y^4$ is divisible by $LT(g) = xy^4$, with quotient $a = 3x$, we add $3x$ to the quotient q (so $q = 3x$), and replace $3x^2y^4$ by $3x^2y^4 - aLT(g) = 0$, at which point the process terminates. The result is the quotient $q = 3x$ and remainder $r = x^3y^3$ and

$$x^3y^3 + 3x^2y^4 = f = qg + r = (3x)(xy^4) + x^3y^3.$$

Note that if we had terminated at the first step because the leading term of f is not divisible by the leading term of g (which terminates the Division Algorithm for polynomials in one variable), then we would have been left with a 'remainder' of f itself, even though 'more' of f is divisible by g. This is the reason for step 2 in the division process (which is not necessary for polynomials in one variable).

(2) Let $f = x^2+x-y^2+y$, and suppose $g_1 = xy+1$ and $g_2 = x+y$. In the first iteration, the leading term x^2 of f is not divisible by the leading term of g_1, but is divisible by the leading term of g_2, so the quotient q_2 is x and the dividend f is replaced by the dividend $f - xg_2 = -xy + x - y^2 + y$. In the second iteration, the leading term of $-xy + x - y^2 + y$ is divisible by $LT(g_1)$, with quotient -1, so $q_1 = -1$ and the dividend is replaced by $(-xy + x - y^2 + y) - (-1)g_1 = x - y^2 + y + 1$. In the third iteration, the leading term of $x - y^2 + y + 1$ is not divisible by the leading term of g_1, but is divisible by the leading term of g_2, with quotient 1, so 1 is added to q_2 (which is now $q_2 = x+1$) and the dividend becomes $(x - y^2 + y + 1) - (1)(g_2) = -y^2 + 1$. The leading term is now $-y^2$, which is not divisible by either $LT(g_1) = xy$ or $LT(g_2) = x$, so $-y^2$ is added to the remainder r (which is now $-y^2$) and the dividend becomes simply 1. Finally, 1 is not divisible by either $LT(g_1)$ or $LT(g_2)$, so is added to the remainder (so r is now $-y^2 + 1$), and the process terminates. The result is

$$q_1 = -1, \qquad q_2 = x + 1, \qquad r = -y^2 + 1 \quad \text{and}$$

$$\begin{aligned} f = x^2 + x - y^2 + y &= (-1)(xy + 1) + (x + 1)(x + y) + (-y^2 + 1) \\ &= q_1 g_1 + q_2 g_2 + r. \end{aligned}$$

(3) Let $f = x^2 + x - y^2 + y$ as in the previous example and interchange the divisors g_1 and g_2: $g_1 = x + y$ and $g_2 = xy + 1$. In this case an easy computation gives

$$q_1 = x - y + 1, \qquad q_2 = 0, \qquad r = 0 \quad \text{and}$$

$$f = x^2 + x - y^2 + y = (x - y + 1)(x + y) = q_1 g_1 + q_2 g_2 + r,$$

showing that the quotients q_i and the remainder r are in general not unique and depend on the order of the divisors $g_1, \ldots, g_m$.

The computation in Example 3 shows that the polynomial $f = x^2 + x - y^2 + y$ is an element of the ideal $I = (x + y, xy + 1)$ since the remainder obtained in this case was 0 (in fact f is just a multiple of the first generator). In Example 2, however, the same polynomial resulted in a nonzero remainder $-y^2 + 1$ when divided by $xy + 1$ and $x + y$, and it was not at all clear from that computation that f was an element of I.

The next theorem shows that if we use a Gröbner basis for the ideal I then these difficulties do not arise: we obtain a *unique* remainder, which in turn can be used to determine whether a polynomial f is an element of the ideal I.

Theorem 23. Fix a monomial ordering on $R = F[x_1, \ldots, x_n]$ and suppose $\{g_1, \ldots, g_m\}$ is a Gröbner basis for the nonzero ideal I in R. Then

(1) Every polynomial $f \in R$ can be written uniquely in the form

$$f = f_I + r$$

where $f_I \in I$ and no nonzero monomial term of the 'remainder' r is divisible by any of the leading terms $LT(g_1), \ldots, LT(g_m)$.

(2) Both f_I and r can be computed by general polynomial division by $g_1, \ldots, g_m$ and are independent of the order in which these polynomials are used in the division.

(3) The remainder r provides a unique representative for the coset of f in the quotient ring $F[x_1, \ldots, x_n]/I$. In particular, $f \in I$ if and only if $r = 0$.

Proof: Letting $f_I = \sum_{i=1}^m q_i g_i \in I$ in the general polynomial division of f by $g_1, \ldots, g_m$ immediately gives a decomposition $f = f_I + r$ for any generators $g_1, \ldots, g_m$. Suppose now that $\{g_1, \ldots, g_m\}$ is a Gröbner basis, and $f = f_I + r = f_I' + r'$. Then $r - r' = f_I' - f_I \in I$, so its leading term $LT(r - r')$ is an element of $LT(I)$, which is the ideal $(LT(g_1), \ldots, LT(g_m))$ since $\{g_1, \ldots, g_m\}$ is a Gröbner basis for I. Every element in this ideal is a sum of multiples of the monomial terms $LT(g_1), \ldots, LT(g_m)$, so is a sum of terms each of which is divisible by one of the $LT(g_i)$. But both r and r', hence also $r - r'$, are sums of monomial terms none of which is divisible by $LT(g_1), \ldots, LT(g_m)$, which is a contradiction unless $r - r' = 0$. It follows that $r = r'$ is unique, hence so is $f_I = f - r$, which proves (1).

We have already seen that f_I and r can be computed algorithmically by polynomial division, and the uniqueness in (1) implies that r is independent of the order in which the polynomials $g_1, \ldots, g_m$ are used in the division. Similarly $f_I = \sum_{i=1}^m q_i g_i$ is uniquely determined (even though the individual quotients q_i are not in general unique), which gives (2).

The first statement in (3) is immediate from the uniqueness in (1). If $r = 0$, then $f = f_I \in I$. Conversely, if $f \in I$, then $f = f + 0$ together with the uniqueness of r implies that $r = 0$, and the final statement of the theorem follows.

As previously mentioned, the importance of Theorem 23, and one of the principal uses of Gröbner bases, is the uniqueness of the representative r, which allows effective computation in the quotient ring $F[x_1, \ldots, x_n]/I$.

We next prove that a set of polynomials in an ideal whose leading terms generate all the leading terms of an ideal is in fact a set of generators for the ideal itself (and so is a Gröbner basis—in some works this is taken as the definition of a Gröbner basis), and this shows in particular that a Gröbner basis always exists.

Proposition 24. Fix a monomial ordering on $R = F[x_1, \ldots, x_n]$ and let I be a nonzero ideal in R.

(1) If $g_1, \ldots, g_m$ are any elements of I such that $LT(I) = (LT(g_1), \ldots, LT(g_m))$, then $\{g_1, \ldots, g_m\}$ is a Gröbner basis for I.

(2) The ideal I has a Gröbner basis.

Proof: Suppose $g_1, \ldots, g_m \in I$ with $LT(I) = (LT(g_1), \ldots, LT(g_m))$. We need to see that $g_1, \ldots, g_m$ generate the ideal I. If $f \in I$, use general polynomial division to write $f = \sum_{i=1}^m q_i g_i + r$ where no nonzero term in the remainder r is divisible by any $LT(g_i)$. Since $f \in I$, also $r \in I$, which means $LT(r)$ is in $LT(I)$. But then $LT(r)$ would be divisible by one of $LT(g_1), \ldots, LT(g_m)$, which is a contradiction unless $r = 0$. Hence $f = \sum_{i=1}^m q_i g_i$ and $g_1, \ldots, g_m$ generate I, so are a Gröbner basis for I, which proves (1).

For (2), note that the ideal $LT(I)$ of leading terms of any ideal I is a monomial ideal generated by all the leading terms of the polynomials in I. By Exercise 1 a finite number of those leading terms suffice to generate $LT(I)$, say $LT(I) = (LT(h_1), \ldots, LT(h_k))$ for some $h_1, \ldots, h_k \in I$. By (1), the polynomials $h_1, \ldots, h_k$ are a Gröbner basis of I, completing the proof.

Proposition 24 proves that Gröbner bases always exist. We next prove a criterion that determines whether a given set of generators of an ideal I is a Gröbner basis, which we then use to provide an algorithm to find a Gröbner basis. The basic idea is very simple: additional elements in $LT(I)$ can arise by taking linear combinations of generators that cancel leading terms, as we saw in taking $yf_1 - xf_2$ in the first example in this section. We shall see that obtaining new leading terms from generators in this simple manner is the only obstruction to a set of generators being a Gröbner basis.

In general, if f_1, f_2 are two polynomials in $F[x_1, \ldots, x_n]$ and M is the monic least common multiple of the monomial terms $LT(f_1)$ and $LT(f_2)$ then we can cancel the leading terms by taking the difference

$$S(f_1, f_2) = \frac{M}{LT(f_1)} f_1 - \frac{M}{LT(f_2)} f_2. \tag{9.1}$$

The next lemma shows that these elementary linear combinations account for all cancellation in leading terms of polynomials of the same multidegree.

Lemma 25. Suppose $f_1, \ldots, f_m \in F[x_1, \ldots, x_n]$ are polynomials with the same multidegree α and that the linear combination $h = a_1 f_1 + \cdots + a_m f_m$ with constants $a_i \in F$ has strictly smaller multidegree. Then

$$h = \sum_{i=2}^{m} b_i S(f_{i-1}, f_i), \quad \text{for some constants } b_i \in F.$$

Proof: Write $f_i = c_i f_i'$ where $c_i \in F$ and f_i' is a monic polynomial of multidegree α. We have

$$h = \sum a_i c_i f_i' = a_1 c_1 (f_1' - f_2') + (a_1 c_1 + a_2 c_2)(f_2' - f_3') + \cdots$$
$$+ (a_1 c_1 + \cdots + a_{m-1} c_{m-1})(f_{m-1}' - f_m') + (a_1 c_1 + \cdots + a_m c_m) f_m'.$$

Note that $f_{i-1}' - f_i' = S(f_{i-1}, f_i)$. Then since h and each $f_{i-1}' - f_i'$ has multidegree strictly smaller than α, we have $a_1 c_1 + \cdots + a_m c_m = 0$, so the last term on the right hand side is 0 and the lemma follows.

The next proposition shows that a set of generators $g_1, \ldots, g_m$ is a Gröbner basis if there are no new leading terms among the differences $S(g_i, g_j)$ not already accounted for by the g_i. This result provides the principal ingredient in an algorithm to construct a Gröbner basis.

For a fixed monomial ordering on $R = F[x_1, \ldots, x_n]$ and ordered set of polynomials $G = \{g_1, \ldots, g_m\}$ in R, write $f \equiv r \bmod G$ if r is the remainder obtained by general polynomial division of $f \in R$ by $g_1, \ldots, g_m$ (in that order).

Proposition 26. *(Buchberger's Criterion)* Let $R = F[x_1, \ldots, x_n]$ and fix a monomial ordering on R. If $I = (g_1, \ldots, g_m)$ is a nonzero ideal in R, then $G = \{g_1, \ldots, g_m\}$ is a Gröbner basis for I if and only if $S(g_i, g_j) \equiv 0 \bmod G$ for $1 \leq i < j \leq m$.

Proof: If $\{g_1, \ldots, g_m\}$ is a Gröbner basis for I, then $S(g_i, g_j) \equiv 0 \bmod G$ by Theorem 23 since each $S(g_i, g_j)$ is an element of I.

Suppose now that $S(g_i, g_j) \equiv 0 \bmod G$ for $1 \leq i < j \leq m$ and take any element $f \in I$. To see that G is a Gröbner basis we need to see that $(LT(g_1), \ldots, LT(g_m))$ contains $LT(f)$. Since $f \in I$, we can write $f = \sum_{i=1}^{m} h_i g_i$ for some polynomials $h_1, \ldots, h_m$. Such a representation is not unique. Among all such representations choose one for which the largest multidegree of any summand (i.e., $\max_{i=1,\ldots,m} \partial(h_i g_i)$) is minimal, say α. It is clear that the multidegree of f is no worse than the largest multidegree of all the summands $h_i g_i$, so $\partial(f) \leq \alpha$. Write

$$\begin{aligned} f = \sum_{i=1}^{m} h_i g_i &= \sum_{\partial(h_i g_i)=\alpha} h_i g_i + \sum_{\partial(h_i g_i)<\alpha} h_i g_i \\ &= \sum_{\partial(h_i g_i)=\alpha} LT(h_i) g_i + \sum_{\partial(h_i g_i)=\alpha} (h_i - LT(h_i)) g_i + \sum_{\partial(h_i g_i)<\alpha} h_i g_i. \end{aligned} \tag{9.2}$$

Suppose that $\partial(f) < \alpha$. Then since the multidegree of the second two sums is also strictly smaller than α it follows that the multidegree of the first sum is strictly smaller than α. If $a_i \in F$ denotes the constant coefficient of the monomial term $LT(h_i)$ then $LT(h_i) = a_i h_i'$ where h_i' is a monomial. We can apply Lemma 25 to $\sum a_i(h_i' g_i)$ to write the first sum above as $\sum b_i S(h_{i-1}' g_{i-1}, h_i' g_i)$ with $\partial(h_{i-1}' g_{i-1}) = \partial(h_i' g_i) = \alpha$. Let $\beta_{i-1,i}$ be the multidegree of the monic least common multiple of $LT(g_{i-1})$ and $LT(g_i)$. Then an easy computation shows that $S(h_{i-1}' g_{i-1}, h_i' g_i)$ is just $S(g_{i-1}, g_i)$ multiplied by the monomial of multidegree $\alpha - \beta_{i-1,i}$. The polynomial $S(g_{i-1}, g_i)$ has multidegree less than $\beta_{i-1,i}$ and, by assumption, $S(g_{i-1}, g_i) \equiv 0 \bmod G$. This means that after general polynomial division of $S(g_{i-1}, g_i)$ by $g_1, \ldots, g_m$, each $S(g_{i-1}, g_i)$ can be written as a sum $\sum q_j g_j$ with $\partial(q_j g_j) < \beta_{i-1,i}$. It follows that each $S(h_{i-1}' g_{i-1}, h_i' g_i)$ is a sum $\sum q_j' g_j$ with $\partial(q_j' g_j) < \alpha$. But then all the sums on the right hand side of equation (2) can be written as a sum of terms of the form $p_i g_i$ with polynomials p_i satisfying $\partial(p_i g_i) < \alpha$. This contradicts the minimality of α and shows that in fact $\partial(f) = \alpha$, i.e., the leading term of f has multidegree α.

If we now take the terms in equation (2) of multidegree α we see that

$$LT(f) = \sum_{\partial(h_i g_i)=\alpha} LT(h_i) LT(g_i),$$

so indeed $LT(f) \in (LT(g_1), \ldots, LT(g_m))$. It follows that $G = \{g_1, \ldots, g_m\}$ is a Gröbner basis.

Buchberger's Algorithm

Buchberger's Criterion can be used to provide an algorithm to find a Gröbner basis for an ideal I, as follows. If $I = (g_1, \ldots, g_m)$ and each $S(g_i, g_j)$ leaves a remainder of 0 when divided by $G = \{g_1, \ldots, g_m\}$ using general polynomial division then G

is a Gröbner basis. Otherwise $S(g_i, g_j)$ has a nonzero remainder r. Increase G by appending the polynomial $g_{m+1} = r$: $G' = \{g_1, \ldots, g_m, g_{m+1}\}$ and begin again (note that this is again a set of generators for I since $g_{m+1} \in I$). It is not hard to check that this procedure terminates after a finite number of steps in a generating set G that satisfies Buchberger's Criterion, hence is a Gröbner basis for I (cf. Exercise 16). Note that once an $S(g_i, g_j)$ yields a remainder of 0 after division by the polynomials in G it also yields a remainder of 0 when additional polynomials are appended to G.

If $\{g_1, \ldots, g_m\}$ is a Gröbner basis for the ideal I and $LT(g_j)$ is divisible by $LT(g_i)$ for some $j \neq i$, then $LT(g_j)$ is not needed as a generator for $LT(I)$. By Proposition 24 we may therefore delete g_j and still retain a Gröbner basis for I. We may also assume without loss that the leading term of each g_i is monic. A Gröbner basis $\{g_1, \ldots, g_m\}$ for I where each $LT(g_i)$ is monic and where $LT(g_j)$ is not divisible by $LT(g_i)$ for $i \neq j$ is called a *minimal Gröbner basis.* While a minimal Gröbner basis is not unique, the number of elements and their leading terms are unique (cf. Exercise 15).

Examples

(1) Choose the lexicographic ordering $x > y$ on $F[x, y]$ and consider the ideal I generated by $f_1 = x^3y - xy^2 + 1$ and $f_2 = x^2y^2 - y^3 - 1$ as in Example 1 at the beginning of this section. To test whether $G = \{f_1, f_2\}$ is a Gröbner basis we compute $S(f_1, f_2) = yf_1 - xf_2 = x + y$, which is its own remainder when divided by $\{f_1, f_2\}$, so G is not a Gröbner basis for I. Set $f_3 = x + y$, and increase the generating set: $G' = \{f_1, f_2, f_3\}$. Now $S(f_1, f_2) \equiv 0 \bmod G'$, and a brief computation yields

$$S(f_1, f_3) = f_1 - x^2yf_3 = -x^2y^2 - xy^2 + 1 \equiv 0 \bmod G'$$

$$S(f_2, f_3) = f_2 - xy^2f_3 = -xy^3 - y^3 - 1 \equiv y^4 - y^3 - 1 \bmod G'.$$

Let $f_4 = y^4 - y^3 - 1$ and increase the generating set to $G'' = \{f_1, f_2, f_3, f_4\}$. The previous 0 remainder is still 0, and now $S(f_2, f_3) \equiv 0 \bmod G''$ by the choice of f_4. Some additional computation yields

$$S(f_1, f_4) \equiv S(f_2, f_4) \equiv S(f_3, f_4) \equiv 0 \bmod G''$$

and so $\{x^3y - xy^2 + 1, x^2y^2 - y^3 - 1, x + y, y^4 - y^3 - 1\}$ is a Gröbner basis for I. In particular, $LT(I)$ is generated by the leading terms of these four polynomials, so $LT(I) = (x^3y, x^2y^2, x, y^4) = (x, y^4)$, as previously mentioned. Then $x + y$ and $y^4 - y^3 - 1$ in I have leading terms generating $LT(I)$, so by Proposition 24, $\{x + y, y^4 - y^3 - 1\}$ gives a minimal Gröbner basis for I:

$$I = (x + y, y^4 - y^3 - 1).$$

This description of I is much simpler than $I = (x^3y - xy^2 + 1, x^2y^2 - y^3 - 1)$.

(2) Choose the lexicographic ordering $y > x$ on $F[x, y]$ and consider the ideal I in the previous example. In this case, $S(f_1, f_2)$ produces a remainder of $f_3 = -x - y$; then $S(f_1, f_3)$ produces a remainder of $f_4 = -x^4 - x^3 + 1$, and then all remainders are 0 with respect to the Gröbner basis $\{x^3y - xy^2 + 1, x^2y^2 - y^3 - 1, -x - y, -x^4 - x^3 + 1\}$. Here $LT(I) = (-xy^2, -y^3, -y, -x^4) = (y, x^4)$, as previously mentioned, and $\{x + y, x^4 + x^3 - 1\}$ gives a minimal Gröbner basis for I with respect to this ordering:

$$I = (x + y, x^4 + x^3 - 1),$$

a different simpler description of I.

In Example 1 above it is easy to check that $\{x+y^4-y^3+y-1,\ y^4-y^3-1\}$ is again a minimal Gröbner basis for I (this is just $\{f_3+f_4,\ f_4\}$), so even with a fixed monomial ordering on $F[x_1, \ldots, x_n]$ a minimal Gröbner basis for an ideal I is not unique. We can obtain an important uniqueness property by strengthening the condition on divisibility by the leading terms of the basis.

Definition. Fix a monomial ordering on $R = F[x_1, \ldots, x_n]$. A Gröbner basis $\{g_1, \ldots, g_m\}$ for the nonzero ideal I in R is called a *reduced Gröbner basis* if

(a) each g_i has monic leading term, i.e., $LT(g_i)$ is monic, $i = 1, \ldots, m$, and

(b) no term in g_j is divisible by $LT(g_i)$ for $j \neq i$.

Note that a reduced Gröbner basis is, in particular, a minimal Gröbner basis. If $G = \{g_1, \ldots, g_m\}$ is a minimal Gröbner basis for I, then the leading term $LT(g_j)$ is not divisible by $LT(g_i)$ for any $i \neq j$. As a result, if we use polynomial division to divide g_j by the other polynomials in G we obtain a remainder g_j' in the ideal I with the same leading term as g_j (the remainder g_j' does not depend on the order of the polynomials used in the division by (2) of Theorem 23). By Proposition 24, replacing g_j by g_j' in G again gives a minimal Gröbner basis for I, and in this basis no term of g_j' is divisible by $LT(g_i)$ for any $i \neq j$. Replacing each element in G by its remainder after division by the other elements in G therefore results in a reduced Gröbner basis for I. The importance of reduced Gröbner bases is that they are unique (for a given monomial ordering), as the next result shows.

Theorem 27. Fix a monomial ordering on $R = F[x_1, \ldots, x_n]$. Then there is a unique reduced Gröbner basis for every nonzero ideal I in R.

Proof: By Exercise 15, two reduced bases have the same number of elements and the same leading terms since reduced bases are also minimal bases. If $G = \{g_1, \ldots, g_m\}$ and $G' = \{g_1', \ldots, g_m'\}$ are two reduced bases for the same nonzero ideal I, then after a possible rearrangement we may assume $LT(g_i) = LT(g_i') = h_i$ for $i = 1, \ldots, m$. For any fixed i, consider the polynomial $f_i = g_i - g_i'$. If f_i is nonzero, then since $f_i \in I$, its leading term must be divisible by some h_j. By definition of a reduced basis, h_j for $j \neq i$ does not divide any of the terms in either g_i or g_i', hence does not divide $LT(f_i)$. But h_i also does not divide $LT(f_i)$ since all the terms in f_i have strictly smaller multidegree. This forces $f_i = 0$, i.e., $g_i = g_i'$ for every i, so $G = G'$.

One application of the uniqueness of the reduced Gröbner basis is a computational method to determine when two ideals in a polynomial ring are equal.

Corollary 28. Let I and J be two ideals in $F[x_1, \ldots, x_n]$. Then $I = J$ if and only if I and J have the same reduced Gröbner basis with respect to any fixed monomial ordering on $F[x_1, \ldots, x_n]$.

Examples

(1) Consider the ideal $I = (h_1, h_2, h_3)$ with $h_1 = x^2+xy^5+y^4$, $h_2 = xy^6-xy^3+y^5-y^2$, and $h_3 = xy^5 - xy^2$ in $F[x, y]$. Using the lexicographic ordering $x > y$ we find

$S(h_1, h_2) \equiv S(h_1, h_3) \equiv 0 \bmod \{h_1, h_2, h_3\}$ and $S(h_2, h_3) \equiv y^5 - y^2 \bmod \{h_1, h_2, h_3\}$. Setting $h_4 = y^5 - y^2$ we find $S(h_i, h_j) \equiv 0 \bmod \{h_1, h_2, h_3, h_4\}$ for $1 \leq i < j \leq 4$, so

$$x^2 + xy^5 + y^4, \quad xy^6 - xy^3 + y^5 - y^2, \quad xy^5 - xy^2, \quad y^5 - y^2$$

is a Gröbner basis for I. The leading terms of this basis are x^2, xy^6, xy^5, y^5. Since y^5 divides both xy^6 and xy^5, we may remove the second and third generators to obtain a minimal Gröbner basis $\{x^2 + xy^5 + y^4, y^5 - y^2\}$ for I. The second term in the first generator is divisible by the leading term y^5 of the second generator, so this is not a reduced Gröbner basis. Replacing $x^2 + xy^5 + y^4$ by its remainder $x^2 + xy^2 + y^4$ after division by the other polynomials in the basis (which in this case is only the polynomial $y^5 - y^2$), we are left with the reduced Gröbner basis $\{x^2 + xy^2 + y^4, y^5 - y^2\}$ for I.

(2) Consider the ideal $J = (h_1, h_2, h_3)$ with $h_1 = xy^3 + y^3 + 1$, $h_2 = x^3y - x^3 + 1$, and $h_3 = x + y$ in $F[x, y]$. Using the lexicographic monomial ordering $x > y$ we find $S(h_1, h_2) \equiv 0 \bmod \{h_1, h_2, h_3\}$ and $S(h_1, h_3) \equiv y^4 - y^3 - 1 \bmod \{h_1, h_2, h_3\}$. Setting $h_4 = y^4 - y^3 - 1$ we find $S(h_i, h_j) \equiv 0 \bmod \{h_1, h_2, h_3, h_4\}$ for $1 \leq i < j \leq 4$, so

$$xy^3 + y^3 + 1, \quad x^3y - x^3 + 1, \quad x + y, \quad y^4 - y^3 - 1$$

is a Gröbner basis for J. The leading terms of this basis are xy^3, x^3y, x, and y^4, so $\{x + y, y^4 - y^3 - 1\}$ is a minimal Gröbner basis for J. In this case none of the terms in $y^4 - y^3 - 1$ are divisible by the leading term of $x + y$ and none of the terms in $x + y$ are divisible by the leading term in $y^4 - y^3 - 1$, so $\{x + y, y^4 - y^3 - 1\}$ is the reduced Gröbner basis for J. This is the basis for the ideal I in Example 1 following Proposition 26, so these two ideals are equal:

$$(x^3y - xy^2 + 1, x^2y^2 - y^3 - 1) = (xy^3 + y^3 + 1, x^3y - x^3 + 1, x + y)$$

(and both are equal to the ideal $(x + y, y^4 - y^3 - 1)$).

Gröbner Bases and Solving Algebraic Equations: Elimination

The theory of Gröbner bases is very useful in explicitly solving systems of algebraic equations, and is the basis by which computer algebra programs attempt to solve systems of equations. Suppose $S = \{f_1, \ldots, f_m\}$ is a collection of polynomials in n variables $x_1, \ldots, x_n$ and we are trying to find the solutions of the system of equations $f_1 = 0$, $f_2 = 0, \ldots, f_m = 0$ (i.e., the common set of zeros of the polynomials in S). If $(a_1, \ldots, a_n)$ is any solution to this system, then every element f of the ideal I generated by S also satisfies $f(a_1, \ldots, a_n) = 0$. Furthermore, it is an easy exercise to see that if $S' = \{g_1, \ldots, g_s\}$ is *any* set of generators for the ideal I then the set of solutions to the system $g_1 = 0, \ldots, g_s = 0$ is the *same* as the original solution set.

In the situation where $f_1, \ldots, f_m$ are *linear* polynomials, a solution to the system of equations can be obtained by successively eliminating the variables $x_1, x_2, \ldots$ by elementary means—using linear combinations of the original equations to eliminate the variable x_1, then using these equations to eliminate x_2, etc., producing a system of equations that can be easily solved (this is "Gauss-Jordan elimination" in linear algebra, cf. the exercises in Section 11.2).

The situation for polynomial equations that are nonlinear is naturally more complicated, but the basic principle is the same. If there is a nonzero polynomial in the

ideal I involving only one of the variables, say $p(x_n)$, then the last coordinate a_n is a solution of $p(x_n) = 0$. If now there is a polynomial in I involving only x_{n-1} and x_n, say $q(x_{n-1}, x_n)$, then the coordinate a_{n-1} would be a solution of $q(x_{n-1}, a_n) = 0$, etc. If we can successively find polynomials in I that eliminate the variables $x_1, x_2, \ldots$ then we will be able to determine all the solutions $(a_1, \ldots, a_n)$ to our original system of equations explicitly.

Finding equations that follow from the system of equations in S, i.e., finding elements of the ideal I that do not involve some of the variables, is referred to as *elimination theory*. The polynomials in I that do not involve the variables $x_1, \ldots, x_i$, i.e., $I \cap F[x_{i+1}, \ldots, x_n]$, is easily seen to be an ideal in $F[x_{i+1}, \ldots, x_n]$ and is given a name.

Definition. If I is an ideal in $F[x_1, \ldots, x_n]$ then $I_i = I \cap F[x_{i+1}, \ldots, x_n]$ is called the i^{th} *elimination ideal* of I with respect to the ordering $x_1 > \cdots > x_n$.

The success of using elimination to solve a system of equations depends on being able to determine the elimination ideals (and, ultimately, on whether these elimination ideals are nonzero).

The following fundamental proposition shows that if the lexicographic monomial ordering $x_1 > \cdots > x_n$ is used to compute a Gröbner basis for I then the elements in the resulting basis not involving the variables $x_1, \ldots, x_i$ not only determine the i^{th} elimination ideal, but in fact give a Gröbner basis for the i^{th} elimination ideal of I.

Proposition 29. (*Elimination*) Suppose $G = \{g_1, \ldots, g_m\}$ is a Gröbner basis for the nonzero ideal I in $F[x_1, \ldots, x_n]$ with respect to the lexicographic monomial ordering $x_1 > \cdots > x_n$. Then $G \cap F[x_{i+1}, \ldots, x_n]$ is a Gröbner basis of the i^{th} elimination ideal $I_i = I \cap F[x_{i+1}, \ldots, x_n]$ of I. In particular, $I \cap F[x_{i+1}, \ldots, x_n] = 0$ if and only if $G \cap F[x_{i+1}, \ldots, x_n] = \emptyset$.

Proof: Denote $G_i = G \cap F[x_{i+1}, \ldots, x_n]$. Then $G_i \subseteq I_i$, so by Proposition 24, to see that G_i is a Gröbner basis of I_i it suffices to see that $LT(G_i)$, the leading terms of the elements in G_i, generate $LT(I_i)$ as an ideal in $F[x_{i+1}, \ldots, x_n]$. Certainly $(LT(G_i)) \subseteq LT(I_i)$ as ideals in $F[x_{i+1}, \ldots, x_n]$. To show the reverse containment, let f be any element in I_i. Then $f \in I$ and since G is a Gröbner basis for I we have

$$LT(f) = a_1(x_1, \ldots, x_n)LT(g_1) + \cdots + a_m(x_1, \ldots, x_n)LT(g_m)$$

for some polynomials $a_1, \ldots, a_m \in F[x_1, \ldots, x_n]$. Writing each polynomial a_i as a sum of monomial terms we see that $LT(f)$ is a sum of monomial terms of the form $ax_1^{s_1} \ldots x_n^{s_n} LT(g_i)$. Since $LT(f)$ involves only the variables $x_{i+1}, \ldots, x_n$, the sum of all such terms containing any of the variables $x_1, \ldots, x_i$ must be 0, so $LT(f)$ is also the sum of those monomial terms only involving $x_{i+1}, \ldots, x_n$. It follows that $LT(f)$ can be written as a $F[x_{i+1}, \ldots, x_n]$-linear combination of some monomial terms $LT(g_t)$ where $LT(g_t)$ does not involve the variables $x_1, \ldots, x_i$. But by the choice of the ordering, if $LT(g_t)$ does not involve $x_1, \ldots, x_i$, then neither do any of the other terms in g_t, i.e., $g_t \in G_i$. Hence $LT(f)$ can be written as a $F[x_{i+1}, \ldots, x_n]$-linear combination of elements $LT(G_i)$, completing the proof.

Note also that Gröbner bases can be used to eliminate any variables simply by using an appropriate monomial ordering.

Examples

(1) The ellipse $2x^2 + 2xy + y^2 - 2x - 2y = 0$ intersects the circle $x^2 + y^2 = 1$ in two points. To find them we compute a Gröbner basis for the ideal $I = (2x^2 + 2xy + y^2 - 2x - 2y, x^2 + y^2 - 1) \subset \mathbb{R}[x, y]$ using the lexicographic monomial order $x > y$ to eliminate x, obtaining $g_1 = 2x + y^2 + 5y^3 - 2$ and $g_2 = 5y^4 - 4y^3$. Hence $5y^4 = 4y^3$ and $y = 0$ or $y = 4/5$. Substituting these values into $g_1 = 0$ and solving for x we find the two intersection points are $(1, 0)$ and $(-3/5, 4/5)$.

Instead using the lexicographic monomial order $y > x$ to eliminate y results in the Gröbner basis $\{y^2 + x^2 - 1, 2yx - 2y + x^2 - 2x + 1, 5x^3 - 7x^2 - x + 3\}$. Then $5x^3 - 7x^2 - x + 3 = (x - 1)^2(5x + 3)$ shows that x is 1 or $-3/5$ and we obtain the same solutions as before, although with more effort.

(2) In the previous example the solutions could also have been found by elementary means. Consider now the solutions in $\mathbb{C}$ to the system of two equations

$$x^3 - 2xy + y^3 = 0 \quad \text{and} \quad x^5 - 2x^2y^2 + y^5 = 0.$$

Computing a Gröbner basis for the ideal generated by $f_1 = x^3 - 2xy + y^3$ and $f_2 = x^5 - 2x^2y^2 + y^5$ with respect to the lexicographic monomial order $x > y$ we obtain the basis

$$\begin{aligned} g_1 &= x^3 - 2xy + y^3 \\ g_2 &= 200xy^2 + 193y^9 + 158y^8 - 45y^7 - 456y^6 + 50y^5 - 100y^4 \\ g_3 &= y^{10} - y^8 - 2y^7 + 2y^6. \end{aligned}$$

Any solution to our original equations would satisfy $g_1 = g_2 = g_3 = 0$. Since $g_3 = y^6(y - 1)^2(y^2 + 2y + 2)$, we have $y = 0$, $y = 1$ or $y = -1 \pm i$. Since $g_1(x, 0) = x^3$ and $g_2(x, 0) = 0$, we see that $(0, 0)$ is the only solution with $y = 0$. Since $g_1(x, 1) = x^3 - 2x + 1$ and $g_2(x, 1) = 200(x - 1)$ have only $x = 1$ as a common zero, the only solution with $y = 1$ is $(1, 1)$. Finally,

$$\begin{aligned} g_1(x, -1 \pm i) &= x^3 + (2 \mp 2i)x + (2 \pm 2i) \\ g_2(x, -1 \pm i) &= -400i(x + 1 \pm i), \end{aligned}$$

and a quick check shows the common zero $x = -1 \mp i$ when $y = -1 \pm i$, respectively. Hence, there are precisely four solutions to the original pair of equations, namely

$$(x, y) = (0, 0), \quad (1, 1), \quad (-1 + i, -1 - i), \quad \text{or} \quad (-1 - i, -1 + i).$$

(3) Consider the solutions in $\mathbb{C}$ to the system of equations

$$\begin{aligned} x + y + z &= 1 \\ x^2 + y^2 + z^2 &= 2 \\ x^3 + y^3 + z^3 &= 3. \end{aligned}$$

The reduced Gröbner basis with respect to the lexicographic ordering $x > y > z$ is

$$\{x + y + z - 1, \ y^2 + yz - y + z^2 - z - (1/2), \ z^3 - z^2 - (1/2)z - (1/6)\}$$

and so z is a root of the polynomial $t^3 - t^2 - (1/2)t - (1/6)$ (by symmetry, also x and y are roots of this same polynomial). For each of the three roots of this polynomial, there are two values of y and one corresponding value of x making the first two polynomials in the Gröbner basis equal to 0. The resulting six solutions are quickly checked to be the three distinct roots of the polynomial $t^3 - t^2 - (1/2)t - (1/6)$ (which is irreducible over $\mathbb{Q}$) in some order.

As the previous examples show, the study of solutions to systems of polynomial equations $f_1 = 0, f_2 = 0, \ldots, f_m = 0$ is intimately related to the study of the ideal $I = (f_1, f_2, \ldots, f_m)$ the polynomials generate in $F[x_1, \ldots, x_n]$. This fundamental connection is the starting point for the important and active branch of mathematics called "algebraic geometry", introduced in Chapter 15, where additional applications of Gröbner bases are given.

We close this section by showing how to compute the basic set-theoretic operations of sums, products and intersections of ideals in polynomial rings. Suppose $I = (f_1, \ldots, f_s)$ and $J = (h_1, \ldots, h_t)$ are two ideals in $F[x_1, \ldots, x_n]$. Then $I + J = (f_1, \ldots, f_s, h_1, \ldots, h_t)$ and $IJ = (f_1h_1, \ldots, f_ih_j, \ldots, f_sh_t)$. The following proposition shows how to compute the intersection of any two ideals.

Proposition 30. If I and J are any two ideals in $F[x_1, \ldots, x_n]$ then $tI + (1 - t)J$ is an ideal in $F[t, x_1, \ldots, x_n]$ and $I \cap J = (tI + (1 - t)J) \cap F[x_1, \ldots, x_n]$. In particular, $I \cap J$ is the first elimination ideal of $tI + (1 - t)J$ with respect to the ordering $t > x_1 > \cdots > x_n$.

Proof: First, tI and $(1-t)J$ are clearly ideals in $F[x_1, \ldots, x_n, t]$, so also their sum $tI + (1-t)J$ is an ideal in $F[x_1, \ldots, x_n, t]$. If $f \in I \cap J$, then $f = tf + (1-t)f$ shows $I \cap J \subseteq (tI + (1-t)J) \cap F[x_1, \ldots, x_n]$. Conversely, suppose $f = tf_1 + (1-t)f_2$ is an element of $F[x_1, \ldots, x_n]$, where $f_1 \in I$ and $f_2 \in J$. Then $t(f_1 - f_2) = f - f_2 \in F[x_1, \ldots, x_n]$ shows that $f_1 - f_2 = 0$ and $f = f_2$, so $f = f_1 = f_2 \in I \cap J$. Since $I \cap J = (tI + (1 - t)J) \cap F[x_1, \ldots, x_n]$, $I \cap J$ is the first elimination ideal of $tI + (1 - t)J$ with respect to the ordering $t > x_1 > \cdots > x_n$.

We have $tI + (1-t)J = (tf_1, \ldots, tf_s, (1-t)h_1, \ldots, (1-t)h_t)$ if $I = (f_1, \ldots, f_s)$ and $J = (h_1, \ldots, h_t)$. By Proposition 29, the elements not involving t in a Gröbner basis for this ideal in $F[t, x_1, \ldots, x_n]$, computed for the lexicographic monomial ordering $t > x_1 > \cdots > x_n$, give a Gröbner basis for the ideal $I \cap J$ in $F[x_1, \ldots, x_n]$.

Example

Let $I = (x, y)^2 = (x^2, xy, y^2)$ and let $J = (x)$. For the lexicographic monomial ordering $t > x > y$ the reduced Gröbner basis for $tI + (1-t)J$ in $F[t, x, y]$ is $\{tx - x, ty^2, x^2, xy\}$ and so $I \cap J = (x^2, xy)$.

EXERCISES

1. Suppose I is an ideal in $F[x_1, \ldots, x_n]$ generated by a (possibly infinite) set $\mathcal{S}$ of polynomials. Prove that a finite subset of the polynomials in $\mathcal{S}$ suffice to generate I. [Use Theorem 21 to write $I = (f_1, \ldots, f_m)$ and then write each $f_i \in I$ using polynomials in $\mathcal{S}$.]
2. Let $\geq$ be any monomial ordering.
   (a) Prove that $LT(fg) = LT(f)LT(g)$ and $\partial(fg) = \partial(f) + \partial(g)$ for any nonzero polynomials f and g.
   (b) Prove that $\partial(f + g) \leq \max(\partial(f), \partial(g))$ with equality if $\partial(f) \neq \partial(g)$.

(c) Prove that $m \geq 1$ for every monomial m.

(d) Prove that if m_1 divides m_2 then $m_2 \geq m_1$. Deduce that the leading term of a polynomial does not divide any of its lower order terms.

3. Prove that if $\geq$ is any total or partial ordering on a nonempty set then the following are equivalent:

(i) Every nonempty subset contains a minimum element.

(ii) There is no infinite strictly decreasing sequence $a_1 > a_2 > a_3 > \cdots$ (this is called the *descending chain condition* or *D.C.C.*).

Deduce that General Polynomial Division always terminates in finitely many steps.

4. Let $\geq$ be a monomial ordering, and for monomials m_1, m_2 define $m_1 \geq_g m_2$ if either $\deg m_1 > \deg m_2$, or $\deg m_1 = \deg m_2$ and $m_1 \geq m_2$.

(a) Prove that $\geq_g$ is also a monomial ordering. (The relation $\geq_g$ is called the *grading* of $\geq$. An ordering in which the most important criterion for comparison is degree is sometimes called a *graded* or a *degree* ordering, so this exercise gives a method for constructing graded orderings.)

(b) The grading of the lexicographic ordering $x_1 > \cdots > x_n$ is called the *grlex* monomial ordering. Show that $x_2^4 > x_1^2x_2 > x_1x_2^2 > x_2^2 > x_1$ with respect to the grlex ordering and $x_1^2x_2 > x_1x_2^2 > x_1 > x_2^4 > x_2^2$ with respect to the lexicographic ordering.

5. The *grevlex* monomial ordering is defined by first choosing an ordering of the variables $\{x_1, x_2, \ldots, x_n\}$, then defining $m_1 \geq m_2$ for monomials m_1, m_2 if either $\deg m_1 > \deg m_2$ or $\deg m_1 = \deg m_2$ and the first exponent of $x_n, x_{n-1}, \ldots, x_1$ (in that order) where m_1 and m_2 differ is *smaller* in m_1.

(a) Prove that grevlex is a monomial ordering that satisfies $x_1 > x_2 > \cdots > x_n$.

(b) Prove that the grevlex ordering on $F[x_1, x_2]$ with respect to $\{x_1, x_2\}$ is the graded lexicographic ordering with $x_1 > x_2$, but that the grevlex ordering on $F[x_1, x_2, x_3]$ is not the grading of any lexicographic ordering.

(c) Show that $x_1x_2^2x_3 > x_1^2x_3^2 > x_2^2x_3^2 > x_2x_3^2 > x_1x_2 > x_2^2 > x_1x_3 > x_3^2 > x_1 > x_2$ for the grevlex monomial ordering with respect to $\{x_1, x_2, x_3\}$.

6. Show that $x^3y > x^3z^2 > x^3z > x^2y^2z > x^2y > xz^2 > y^2z^2 > y^2z$ with respect to the lexicographic monomial ordering $x > y > z$. Show that for the corresponding grlex monomial ordering $x^3z^2 > x^2y^2z > x^3y > x^3z > y^2z^2 > x^2y > xz^2 > y^2z$, and that $x^2y^2z > x^3z^2 > x^3y > x^3z > y^2z^2 > x^2y > y^2z > xz^2$ for the grevlex monomial ordering with respect to $\{x, y, z\}$.

7. Order the monomials $x^2z, x^2y^2z, xy^2z, x^3y, x^3z^2, x^2, x^2yz^2, x^2z^2$ for the lexicographic monomial ordering $x > y > z$, for the corresponding grlex monomial order, and for the grevlex monomial ordering with respect to $\{x, y, z\}$.

8. Show there are $n!$ distinct lexicographic monomial orderings on $F[x_1, \ldots, x_n]$. Show similarly that there are $n!$ distinct grlex and grevlex monomial orderings.

9. It can be shown that any monomial ordering on $F[x_1, \ldots, x_n]$ may be obtained as follows. For $k \leq n$ let $v_1, v_2, \ldots, v_k$ be nonzero vectors in Euclidean n-space, $\mathbb{R}^n$, that are pairwise orthogonal: $v_i \cdot v_j = 0$ for all $i \neq j$, where $\cdot$ is the usual dot product, and suppose also that all the coordinates of v_1 are nonnegative. Define an order, $\geq$, on monomials by $m_1 > m_2$ if and only if for some $t \leq k$ we have $v_i \cdot \partial(m_1) = v_i \cdot \partial(m_2)$ for all $i \in \{1, 2, \ldots, t-1\}$ and $v_t \cdot \partial(m_1) > v_t \cdot \partial(m_2)$.

(a) Let $k = n$ and let $v_i = (0, \ldots, 0, 1, 0, \ldots, 0)$ with 1 in the i^{th} position. Show that $\geq$ defines the lexicographic order with $x_1 > x_2 > \cdots > x_n$.

(b) Let $k = n$ and define $v_1 = (1, 1, \ldots, 1)$ and $v_i = (1, 1, \ldots, 1, -n+i-1, 0, \ldots, 0)$,

where there are $i-2$ trailing zeros, $2 \le i \le n$. Show that $\ge$ defines the grlex order with respect to $\{x_1, \ldots, x_n\}$.

10. Suppose I is a monomial ideal generated by monomials $m_1, \ldots, m_k$. Prove that the polynomial $f \in F[x_1, \ldots, x_n]$ is in I if and only if every monomial term f_i of f is a multiple of one of the m_j. [For polynomials $a_1, \ldots, a_k \in F[x_1, \ldots, x_n]$ expand the polynomial $a_1m_1 + \cdots + a_km_k$ and note that every monomial term is a multiple of at least one of the m_j.] Show that $x^2yz+3xy^2$ is an element of the ideal $I = (xyz, y^2) \subset F[x, y, z]$ but is not an element of the ideal $I' = (xz^2, y^2)$.

11. Fix a monomial ordering on $R = F[x_1, \ldots, x_n]$ and suppose $\{g_1, \ldots, g_m\}$ is a Gröbner basis for the ideal I in R. Prove that $h \in LT(I)$ if and only if h is a sum of monomial terms each divisible by some $LT(g_i)$, $1 \le i \le m$. [Use the previous exercise.]

12. Suppose I is a monomial ideal with monomial generators $g_1, \ldots, g_m$. Use the previous exercise to prove directly that $\{g_1, \ldots, g_m\}$ is a Gröbner basis for I.

13. Suppose I is a monomial ideal with monomial generators $g_1, \ldots, g_m$. Use Buchberger's Criterion to prove that $\{g_1, \ldots, g_m\}$ is a Gröbner basis for I.

14. Suppose I is a monomial ideal in $R = F[x_1, \ldots, x_n]$ and suppose $\{m_1, \ldots, m_k\}$ is a minimal set of monomials generating I, i.e., each m_i is a monomial and no proper subset of $\{m_1, \ldots, m_k\}$ generates I. Prove that the m_i, $1 \le i \le k$ are unique. [Use Exercise 10.]

15. Fix a monomial ordering on $R = F[x_1, \ldots, x_n]$.

(a) Prove that $\{g_1, \ldots, g_m\}$ is a minimal Gröbner basis for the ideal I in R if and only if $\{LT(g_1), \ldots, LT(g_m)\}$ is a minimal generating set for $LT(I)$.

(b) Prove that the leading terms of a minimal Gröbner basis for I are uniquely determined and the number of elements in any two minimal Gröbner bases for I is the same. [Use (a) and the previous exercise.]

16. Fix a monomial ordering on $F[x_1, \ldots, x_n]$ and suppose $G = \{g_1, \ldots, g_m\}$ is a set of generators for the nonzero ideal I. Show that if $S(g_i, g_j) \not\equiv 0 \bmod G$ then the ideal $(LT(g_1), \ldots, LT(g_m), LT(S(g_i, g_j))$ is strictly larger than the ideal $(LT(g_1), \ldots, LT(g_m))$. Conclude that the algorithm for computing a Gröbner basis described following Proposition 26 terminates after a finite number of steps. [Use Exercise 1.]

17. Fix the lexicographic ordering $x > y$ on $F[x, y]$. Use Buchberger's Criterion to show that $\{x^2y - y^2, x^3 - xy\}$ is a Gröbner basis for the ideal $I = (x^2y - y^2, x^3 - xy)$.

18. Show $\{x - y^3, y^5 - y^6\}$ is the reduced Gröbner basis for the ideal $I = (x - y^3, -x^2 + xy^2)$ with respect to the lexicographic ordering defined by $x > y$ in $F[x, y]$.

19. Fix the lexicographic ordering $x > y$ on $F[x, y]$.

(a) Show that $\{x^3 - y, x^2y - y^2, xy^2 - y^2, y^3 - y^2\}$ is the reduced Gröbner basis for the ideal $I = (-x^3 + y, x^2y - y^2)$.

(b) Determine whether the polynomial $f = x^6 - x^5y$ is an element of the ideal I.

20. Fix the lexicographic ordering $x > y > z$ on $F[x, y, z]$. Show that $\{x^2 + xy + z, xyz + z^2, xz^2, z^3\}$ is the reduced Gröbner basis for the ideal $I = (x^2 + xy + z, xyz + z^2)$ and in particular conclude that the leading term ideal $LT(I)$ requires four generators.

21. Fix the lexicographic ordering $x > y$ on $F[x, y]$. Use Buchberger's Criterion to show that $\{x^2y - y^2, x^3 - xy\}$ is a Gröbner basis for the ideal $I = (x^2y - y^2, x^3 - xy)$.

22. Let $I = (x^2 - y, x^2y - z)$ in $F[x, y, z]$.

(a) Show that $\{x^2 - y, y^2 - z\}$ is the reduced Gröbner basis for I with respect to the lexicographic ordering defined by $x > y > z$.

(b) Show that $\{x^2 - y, z - y^2\}$ is the reduced Gröbner basis for I with respect to the

lexicographic ordering defined by $z > x > y$ (note these are essentially the same polynomials as in (a)).

(c) Show that $\{y - x^2, z - x^4\}$ is the reduced Gröbner basis for I with respect to the lexicographic ordering defined by $z > y > x$.

23. Show that the ideals $I = (x^2y + xy^2 - 2y, x^2 + xy - x + y^2 - 2y, xy^2 - x - y + y^3)$ and $J = (x - y^2, xy - y, x^2 - y)$ in $F[x, y]$ are equal.

24. Use reduced Gröbner bases to show that the ideal $I = (x^3 - yz, yz + y)$ and the ideal $J = (x^3z + x^3, x^3 + y)$ in $F[x, y, z]$ are equal.

25. Show that the reduced Gröbner basis using the lexicographic ordering $x > y$ for the ideal $I = (x^2 + xy^2, x^2 - y^3, y^3 - y^2)$ is $\{x^2 - y^2, y^3 - y^2, xy^2 + y^2\}$.

26. Show that the reduced Gröbner basis for the ideal $I = (xy + y^2, x^2y + xy^2 + x^2)$ is $\{x^2, xy + y^2, y^3\}$ with respect to the lexicographic ordering $x > y$ and is $\{y^2 + yx, x^2\}$ with respect to the lexicographic ordering $y > x$.

There are generally substantial differences in computational complexity when using different monomial orders. The grevlex monomial ordering often provides the most efficient computation and produces simpler polynomials.

27. Show that $\{x^3 - y^3, x^2 + xy^2 + y^4, x^2y + xy^3 + y^2\}$ is a reduced Gröbner basis for the ideal I in the example following Corollary 28 with respect to the grlex monomial ordering. (Note that while this gives three generators for I rather than two for the lexicographic ordering as in the example, the degrees are smaller.)

28. Let $I = (x^4 - y^4 + z^3 - 1, x^3 + y^2 + z^2 - 1)$. Show that there are five elements in a reduced Gröbner basis for I with respect to the lexicographic ordering with $x > y > z$ (the maximum degree among the five generators is 12 and the maximum number of monomial terms among the five generators is 35), that there are two elements for the lexicographic ordering $y > z > x$ (maximum degree is 6 and maximum number of terms is 8), and that $\{x^3 + y^2 + z^2 - 1, xy^2 + xz^2 - x + y^4 - z^3 + 1\}$ is the reduced Gröbner basis for the grevlex monomial ordering.

29. Solve the system of equations $x^2 - yz = 3$, $y^2 - xz = 4$, $z^2 - xy = 5$ over $\mathbb{C}$.

30. Find a Gröbner basis for the ideal $I = (x^2 + xy + y^2 - 1, x^2 + 4y^2 - 4)$ for the lexicographic ordering $x > y$ and use it to find the four points of intersection of the ellipse $x^2 + xy + y^2 = 1$ with the ellipse $x^2 + 4y^2 = 4$ in $\mathbb{R}^2$.

31. Use Gröbner bases to find all six solutions to the system of equations $2x^3 + 2x^2y^2 + 3y^3 = 0$ and $3x^5 + 2x^3y^3 + 2y^5 = 0$ over $\mathbb{C}$.

32. Use Gröbner bases to show that $(x, z) \cap (y^2, x - yz) = (xy, x - yz)$ in $F[x, y, z]$.

33. Use Gröbner bases to compute the intersection of the ideals $(x^3y - xy^2 + 1, x^2y^2 - y^3 - 1)$ and $(x^2 - y^2, x^3 + y^3)$ in $F[x, y]$.

The following four exercises deal with the *ideal quotient* of two ideals I and J in a ring R.

Definition. The *ideal quotient* $(I : J)$ of two ideals I, J in a ring R is the ideal

$$(I : J) = \{r \in R \mid rJ \in I\}.$$

34. (a) Suppose R is an integral domain, $0 \neq f \in R$ and I is an ideal in R. Show that if $\{g_1, \ldots, g_s\}$ are generators for the ideal $I \cap (f)$, then $\{g_1/f, \ldots, g_s/f\}$ are generators for the ideal quotient $(I : (f))$.

(b) If I is an ideal in the commutative ring R and $f_1, \ldots, f_s \in R$, show that the ideal quotient $(I : (f_1, \ldots f_s))$ is the ideal $\cap_{i=1}^{s} (I : (f_i))$.

35. If $I = (x^2y + z^3, x + y^3 - z, 2y^4z - yz^2 - z^3)$ and $J = (x^2y^5, x^3z^4, y^3z^7)$ in $\mathbb{Q}[x, y, z]$ show $(I : J)$ is the ideal $(z^2, y + z, x - z)$. [Use the previous exercise and Proposition 30.]

36. Suppose that K is an ideal in R, that I is an ideal containing K, and J is any ideal. If $\overline{I}$ and $\overline{J}$ denote the images of I and J in the quotient ring R/K, show that $\overline{(I : J)} = (\overline{I} : \overline{J})$ where $\overline{(I : J)}$ is the image in R/K of the ideal quotient $(I : J)$.

37. Let K be the ideal $(y^5 - z^4)$ in $R = \mathbb{Q}[y, z]$. For each of the following pairs of ideals I and J, use the previous two exercises together with Proposition 30 to verify the ideal quotients $(\overline{I} : \overline{J})$ in the ring R/K:

i. $I = (y^3, y^5 - z^4)$, $J = (z)$, $(\overline{I} : \overline{J}) = (\bar{y}^3, \bar{z}^3)$.

ii. $I = (y^3, z, y^5 - z^4)$, $J = (y)$, $(\overline{I} : \overline{J}) = (\bar{y}^2, \bar{z})$.

iii. $I = (y, y^3, z, y^5 - z^4)$, $J = (1)$, $(\overline{I} : \overline{J}) = (\bar{y}, \bar{z})$.

Exercises 38 to 44 develop some additional elementary properties of monomial ideals in $F[x_1, \ldots, x_n]$. It follows from Hilbert's Basis Theorem that ideals are finitely generated, however one need not assume this in these exercises—the arguments are the same for finitely or infinitely generated ideals. These exercises may be used to give an independent proof of Hilbert's Basis Theorem (Exercise 44). In these exercises, M and N are monomial ideals with monomial generators $\{m_i \mid i \in I\}$ and $\{n_j \mid j \in J\}$ for some index sets I and J respectively.

38. Prove that the sum and product of two monomial ideals is a monomial ideal by showing that $M + N = (m_i,\ n_j \mid i \in I,\ j \in J)$, and $MN = (m_i n_j \mid i \in I,\ j \in J)$.

39. Show that if $\{M_s \mid s \in S\}$ is any nonempty collection of monomial ideals that is totally ordered under inclusion then $\cup_{s \in S} M_s$ is a monomial ideal. (In particular, the union of any increasing sequence of monomial ideals is a monomial ideal, cf. Exercise 19, Section 7.3.)

40. Prove that the intersection of two monomial ideals is a monomial ideal by showing that $M \cap N = (e_{i,j} \mid i \in I,\ j \in J)$, where $e_{i,j}$ is the least common multiple of m_i and n_j. [Use Exercise 10.]

41. Prove that for any monomial n, the ideal quotient $(M : (n))$ is $(m_i/d_i \mid i \in I)$, where d_i is the greatest common divisor of m_i and n (cf. Exercise 34). Show that if N is finitely generated, then the ideal quotient $(M : N)$ of two monomial ideals is a monomial ideal.

42. **(a)** Show that M is a monomial prime ideal if and only if $M = (S)$ for some subset of S of $\{x_1, x_2, \ldots, x_n\}$. (In particular, there are only finitely many monomial prime ideals, and each is finitely generated.)

(b) Show that $(x_1, \ldots, x_n)$ is the only monomial maximal ideal.

43. *(Dickson's Lemma*—a special case of Hilbert's Basis Theorem) Prove that every monomial ideal in $F[x_1, \ldots, x_n]$ is finitely generated as follows.
Let $\mathcal{S} = \{N \mid N$ is a monomial ideal that is not finitely generated$\}$, and assume by way of contradiction $\mathcal{S} \neq \emptyset$.

(a) Show that $\mathcal{S}$ contains a maximal element M. [Use Exercise 30 and Zorn's Lemma.]

(b) Show that there are monomials x, y not in M with $xy \in M$. [Use Exercise 33(a).]

(c) For x as in (b), show that M contains a finitely generated monomial ideal M_0 such that $M_0 + (x) = M + (x)$ and $M = M_0 + (x)(M : (x))$, where $(M : (x))$ is the (monomial) ideal defined in Exercise 32, and $(x)(M : (x))$ is the product of these two ideals. Deduce that M is finitely generated, a contradiction which proves $\mathcal{S} = \emptyset$. [Use the maximality of M and previous exercises.]

44. If I is a nonzero ideal in $F[x_1, \ldots, x_n]$, use Dickson's Lemma to prove that $LT(I)$ is finitely generated. Conclude that I has a Gröbner basis and deduce Hilbert's Basis Theorem. [cf. Proposition 24.]

45. *(n-colorings of graphs)* A finite *graph* $\mathcal{G}$ of size N is a set of vertices $i \in \{1, 2, \ldots, N\}$ and a collection of edges (i, j) connecting vertex i with vertex j. An *n-coloring* of $\mathcal{G}$ is an assignment of one of n colors to each vertex in such a way that vertices connected by an edge have distinct colors. Let F be any field containing at least n elements. If we introduce a variable x_i for each vertex i and represent the n colors by choosing a set S of n distinct elements from F, then an n-coloring of $\mathcal{G}$ is equivalent to assigning a value $x_i = \alpha_i$ for each $i = 1, 2, \ldots, N$ where $\alpha_i \in S$ and $\alpha_i \neq \alpha_j$ if (i, j) is an edge in $\mathcal{G}$. If $f(x) = \prod_{\alpha \in S}(x - \alpha)$ is the polynomial in $F[x]$ of degree n whose roots are the elements in S, then $x_i = \alpha_i$ for some $\alpha_i \in S$ is equivalent to the statement that x_i is a solution to the equation $f(x_i) = 0$. The statement $\alpha_i \neq \alpha_j$ is then the statement that $f(x_i) = f(x_j)$ but $x_i \neq x_j$, so x_i and x_j satisfy the equation $g(x_i, x_j) = 0$, where $g(x_i, x_j)$ is the polynomial $(f(x_i) - f(x_j))/(x_i - x_j)$ in $F[x_i, x_j]$. It follows that finding an *n-coloring* of $\mathcal{G}$ is equivalent to solving the system of equations

$$\begin{cases} f(x_i) = 0, & \text{for } i = 1, 2, \ldots, N, \\ g(x_i, x_j) = 0, & \text{for all edges } (i, j) \text{ in } \mathcal{G} \end{cases}$$

(note also we may use any polynomial g satisfying $\alpha_i \neq \alpha_j$ if $g(\alpha_i, \alpha_j) = 0$). It follows by "Hilbert's Nullstellensatz" (cf. Corollary 33 in Section 15.3) that this system of equations has a solution, hence $\mathcal{G}$ has an n-coloring, unless the ideal I in $F[x_1, x_2, \ldots, x_N]$ generated by the polynomials $f(x_i)$ for $i = 1, 2, \ldots, N$, together with the polynomials $g(x_i, x_j)$ for all the edges (i, j) in the graph $\mathcal{G}$, is not a proper ideal. This in turn is equivalent to the statement that the reduced Gröbner basis for I (with respect to any monomial ordering) is simply $\{1\}$. Further, when an n-coloring does exist, solving this system of equations as in the examples following Proposition 29 provides an explicit coloring for $\mathcal{G}$.

There are many possible choices of field F and set S. For example, use any field F containing a set S of distinct n^{th} roots of unity, in which case $f(x) = x^n - 1$ and we may take $g(x_i, x_j) = (x_i^n - x_j^n)/(x_i - x_j) = x_i^{n-1} + x_i^{n-2}x_j + \cdots + x_i x_j^{n-2} + x_j^{n-1}$, or use any subset S of $F = \mathbb{F}_p$ with a prime $p \geq n$ (in the special case $n = p$, then, by Fermat's Little Theorem, we have $f(x) = x^p - x$ and $g(x_i, x_j) = (x_i - x_j)^{p-1} - 1$).

(a) Consider a possible 3-coloring of the graph $\mathcal{G}$ with eight vertices and 14 edges $(1, 3)$, $(1, 4)$, $(1, 5)$, $(2, 4)$, $(2, 7)$, $(2, 8)$, $(3, 4)$, $(3, 6)$, $(3, 8)$, $(4, 5)$, $(5, 6)$, $(6, 7)$, $(6, 8)$, $(7, 8)$. Take $F = \mathbb{F}_3$ with 'colors' $0, 1, 2 \in \mathbb{F}_3$ and suppose vertex 1 is colored by 0. In this case $f(x) = x(x - 1)(x - 2) = x^3 - x \in \mathbb{F}_3[x]$ and $g(x_i, x_j) = x_i^2 + x_i x_j + x_j^2 - 1$. If I is the ideal generated by x_1, $x_i^3 - x_i$, $2 \leq i \leq 8$ and $g(x_i, x_j)$ for the edges (i, j) in $\mathcal{G}$, show that the reduced Gröbner basis for I with respect to the lexicographic monomial ordering $x_1 > x_2 > \cdots > x_8$ is $\{x_1, x_2, x_3 + x_8, x_4 + 2x_8, x_5 + x_8, x_6, x_7 + x_8, x_8^2 + 2\}$. Deduce that $\mathcal{G}$ has two distinct 3-colorings, determined by the coloring of vertex 8 (which must be colored by a nonzero element in $\mathbb{F}_3$), and exhibit the colorings of $\mathcal{G}$.

Show that if the edge $(3, 7)$ is added to $\mathcal{G}$ then the graph cannot be 3-colored.

(b) Take $F = \mathbb{F}_5$ with four 'colors' $1, 2, 3, 4 \in \mathbb{F}_5$, so $f(x) = x^4 - 1$ and we may use $g(x_i, x_j) = x_i^3 + x_i^2 x_j + x_i x_j^2 + x_j^3$. Show that the graph $\mathcal{G}$ with five vertices having 9 edges $(1, 3)$, $(1, 4)$, $(1, 5)$, $(2, 3)$, $(2, 4)$, $(2, 5)$, $(3, 4)$, $(3, 5)$, $(4, 5)$ (the "complete graph on five vertices" with one edge removed) can be 4-colored but cannot be 3-colored.

(c) Use Gröbner bases to show that the graph $\mathcal{G}$ with nine vertices and 22 edges $(1, 4)$, $(1, 6)$, $(1, 7)$, $(1, 8)$, $(2, 3)$, $(2, 4)$, $(2, 6)$, $(2, 7)$, $(3, 5)$, $(3, 7)$, $(3, 9)$, $(4, 5)$, $(4, 6)$, $(4, 7)$, $(4, 9)$, $(5, 6)$, $(5, 7)$, $(5, 8)$, $(5, 9)$, $(6, 7)$, $(6, 9)$, $(7, 8)$ has precisely four 4-colorings up to a permutation of the colors (so a total of 96 total 4-colorings). Show that if the edge $(1, 5)$ is added then $\mathcal{G}$ cannot be 4-colored.

Part III

MODULES AND VECTOR SPACES

In Part III we study the mathematical objects called modules. The use of modules was pioneered by one of the most prominent mathematicians of the first part of this century, Emmy Noether, who led the way in demonstrating the power and elegance of this structure. We shall see that vector spaces are just special types of modules which arise when the underlying ring is a field. If R is a ring, the definition of an R-module M is closely analogous to the definition of a group action where R plays the role of the group and M the role of the set. The additional axioms for a module require that M itself have more structure (namely that M be an abelian group). Modules are the "representation objects" for rings, i.e., they are, by definition, algebraic objects on which rings act. As the theory develops it will become apparent how the structure of the ring R (in particular, the structure and wealth of its ideals) is reflected by the structure of its modules and vice versa in the same way that the structure of the collection of normal subgroups of a group was reflected by its permutation representations.

CHAPTER 10

Introduction to Module Theory

10.1 BASIC DEFINITIONS AND EXAMPLES

We start with the definition of a module.

Definition. Let R be a ring (not necessarily commutative nor with 1). A *left R-module* or a *left module over R* is a set M together with

(1) a binary operation $+$ on M under which M is an abelian group, and

(2) an action of R on M (that is, a map $R \times M \to M$) denoted by rm, for all $r \in R$ and for all $m \in M$ which satisfies

(a) $(r+s)m = rm + sm$, for all $r, s \in R, m \in M$,

(b) $(rs)m = r(sm)$, for all $r, s \in R, m \in M$, and

(c) $r(m+n) = rm + rn$, for all $r \in R, m, n \in M$.

If the ring R has a 1 we impose the additional axiom:

(d) $1m = m$, for all $m \in M$.

The descriptor "left" in the above definition indicates that the ring elements appear on the left; "right" R-modules can be defined analogously. If the ring R is *commutative* and M is a left R-module we can make M into a right R-module by defining $mr = rm$ for $m \in M$ and $r \in R$. If R is not commutative, axiom 2(b) in general will not hold with this definition (so not every left R-module is also a right R-module). Unless explicitly mentioned otherwise the term "module" will always mean "left module." Modules satisfying axiom 2(d) are called *unital* modules and in this book all our modules will be unital (this is to avoid "pathologies" such as having $rm = 0$ for all $r \in R$ and $m \in M$).

When R is a field F the axioms for an R-module are precisely the same as those for a vector space over F, so that

modules over a field F and vector spaces over F are the same.

Before giving other examples of R-modules we record the obvious definition of submodules.

Definition. Let R be a ring and let M be an R-module. An *R-submodule* of M is a subgroup N of M which is closed under the action of ring elements, i.e., $rn \in N$, for all $r \in R, n \in N$.

Submodules of M are therefore just subsets of M which are themselves modules under the restricted operations. In particular, if $R = F$ is a field, submodules are the same as subspaces. Every R-module M has the two submodules M and 0 (the latter is called the *trivial submodule*).

Examples

(1) Let R be any ring. Then $M = R$ is a left R-module, where the action of a ring element on a module element is just the usual multiplication in the ring R (similarly, R is a right module over itself). In particular, every field can be considered as a (1-dimensional) vector space over itself. When R is considered as a left module over itself in this fashion, the submodules of R are precisely the left ideals of R (and if R is considered as a right R-module over itself, its submodules are the right ideals). Thus if R is not commutative it has a left and right module structure over itself and these structures may be different (e.g., the submodules may be different) — Exercise 21 at the end of this section gives a specific example of this.

(2) Let $R = F$ be a field. As noted above, every vector space over F is an F-module and vice versa. Let $n \in \mathbb{Z}^+$ and let

$$F^n = \{(a_1, a_2, \ldots, a_n) \mid a_i \in F, \text{ for all } i\}$$

(called *affine n-space over F*). Make F^n into a vector space by defining addition and scalar multiplication componentwise:

$$(a_1, a_2, \ldots, a_n) + (b_1, b_2, \ldots, b_n) = (a_1 + b_1, a_2 + b_2, \ldots, a_n + b_n)$$
$$\alpha(a_1, \ldots, a_n) = (\alpha a_1, \ldots, \alpha a_n), \qquad \alpha \in F.$$

As in the case of Euclidean n-space (i.e., when $F = \mathbb{R}$), affine n-space is a vector space of dimension n over F (we shall discuss the notion of dimension more thoroughly in the next chapter).

(3) Let R be a ring with 1 and let $n \in \mathbb{Z}^+$. Following Example 2 define

$$R^n = \{(a_1, a_2, \ldots, a_n) \mid a_i \in R, \text{ for all } i\}.$$

Make R^n into an R-module by componentwise addition and multiplication by elements of R in the same manner as when R was a field. The module R^n is called *the free module of rank n over R*. (We shall see shortly that free modules have the same "universal property" in the context of R-modules that free groups were seen to have in Section 6.3. We shall also soon discuss direct products of R-modules.) An obvious submodule of R^n is given by the i^{th} component, namely the set of n-tuples with arbitrary ring elements in the i^{th} component and zeros in the j^{th} component for all $j \neq i$.

(4) The same abelian group may have the structure of an R-module for a number of different rings R and each of these module structures may carry useful information. Specifically, if M is an R-module and S is a subring of R with $1_S = 1_R$, then M is automatically an S-module as well. For instance the field $\mathbb{R}$ is an $\mathbb{R}$-module, a $\mathbb{Q}$-module and a $\mathbb{Z}$-module.

(5) If M is an R-module and for some (2-sided) ideal I of R, $am = 0$, for all $a \in I$ and all $m \in M$, we say M is *annihilated by I*. In this situation we can make M into an (R/I)-module by defining an action of the quotient ring R/I on M as follows: for each $m \in M$ and coset $r + I$ in R/I let

$$(r + I)m = rm.$$

Since $am = 0$ for all $a \in I$ and all $m \in M$ this is well defined and one easily checks that it makes M into an (R/I)-module. In particular, when I is a maximal ideal in the commutative ring R and $IM = 0$, then M is a vector space over the field R/I (cf. the following example).

The next example is of sufficient importance as to be singled out. It will form the basis for our proof of the Fundamental Theorem of Finitely Generated Abelian Groups in Chapter 12.

Example: ($\mathbb{Z}$-modules)

Let R = $\mathbb{Z}$, let A be *any* abelian group (finite or infinite) and write the operation of A as $+$. Make A into a $\mathbb{Z}$-module as follows: for any $n \in \mathbb{Z}$ and $a \in A$ define

$$na = \begin{cases} a + a + \cdots + a \quad (n \text{ times}) & \text{if } n > 0 \\ 0 & \text{if } n = 0 \\ -a - a - \cdots - a \quad (-n \text{ times}) & \text{if } n < 0 \end{cases}$$

(here 0 is the identity of the additive group A). This definition of an action of the integers on A makes A into a $\mathbb{Z}$-module, and the module axioms show that this is the only possible action of $\mathbb{Z}$ on A making it a (unital) $\mathbb{Z}$-module. Thus every abelian group is a $\mathbb{Z}$-module. Conversely, if M is any $\mathbb{Z}$-module, a fortiori M is an abelian group, so

$\mathbb{Z}$-modules are the same as abelian groups.

Furthermore, it is immediate from the definition that

$\mathbb{Z}$-submodules are the same as subgroups.

Note that for the cyclic group $\langle a \rangle$ written multiplicatively the additive notation na becomes a^n, that is, we have all along been using the fact that $\langle a \rangle$ is a right $\mathbb{Z}$-module (checking that this "exponential" notation satisfies the usual laws of exponents is equivalent to checking the $\mathbb{Z}$-module axioms — this was given as an exercise at the end of Section 1.1). Note that since $\mathbb{Z}$ is commutative these definitions of left and right actions by ring elements give the same module structure.

If A is an abelian group containing an element x of finite order n then $nx = 0$. Thus, in contrast to vector spaces, a $\mathbb{Z}$-module may have nonzero elements x such that $nx = 0$ for some nonzero ring element n. In particular, if A has order m, then by Lagrange's Theorem (Corollary 9, Section 3.2) $mx = 0$, for *all* $x \in A$. Note that then A is a module over $\mathbb{Z}/m\mathbb{Z}$.

In particular, if p is a prime and A is an abelian group (written additively) such that $px = 0$, for all $x \in A$, then (as noted in Example 5) A is a $\mathbb{Z}/p\mathbb{Z}$-module, i.e., can be considered as a vector space over the field $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$. For instance, the Klein 4-group is a (2-dimensional) vector space over $\mathbb{F}_2$. These groups are the *elementary abelian p-groups* discussed in Section 4.4 (see, in particular, Proposition 17(3)).

The next example is also of fundamental importance and will form the basis for our study of canonical forms of matrices in Sections 12.2 and 12.3.

Example: ($F[x]$-modules)

Let F be a field, let x be an indeterminate and let R be the polynomial ring $F[x]$. Let V be a vector space over F and let T be a linear transformation from V to V (we shall review the theory of linear transformations in the next chapter — for the purposes of this example one only needs to know the definition of a linear transformation). We have already seen that V is an F-module; the linear map T will enable us to make V into an $F[x]$-module.

First, for the nonnegative integer n, define

$$T^0 = I,$$

$$\vdots$$

$$T^n = T \circ T \circ \cdots \circ T \qquad (n \text{ times})$$

where I is the identity map from V to V and $\circ$ denotes function composition (which makes sense because the domain and codomain of T are the same). Also, for any two linear transformations A, B from V to V and elements $\alpha, \beta \in F$, let $\alpha A + \beta B$ be defined by

$$(\alpha A + \beta B)(v) = \alpha(A(v)) + \beta(B(v))$$

(i.e., addition and scalar multiplication of linear transformations are defined pointwise). Then $\alpha A + \beta B$ is easily seen to be a linear transformation from V to V, so that linear combinations of linear transformations are again linear transformations.

We now define the action of any polynomial in x on V. Let $p(x)$ be the polynomial

$$p(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0,$$

where $a_0, \ldots, a_n \in F$. For each $v \in V$ define an action of the ring element $p(x)$ on the module element v by

$$\begin{aligned} p(x)v &= (a_n T^n + a_{n-1}T^{n-1} + \cdots + a_1 T + a_0)(v) \\ &= a_n T^n(v) + a_{n-1}T^{n-1}(v) + \cdots + a_1 T(v) + a_0 v \end{aligned}$$

(i.e., $p(x)$ acts by substituting the linear transformation T for x in $p(x)$ and applying the resulting linear transformation to v). Put another way, x acts on V as the linear transformation T and we extend this to an action of all of $F[x]$ on V in a natural way. It is easy to check that this definition of an action of $F[x]$ on V satisfies all the module axioms and makes V into an $F[x]$-module.

The field F is naturally a subring of $F[x]$ (the constant polynomials) and the action of these field elements is by definition the same as their action when viewed as constant polynomials. In other words, the definition of the $F[x]$ action on V is consistent with the given action of the field F on the vector space V, i.e., the definition *extends* the action of F to an action of the larger ring $F[x]$.

The way $F[x]$ acts on V depends on the choice of T so that there are in general many different $F[x]$-module structures on the same vector space V. For instance, if $T = 0$, and $p(x)$, v are as above, then $p(x)v = a_0 v$, that is, the polynomial $p(x)$ acts on v simply by multiplying by the constant term of $p(x)$, so that the $F[x]$-module structure is just the F-module structure. If, on the other hand, T is the identity transformation (so $T^n(v) = v$, for all n and v), then $p(x)v = a_n v + a_{n-1}v + \cdots + a_0 v = (a_n + \cdots + a_0)v$, so that now $p(x)$ multiplies v by the sum of the coefficients of $p(x)$.

To give another specific example, let V be affine n-space F^n and let T be the "shift operator"

$$T(x_1, x_2, \ldots, x_n) = (x_2, x_3, \ldots, x_n, 0).$$

Let e_i be the usual i^{th} basis vector $(0, 0, \ldots, 0, 1, 0, \ldots, 0)$ where the 1 is in position i. Then

$$T^k(e_i) = \begin{cases} e_{i-k} & \text{if } i > k \\ 0 & \text{if } i \leq k \end{cases}$$

so for example, if $m < n$,

$$(a_m x^m + a_{m-1}x^{m-1} + \cdots + a_0)e_n = (0, \ldots, 0, a_m, a_{m-1}, \ldots, a_0).$$

From this we can determine the action of any polynomial on any vector.

The construction of an $F[x]$-module from a vector space V over F and a linear transformation T from V to V in fact describes *all* $F[x]$-modules; namely, an $F[x]$-module is a vector space together with a linear transformation which specifies the action of x. This is because if V is any $F[x]$-module, then V is an F-module and the action of the ring element x on V is a linear transformation from V to V. The axioms for a module ensure that the actions of F and x on V uniquely determine the action of any element of $F[x]$ on V. Thus there is a bijection between the collection of $F[x]$-modules and the collection of pairs V, T

$$\left\{ V \text{ an } F[x]\text{-module} \right\} \quad \longleftrightarrow \quad \left\{ \begin{array}{c} V \text{ a vector space over } F \\ \text{and} \\ T : V \to V \text{ a linear transformation} \end{array} \right\}$$

given by

the element x acts on V as the linear transformation T.

Now we consider $F[x]$-submodules of V where, as above, V is any $F[x]$-module and T is the linear transformation from V to V given by the action of x. An $F[x]$-submodule W of V must first be an F-submodule, i.e., W must be a vector subspace of V. Secondly, W must be sent to itself under the action of the ring element x, i.e., we must have $T(w) \in W$, for all $w \in W$. Any vector subspace U of V such that $T(U) \subseteq U$ is called *T-stable* or *T-invariant*. If U is any T-stable subspace of V it follows that $T^n(U) \subseteq U$, for all $n \in \mathbb{Z}^+$ (for example, $T(U) \subseteq U$ implies $T^2(U) = T(T(U)) \subseteq T(U) \subseteq U$). Moreover any linear combination of powers of T then sends U into U so that U is also stable by the action of *any* polynomial in T. Thus U is an $F[x]$-submodule of V. This shows that

the $F[x]$-submodules of V are precisely the T-stable subspaces of V.

In terms of the bijection above,

$$\left\{ W \text{ an } F[x]\text{-submodule} \right\} \quad \longleftrightarrow \quad \left\{ \begin{array}{c} W \text{ a subspace of } V \\ \text{and} \\ W \text{ is } T\text{-stable} \end{array} \right\}$$

which gives a complete dictionary between $F[x]$-modules V and vector spaces V together with a given linear transformation T from V to V.

For instance, if T is the shift operator defined on affine n-space above and k is any integer in the range $0 \leq k \leq n$, then the subspace

$$U_k = \{ (x_1, x_2, \ldots, x_k, 0, \ldots, 0) \mid x_i \in F \}$$

is clearly T-stable so is an $F[x]$-submodule of V.

We emphasize that an abelian group M may have many different R-module structures, even if the ring R does not vary (in the same way that a given group G may act in many ways as a permutation group on some fixed set Ω). We shall see that the structure of an R-module is reflected by the ideal structure of R. When R is a field (the subject of the next chapter) all R-modules will be seen to be products of copies of R (as in Example 3 above).

We shall see in Chapter 12 that the relatively simple ideal structure of the ring $F[x]$ (recall that $F[x]$ is a Principal Ideal Domain) forces the $F[x]$-module structure of V to be correspondingly uncomplicated, and this in turn provides a great deal of information about the linear transformation T (in particular, gives some nice matrix representations for T: its *rational canonical form* and its *Jordan canonical form*). Moreover, the same arguments which *classify* finitely generated $F[x]$-modules apply to any Principal Ideal Domain R, and when these are invoked for $R = \mathbb{Z}$, we obtain the Fundamental Theorem of Finitely Generated Abelian Groups. These results generalize the theorem that every finite dimensional vector space has a basis.

In Part VI of the book we shall study modules over certain noncommutative rings (group rings) and see that this theory in some sense generalizes both the study of $F[x]$-modules in Chapter 12 and the notion of a permutation representation of a finite group.

We establish a submodule criterion analogous to that for subgroups of a group in Section 2.1.

Proposition 1. *(The Submodule Criterion)* Let R be a ring and let M be an R-module. A subset N of M is a submodule of M if and only if

(1) $N \neq \emptyset$, and

(2) $x + ry \in N$ for all $r \in R$ and for all $x, y \in N$.

Proof: If N is a submodule, then $0 \in N$ so $N \neq \emptyset$. Also N is closed under addition and is sent to itself under the action of elements of R. Conversely, suppose (1) and (2) hold. Let $r = -1$ and apply the subgroup criterion (in additive form) to see that N is a subgroup of M. In particular, $0 \in N$. Now let $x = 0$ and apply hypothesis (2) to see that N is sent to itself under the action of R. This establishes the proposition.

We end this section with an important definition and some examples.

Definition. Let R be a commutative ring with identity. An *R-algebra* is a ring A with identity together with a ring homomorphism $f : R \to A$ mapping 1_R to 1_A such that the subring $f(R)$ of A is contained in the center of A.

If A is an R-algebra then it is easy to check that A has a natural left and right (unital) R-module structure defined by $r \cdot a = a \cdot r = f(r)a$ where $f(r)a$ is just the multiplication in the ring A (and this is the same as $af(r)$ since by assumption $f(r)$ lies in the center of A). In general it is possible for an R-algebra A to have other left (or right) R-module structures, but unless otherwise stated, this natural module structure on an algebra will be assumed.

Definition. If A and B are two R-algebras, an *R-algebra homomorphism* (or isomorphism) is a ring homomorphism (isomorphism, respectively) $\varphi : A \to B$ mapping 1_A to 1_B such that $\varphi(r \cdot a) = r \cdot \varphi(a)$ for all $r \in R$ and $a \in A$.

Examples

Let R be a commutative ring with 1.

(1) Any ring with identity is a $\mathbb{Z}$-algebra.

(2) For any ring A with identity, if R is a subring of the center of A containing the identity of A then A is an R-algebra. In particular, a commutative ring A containing 1 is an R-algebra for any subring R of A containing 1. For example, the polynomial ring $R[x]$ is an R-algebra, the polynomial ring over R in any number of variables is an R-algebra, and the group ring RG for a finite group G is an R-algebra (cf. Section 7.2).

(3) If A is an R-algebra then the R-module structure of A depends only on the subring $f(R)$ contained in the center of A as in the previous example. If we replace R by its image $f(R)$ we see that "up to a ring homomorphism" every algebra A arises from a subring of the center of A that contains 1_A.

(4) A special case of the previous example occurs when $R = F$ is a *field*. In this case F is *isomorphic* to its image under f, so we can identify F itself as a subring of A. Hence, saying that A is an algebra over a field F is the same as saying that the ring A contains the field F in its center and the identity of A and of F are the same (this last condition is necessary, cf. Exercise 23).

Suppose that A is an R-algebra. Then A is a ring with identity that is a (unital) left R-module satisfying $r \cdot (ab) = (r \cdot a)b = a(r \cdot b)$ for all $r \in R$ and $a, b \in A$ (these are all equal to the product $f(r)ab$ in the ring A—recall that $f(R)$ is contained in the center of A). Conversely, these conditions on a ring A define an R-algebra, and are sometimes used as the definition of an R-algebra (cf. Exercise 22).

EXERCISES

In these exercises R is a ring with 1 and M is a left R-module.

1. Prove that $0m = 0$ and $(-1)m = -m$ for all $m \in M$.

2. Prove that $R^\times$ and M satisfy the two axioms in Section 1.7 for a *group action* of the multiplicative group $R^\times$ on the set M.

3. Assume that $rm = 0$ for some $r \in R$ and some $m \in M$ with $m \neq 0$. Prove that r does not have a left inverse (i.e., there is no $s \in R$ such that $sr = 1$).

4. Let M be the module R^n described in Example 3 and let $I_1, I_2, \ldots, I_n$ be left ideals of R. Prove that the following are submodules of M:

(a) $\{(x_1, x_2, \ldots, x_n) \mid x_i \in I_i\}$

(b) $\{(x_1, x_2, \ldots, x_n) \mid x_i \in R \text{ and } x_1 + x_2 + \cdots + x_n = 0\}$.

5. For any left ideal I of R define

$$IM = \{\sum_{\text{finite}} a_i m_i \mid a_i \in I,\ m_i \in M\}$$

to be the collection of all finite sums of elements of the form am where $a \in I$ and $m \in M$. Prove that IM is a submodule of M.

6. Show that the intersection of any nonempty collection of submodules of an R-module is a submodule.

7. Let $N_1 \subseteq N_2 \subseteq \cdots$ be an ascending chain of submodules of M. Prove that $\cup_{i=1}^{\infty} N_i$ is a submodule of M.

8. An element m of the R-module M is called a *torsion element* if $rm = 0$ for some nonzero element $r \in R$. The set of torsion elements is denoted

$$\text{Tor}(M) = \{m \in M \mid rm = 0 \text{ for some nonzero } r \in R\}.$$

   (a) Prove that if R is an integral domain then $\text{Tor}(M)$ is a submodule of M (called the *torsion* submodule of M).
   (b) Give an example of a ring R and an R-module M such that $\text{Tor}(M)$ is not a submodule. [Consider the torsion elements in the R-module R.]
   (c) If R has zero divisors show that every nonzero R-module has nonzero torsion elements.

9. If N is a submodule of M, the *annihilator of N in R* is defined to be $\{r \in R \mid rn = 0 \text{ for all } n \in N\}$. Prove that the annihilator of N in R is a 2-sided ideal of R.

10. If I is a right ideal of R, the *annihilator of I in M* is defined to be $\{m \in M \mid am = 0 \text{ for all } a \in I\}$. Prove that the annihilator of I in M is a submodule of M.

11. Let M be the abelian group (i.e., $\mathbb{Z}$-module) $\mathbb{Z}/24\mathbb{Z} \times \mathbb{Z}/15\mathbb{Z} \times \mathbb{Z}/50\mathbb{Z}$.
   (a) Find the annihilator of M in $\mathbb{Z}$ (i.e., a generator for this principal ideal).
   (b) Let $I = 2\mathbb{Z}$. Describe the annihilator of I in M as a direct product of cyclic groups.

12. In the notation of the preceding exercises prove the following facts about annihilators.
   (a) Let N be a submodule of M and let I be its annihilator in R. Prove that the annihilator of I in M contains N. Give an example where the annihilator of I in M does not equal N.
   (b) Let I be a right ideal of R and let N be its annihilator in M. Prove that the annihilator of N in R contains I. Give an example where the annihilator of N in R does not equal I.

13. Let I be an ideal of R. Let M' be the subset of elements a of M that are annihilated by some power, I^k, of the ideal I, where the power may depend on a. Prove that M' is a submodule of M. [Use Exercise 7.]

14. Let z be an element of the center of R, i.e., $zr = rz$ for all $r \in R$. Prove that zM is a submodule of M, where $zM = \{zm \mid m \in M\}$. Show that if R is the ring of 2×2 matrices over a field and e is the matrix with a 1 in position 1,1 and zeros elsewhere then eR is *not* a left R-submodule (where $M = R$ is considered as a left R-module as in Example 1) — in this case the matrix e is not in the center of R.

15. If M is a finite abelian group then M is naturally a $\mathbb{Z}$-module. Can this action be extended to make M into a $\mathbb{Q}$-module?

16. Prove that the submodules U_k described in the example of $F[x]$-modules are all of the $F[x]$-submodules for the shift operator.

17. Let T be the shift operator on the vector space V and let $e_1, \ldots, e_n$ be the usual basis vectors described in the example of $F[x]$-modules. If $m \geq n$ find $(a_m x^m + a_{m-1}x^{m-1} + \cdots + a_0)e_n$.

18. Let $F = \mathbb{R}$, let $V = \mathbb{R}^2$ and let T be the linear transformation from V to V which is rotation clockwise about the origin by $\pi/2$ radians. Show that V and 0 are the only $F[x]$-submodules for this T.

19. Let $F = \mathbb{R}$, let $V = \mathbb{R}^2$ and let T be the linear transformation from V to V which is projection onto the y-axis. Show that V, 0, the x-axis and the y-axis are the only $F[x]$-submodules for this T.

20. Let $F = \mathbb{R}$, let $V = \mathbb{R}^2$ and let T be the linear transformation from V to V which is rotation clockwise about the origin by π radians. Show that *every* subspace of V is an

$F[x]$-submodule for this T.

21. Let $n \in \mathbb{Z}^+$, $n > 1$ and let R be the ring of $n \times n$ matrices with entries from a field F. Let M be the set of $n \times n$ matrices with arbitrary elements of F in the first column and zeros elsewhere. Show that M is a submodule of R when R is considered as a left module over itself, but M is not a submodule of R when R is considered as a right R-module.

22. Suppose that A is a ring with identity 1_A that is a (unital) left R-module satisfying $r \cdot (ab) = (r \cdot a)b = a(r \cdot b)$ for all $r \in R$ and $a, b \in A$. Prove that the map $f : R \to A$ defined by $f(r) = r \cdot 1_A$ is a ring homomorphism mapping 1_R to 1_A and that $f(R)$ is contained in the center of A. Conclude that A is an R-algebra and that the R-module structure on A induced by its algebra structure is precisely the original R-module structure.

23. Let A be the direct product ring $\mathbb{C} \times \mathbb{C}$ (cf. Section 7.6). Let τ_1 denote the identity map on $\mathbb{C}$ and let τ_2 denote complex conjugation. For any pair $p, q \in \{1, 2\}$ (not necessarily distinct) define

$$f_{p,q} : \mathbb{C} \to \mathbb{C} \times \mathbb{C} \qquad \text{by} \qquad f_{p,q}(z) = (\tau_p(z), \tau_q(z)).$$

So, for example, $f_{2,1} : z \mapsto (\bar{z}, z)$, where $\bar{z}$ is the complex conjugate of z, i.e., $\tau_2(z)$.

(a) Prove that each $f_{p,q}$ is an injective ring homomorphism, and that they all agree on the subfield $\mathbb{R}$ of $\mathbb{C}$. Deduce that A has four distinct $\mathbb{C}$-algebra structures. Explicitly give the action $z \cdot (u, v)$ of a complex number z on an ordered pair in A in each case.

(b) Prove that if $f_{p,q} \neq f_{p',q'}$ then the identity map on A is *not* a $\mathbb{C}$-algebra homomorphism from A considered as a $\mathbb{C}$-algebra via $f_{p,q}$ to A considered a $\mathbb{C}$-algebra via $f_{p',q'}$ (although the identity is an $\mathbb{R}$-algebra isomorphism).

(c) Prove that for any pair p, q there is some ring isomorphism from A to itself such that A is isomorphic as a $\mathbb{C}$-algebra via $f_{p,q}$ to A considered as $\mathbb{C}$-algebra via $f_{1,1}$ (the "natural" $\mathbb{C}$-algebra structure on A).

Remark: In the preceding exercise $A = \mathbb{C} \times \mathbb{C}$ is not a $\mathbb{C}$-algebra over either of the direct factor component copies of $\mathbb{C}$ (for example the subring $\mathbb{C} \times 0 \cong \mathbb{C}$) since it is not a unital module over these copies of $\mathbb{C}$ (the 1 of these subrings is not the same as the 1 of A).

10.2 QUOTIENT MODULES AND MODULE HOMOMORPHISMS

This section contains the basic theory of quotient modules and module homomorphisms.

Definition. Let R be a ring and let M and N be R-modules.

(1) A map $\varphi : M \to N$ is an *R-module homomorphism* if it respects the R-module structures of M and N, i.e.,

(a) $\varphi(x + y) = \varphi(x) + \varphi(y)$, for all $x, y \in M$ and

(b) $\varphi(rx) = r\varphi(x)$, for all $r \in R$, $x \in M$.

(2) An R-module homomorphism is an *isomorphism (of R-modules)* if it is both injective and surjective. The modules M and N are said to be *isomorphic*, denoted $M \cong N$, if there is some R-module isomorphism $\varphi : M \to N$.

(3) If $\varphi : M \to N$ is an R-module homomorphism, let $\ker \varphi = \{m \in M \mid \varphi(m) = 0\}$ (the *kernel* of φ) and let $\varphi(M) = \{n \in N \mid n = \varphi(m) \text{ for some } m \in M\}$ (the *image* of φ, as usual).

(4) Let M and N be R-modules and define $\text{Hom}_R(M, N)$ to be the set of all R-module homomorphisms from M into N.

Any R-module homomorphism is also a homomorphism of the additive groups, but not every group homomorphism need be a module homomorphism (because condition (b) may not be satisfied). The unqualified term "isomorphism" when applied to R-modules will always mean R-module isomorphism. When the symbol $\cong$ is used without qualification it will denote an isomorphism of the respective structures (which will be evident from the context).

It is an easy exercise using the submodule criterion (Proposition 1) to show that kernels and images of R-module homomorphisms are submodules.

Examples

(1) If R is a ring and $M = R$ is a module over itself, then R-module homomorphisms (even from R to itself) need not be ring homomorphisms and ring homomorphisms need not be R-module homomorphisms. For example, when $R = \mathbb{Z}$ the $\mathbb{Z}$-module homomorphism $x \mapsto 2x$ is not a ring homomorphism (1 does not map to 1). When $R = F[x]$ the ring homomorphism $\varphi : f(x) \mapsto f(x^2)$ is not an $F[x]$-module homomorphism (if it were, we would have $x^2 = \varphi(x) = \varphi(x \cdot 1) = x\varphi(1) = x$).

(2) Let R be a ring, let $n \in \mathbb{Z}^+$ and let $M = R^n$. One easily checks that for each $i \in \{1, \ldots, n\}$ the projection map

$$\pi_i : R^n \to R \quad \text{by} \quad \pi_i(x_1, \ldots, x_n) = x_i$$

is a surjective R-module homomorphism with kernel equal to the submodule of n-tuples which have a zero in position i.

(3) If R is a field, R-module homomorphisms are called *linear transformations*. These will be studied extensively in Chapter 11.

(4) For the ring $R = \mathbb{Z}$ the action of ring elements (integers) on any $\mathbb{Z}$-module amounts to just adding and subtracting within the (additive) abelian group structure of the module so that in this case condition (b) of a homomorphism is implied by condition (a). For example, $\varphi(2x) = \varphi(x + x) = \varphi(x) + \varphi(x) = 2\varphi(x)$, etc. It follows that

$\mathbb{Z}$-module homomorphisms are the same as abelian group homomorphisms.

(5) Let R be a ring, let I be a 2-sided ideal of R and suppose M and N are R-modules annihilated by I (i.e., $am = 0$ and $an = 0$ for all $a \in I$, $n \in N$ and $m \in M$). Any R-module homomorphism from N to M is then automatically a homomorphism of (R/I)-modules (see Example 5 of Section 1). In particular, if A is an additive abelian group such that for some prime p, $px = 0$ for all $x \in A$, then any group homomorphism from A to itself is a $\mathbb{Z}/p\mathbb{Z}$-module homomorphism, i.e., is a linear transformation over the field $\mathbb{F}_p$. In particular, the group of all (group) automorphisms of A is the group of invertible linear transformations from A to itself: $GL(A)$.

Proposition 2. Let M, N and L be R-modules.

(1) A map $\varphi : M \to N$ is an R-module homomorphism if and only if $\varphi(rx + y) = r\varphi(x) + \varphi(y)$ for all $x, y \in M$ and all $r \in R$.

(2) Let φ, ψ be elements of $\mathrm{Hom}_R(M, N)$. Define $\varphi + \psi$ by

$$(\varphi + \psi)(m) = \varphi(m) + \psi(m) \qquad \text{for all } m \in M.$$

Then $\varphi + \psi \in \mathrm{Hom}_R(M, N)$ and with this operation $\mathrm{Hom}_R(M, N)$ is an abelian group. If R is a commutative ring then for $r \in R$ define $r\varphi$ by

$$(r\varphi)(m) = r(\varphi(m)) \qquad \text{for all } m \in M.$$

Then $r\varphi \in \mathrm{Hom}_R(M, N)$ and with this action of the commutative ring R the abelian group $\mathrm{Hom}_R(M, N)$ is an R-module.

(3) If $\varphi \in \mathrm{Hom}_R(L, M)$ and $\psi \in \mathrm{Hom}_R(M, N)$ then $\psi \circ \varphi \in \mathrm{Hom}_R(L, N)$.

(4) With addition as above and multiplication defined as function composition, $\mathrm{Hom}_R(M, M)$ is a ring with 1. When R is commutative $\mathrm{Hom}_R(M, M)$ is an R-algebra.

Proof: (1) Certainly $\varphi(rx+y) = r\varphi(x)+\varphi(y)$ if φ is an R-module homomorphism. Conversely, if $\varphi(rx + y) = r\varphi(x) + \varphi(y)$, take $r = 1$ to see that φ is additive and take $y = 0$ to see that φ commutes with the action of R on M (i.e., is *homogeneous*).

(2) It is straightforward to check that all the abelian group and R-module axioms hold with these definitions — the details are left as an exercise. We note that the commutativity of R is used to show that $r\varphi$ satisfies the second axiom of an R-module homomorphism, namely,

$$\begin{aligned}
(r_1\varphi)(r_2m) &= r_1\varphi(r_2m) && \text{(by definition of } r_1\varphi) \\
&= r_1r_2(\varphi(m)) && \text{(since } \varphi \text{ is a homomorphism)} \\
&= r_2r_1\varphi(m) && \text{(since } R \text{ is commutative)} \\
&= r_2(r_1\varphi)(m) && \text{(by definition of } r_1\varphi).
\end{aligned}$$

Verification of the axioms relies ultimately on the hypothesis that N is an R-module. The domain M could in fact be any set — it does not have to be an R-module nor an abelian group.

(3) Let φ and ψ be as given and let $r \in R$, $x, y \in L$. Then

$$\begin{aligned}
(\psi \circ \varphi)(rx + y) &= \psi(\varphi(rx + y)) \\
&= \psi(r\varphi(x) + \varphi(y)) && \text{(by (1) applied to } \varphi) \\
&= r\psi(\varphi(x)) + \psi(\varphi(y)) && \text{(by (1) applied to } \psi) \\
&= r(\psi \circ \varphi)(x) + (\psi \circ \varphi)(y)
\end{aligned}$$

so, by (1), $\psi \circ \varphi$ is an R-module homomorphism.

(4) Note that since the domain and codomain of the elements of $\mathrm{Hom}_R(M, M)$ are the same, function composition is defined. By (3), it is a binary operation on $\mathrm{Hom}_R(M, M)$. As usual, function composition is associative. The remaining ring axioms are straightforward to check — the details are left as an exercise. The identity function, I, (as usual, $I(x) = x$, for all $x \in M$) is seen to be the multiplicative identity of $\mathrm{Hom}_R(M, M)$. If R is commutative, then (2) shows that the ring $\mathrm{Hom}_R(M, M)$ is a left R-module and defining $\varphi r = r\varphi$ for all $\varphi \in \mathrm{Hom}_R(M, M)$ and $r \in R$ makes $\mathrm{Hom}_R(M, M)$ into an R-algebra.

Definition. The ring $\mathrm{Hom}_R(M, M)$ is called the *endomorphism ring of* M and will often be denoted by $\mathrm{End}_R(M)$, or just $\mathrm{End}(M)$ when the ring R is clear from the context. Elements of $\mathrm{End}(M)$ are called *endomorphisms*.

When R is commutative there is a natural map from R into End(M) given by $r \mapsto rI$, where the latter endomorphism of M is just multiplication by r on M (cf. Exercise 7). The image of R is contained in the center of End(M) so if R has an identity, End(M) is an R-algebra. The ring homomorphism (cf. Exercise 7) from R to $\text{End}_R(M)$ may not be injective since for some r we may have $rm = 0$ for all $m \in M$ (e.g., $R = \mathbb{Z}$, $M = \mathbb{Z}/2\mathbb{Z}$, and $r = 2$). When R is a field, however, this map is injective (in general, no unit is in the kernel of this map) and the copy of R in $\text{End}_R(M)$ is called the (subring of) *scalar transformations*.

Next we prove that every submodule N of an R-module M is "normal" in the sense that we can *always* form the quotient module M/N, and the natural projection $\pi : M \to M/N$ is an R-module homomorphism with kernel N. The proof of this fact and, more generally, the subsequent proofs of the isomorphism theorems for modules follow easily from the corresponding facts for groups. The reason for this is because a module is first of all an *abelian* group and so *every* submodule is automatically a normal subgroup and any module homomorphism is, in particular, a homomorphism of abelian groups, all of which we have already considered in Chapter 3. What remains to be proved in order to extend results on abelian groups to corresponding results on modules is to check that the action of R is compatible with these group quotients and homomorphisms. For example, the map π above was shown to be a group homomorphism in Chapter 3 but the abelian group M/N must be shown to be an R-module (i.e., to have an action by R) and property (b) in the definition of a module homomorphism must be checked for π.

Proposition 3. Let R be a ring, let M be an R-module and let N be a submodule of M. The (additive, abelian) quotient group M/N can be made into an R-module by defining an action of elements of R by

$$r(x + N) = (rx) + N, \qquad \text{for all } r \in R,\ x + N \in M/N.$$

The natural projection map $\pi : M \to M/N$ defined by $\pi(x) = x + N$ is an R-module homomorphism with kernel N.

Proof: Since M is an abelian group under $+$ the quotient group M/N is defined and is an abelian group. To see that the action of the ring element r on the coset $x+N$ is well defined, suppose $x+N = y+N$, i.e., $x-y \in N$. Since N is a (left) R-submodule, $r(x - y) \in N$. Thus $rx - ry \in N$ and $rx + N = ry + N$, as desired. Now since the operations in M/N are "compatible" with those of M, the axioms for an R-module are easily checked in the same way as was done for quotient groups. For example, axiom 2(b) holds as follows: for all $r_1, r_2 \in R$ and $x + N \in M/N$, by definition of the action of ring elements on elements of M/N

$$\begin{aligned}(r_1r_2)(x + N) &= (r_1r_2x) + N \\ &= r_1(r_2x + N) \\ &= r_1(r_2(x + N)).\end{aligned}$$

The other axioms are similarly checked — the details are left as an exercise. Finally, the natural projection map π described above is, in particular, the natural projection of the abelian group M onto the abelian group M/N hence is a group homomorphism with kernel N. The kernel of any module homomorphism is the same as its kernel when viewed as a homomorphism of the abelian group structures. It remains only to show π is a module homomorphism, i.e., $\pi(rm) = r\pi(m)$. But

$$\begin{aligned}\pi(rm) &= rm + N \\ &= r(m + N) \qquad \text{(by definition of the action of } R \text{ on } M/N) \\ &= r\pi(m).\end{aligned}$$

This completes the proof.

All the isomorphism theorems stated for groups also hold for R-modules. The proofs are similar to that of Proposition 3 above in that they begin by invoking the corresponding theorem for groups and then prove that the group homomorphisms are also R-module homomorphisms. To state the Second Isomorphism Theorem we need the following.

Definition. Let A, B be submodules of the R-module M. The *sum* of A and B is the set

$$A + B = \{a + b \mid a \in A, b \in B\}.$$

One can easily check that the sum of two submodules A and B is a submodule and is the smallest submodule which contains both A and B.

Theorem 4. (Isomorphism Theorems)

(1) *(The First Isomorphism Theorem for Modules)* Let M, N be R-modules and let $\varphi : M \to N$ be an R-module homomorphism. Then $\ker\varphi$ is a submodule of M and $M/\ker\varphi \cong \varphi(M)$.

(2) *(The Second Isomorphism Theorem)* Let A, B be submodules of the R-module M. Then $(A + B)/B \cong A/(A \cap B)$.

(3) *(The Third Isomorphism Theorem)* Let M be an R-module, and let A and B be submodules of M with $A \subseteq B$. Then $(M/A)/(B/A) \cong M/B$.

(4) *(The Fourth or Lattice Isomorphism Theorem)* Let N be a submodule of the R-module M. There is a bijection between the submodules of M which contain N and the submodules of M/N. The correspondence is given by $A \leftrightarrow A/N$, for all $A \supseteq N$. This correspondence commutes with the processes of taking sums and intersections (i.e., is a lattice isomorphism between the lattice of submodules of M/N and the lattice of submodules of M which contain N).

Proof: Exercise.

EXERCISES

In these exercises R is a ring with 1 and M is a left R-module.

1. Use the submodule criterion to show that kernels and images of R-module homomorphisms are submodules.
2. Show that the relation "is R-module isomorphic to" is an equivalence relation on any set of R-modules.
3. Give an explicit example of a map from one R-module to another which is a group homomorphism but not an R-module homomorphism.
4. Let A be any $\mathbb{Z}$-module, let a be any element of A and let n be a positive integer. Prove that the map $\varphi_a : \mathbb{Z}/n\mathbb{Z} \to A$ given by $\varphi(\bar{k}) = ka$ is a well defined $\mathbb{Z}$-module homomorphism if and only if $na = 0$. Prove that $\mathrm{Hom}_{\mathbb{Z}}(\mathbb{Z}/n\mathbb{Z}, A) \cong A_n$, where $A_n = \{a \in A \mid na = 0\}$ (so A_n is the annihilator in A of the ideal (n) of $\mathbb{Z}$ — cf. Exercise 10, Section 1).
5. Exhibit all $\mathbb{Z}$-module homomorphisms from $\mathbb{Z}/30\mathbb{Z}$ to $\mathbb{Z}/21\mathbb{Z}$.
6. Prove that $\mathrm{Hom}_{\mathbb{Z}}(\mathbb{Z}/n\mathbb{Z}, \mathbb{Z}/m\mathbb{Z}) \cong \mathbb{Z}/(n, m)\mathbb{Z}$.
7. Let z be a fixed element of the center of R. Prove that the map $m \mapsto zm$ is an R-module homomorphism from M to itself. Show that for a commutative ring R the map from R to $\mathrm{End}_R(M)$ given by $r \mapsto rI$ is a ring homomorphism (where I is the identity endomorphism).
8. Let $\varphi : M \to N$ be an R-module homomorphism. Prove that $\varphi(\mathrm{Tor}(M)) \subseteq \mathrm{Tor}(N)$ (cf. Exercise 8 in Section 1).
9. Let R be a commutative ring. Prove that $\mathrm{Hom}_R(R, M)$ and M are isomorphic as left R-modules. [Show that each element of $\mathrm{Hom}_R(R, M)$ is determined by its value on the identity of R.]
10. Let R be a commutative ring. Prove that $\mathrm{Hom}_R(R, R)$ and R are isomorphic as rings.
11. Let $A_1, A_2, \ldots, A_n$ be R-modules and let B_i be a submodule of A_i for each $i = 1, 2, \ldots, n$. Prove that
$$(A_1 \times \cdots \times A_n)/(B_1 \times \cdots \times B_n) \cong (A_1/B_1) \times \cdots \times (A_n/B_n).$$
[Recall Exercise 14 in Section 5.1.]
12. Let I be a left ideal of R and let n be a positive integer. Prove
$$R^n/IR^n \cong R/IR \times \cdots \times R/IR \quad (n \text{ times})$$
where IR^n is defined as in Exercise 5 of Section 1. [Use the preceding exercise.]
13. Let I be a nilpotent ideal in a commutative ring R (cf. Exercise 37, Section 7.3), let M and N be R-modules and let $\varphi : M \to N$ be an R-module homomorphism. Show that if the induced map $\overline{\varphi} : M/IM \to N/IN$ is surjective, then φ is surjective.
14. Let $R = \mathbb{Z}[x]$ be the ring of polynomials in x and let $A = \mathbb{Z}[t_1, t_2, \ldots]$ be the ring of polynomials in the independent indeterminates $t_1, t_2, \ldots$. Define an action of R on A as follows: 1) let $1 \in R$ act on A as the identity, 2) for $n \geq 1$ let $x^n \circ 1 = t_n$, let $x^n \circ t_i = t_{n+i}$ for $i = 1, 2, \ldots$, and let x^n act as 0 on monomials in A of (total) degree at least two, and 3) extend $\mathbb{Z}$-linearly, i.e., so that the module axioms 2(a) and 2(c) are satisfied.
   (a) Show that $x^{p+q} \circ t_i = x^p \circ (x^q \circ t_i) = t_{p+q+i}$ and use this to show that under this action the ring A is a (unital) R-module.
   (b) Show that the map $\varphi : R \to A$ defined by $\varphi(r) = r \circ 1_A$ is an R-module homomorphism of the ring R into the ring A mapping 1_R to 1_A, but is not a ring homomorphism from R to A.

10.3 GENERATION OF MODULES, DIRECT SUMS, AND FREE MODULES

Let R be a ring with 1. As in the preceding sections the term "module" will mean "left module." We first extend the notion of the sum of two submodules to sums of any finite number of submodules and define the submodule generated by a subset.

Definition. Let M be an R-module and let $N_1, \ldots, N_n$ be submodules of M.

(1) The *sum* of $N_1, \ldots, N_n$ is the set of all finite sums of elements from the sets N_i: $\{a_1 + a_2 + \cdots + a_n \mid a_i \in N_i \text{ for all } i\}$. Denote this sum by $N_1 + \cdots + N_n$.

(2) For any subset A of M let

$$RA = \{r_1a_1 + r_2a_2 + \cdots + r_ma_m \mid r_1, \ldots, r_m \in R,\ a_1, \ldots, a_m \in A,\ m \in \mathbb{Z}^+\}$$

(where by convention $RA = \{0\}$ if $A = \emptyset$). If A is the finite set $\{a_1, a_2, \ldots, a_n\}$ we shall write $Ra_1 + Ra_2 + \cdots + Ra_n$ for RA. Call RA the *submodule of M generated by A*. If N is a submodule of M (possibly $N = M$) and $N = RA$, for some subset A of M, we call A a *set of generators* or *generating set* for N, and we say N *is generated by* A.

(3) A submodule N of M (possibly $N = M$) is *finitely generated* if there is some finite subset A of M such that $N = RA$, that is, if N is generated by some finite subset.

(4) A submodule N of M (possibly $N = M$) is *cyclic* if there exists an element $a \in M$ such that $N = Ra$, that is, if N is generated by one element:

$$N = Ra = \{ra \mid r \in R\}.$$

Note that these definitions do not require that the ring R contain a 1, however this condition ensures that A is contained in RA. It is easy to see using the Submodule Criterion that for any subset A of M, RA is indeed a submodule of M and is the smallest submodule of M which contains A (i.e., any submodule of M which contains A also contains RA). In particular, for submodules $N_1, \ldots, N_n$ of M, $N_1 + \cdots + N_n$ is just the submodule generated by the set $N_1 \cup \cdots \cup N_n$ and is the smallest submodule of M containing N_i, for all i. If $N_1, \ldots, N_n$ are generated by sets $A_1, \ldots, A_n$ respectively, then $N_1 + \cdots + N_n$ is generated by $A_1 \cup \cdots \cup A_n$. Note that cyclic modules are, a fortiori, finitely generated.

A submodule N of an R-module M may have many different generating sets (for instance the set N itself always generates N). If N is finitely generated, then there is a smallest nonnegative integer d such that N is generated by d elements (and no fewer). Any generating set consisting of d elements will be called a *minimal set of generators for N* (it is not unique in general). If N is not finitely generated, it need not have a minimal generating set.

The process of generating submodules of an R-module M by taking subsets A of M and forming all finite "R-linear combinations" of elements of A will be our primary way of producing submodules (this notion is perhaps familiar from vector space theory where it is referred to as taking the *span* of A). The obstruction which made the analogous process so difficult for groups in general was the noncommutativity of group

operations. For abelian groups, G, however, it was much simpler to control the subgroup $\langle A \rangle$ generated by A, for a subset A of G (see Section 2.4 for the complete discussion of this). The situation for R-modules is similar to that of abelian groups (even if R is a noncommutative ring) because we can always collect "like terms" in elements of A, i.e., terms such as $r_1a_1 + r_2a_2 + s_1a_1$ can always be simplified to $(r_1 + s_1)a_1 + r_2a_2$. This again reflects the underlying abelian group structure of modules.

Examples

(1) Let $R = \mathbb{Z}$ and let M be any R-module, that is, any abelian group. If $a \in M$, then $\mathbb{Z}a$ is just the cyclic subgroup of M generated by a: $\langle a \rangle$ (compare Definition 4 above with the definition of a cyclic group). More generally, M is generated as a $\mathbb{Z}$-module by a set A if and only if M is generated as a group by A (that is, the action of ring elements in this instance produces no elements that cannot already be obtained from A by addition and subtraction). The definition of finitely generated for $\mathbb{Z}$-modules is identical to that for abelian groups found in Chapter 5.

(2) Let R be a ring with 1 and let M be the (left) R-module R itself. Note that R is a finitely generated, in fact cyclic, R-module because $R = R1$ (i.e., we can take $A = \{1\}$). Recall that the submodules of R are precisely the left ideals of R, so saying I is a cyclic R-submodule of the left R-module R is the same as saying I is a principal ideal of R (usually the term "principal ideal" is used in the context of commutative rings). Also, saying I is a finitely generated R-submodule of R is the same as saying I is a finitely generated ideal. When R is a commutative ring we often write AR or aR for the submodule (ideal) generated by A or a respectively, as we have been doing for $\mathbb{Z}$ when we wrote $n\mathbb{Z}$. In this situation $AR = RA$ and $aR = Ra$ (elementwise as well). Thus a Principal Ideal Domain is a (commutative) integral domain R with identity in which every R-submodule of R is cyclic.

Submodules of a finitely generated module need not be finitely generated: take M to be the cyclic R-module R itself where R is the polynomial ring in infinitely many variables $x_1, x_2, x_3, \ldots$ with coefficients in some field F. The submodule (i.e., 2-sided ideal) generated by $\{x_1, x_2, \ldots\}$ cannot be generated by any finite set (note that one must show that *no* finite subset of this ideal will generate it).

(3) Let R be a ring with 1 and let M be the free module of rank n over R, as described in the first section. For each $i \in \{1, 2, \ldots, n\}$ let $e_i = (0, 0, \ldots, 0, 1, 0, \ldots, 0)$, where the 1 appears in position i. Since

$$(s_1, s_2, \ldots, s_n) = \sum_{i=1}^{n} s_i e_i$$

it is clear that M is generated by $\{e_1, \ldots, e_n\}$. If R is commutative then this is a *minimal* generating set (cf. Exercises 2 and 27).

(4) Let F be a field, let x be an indeterminate, let V be a vector space over F and let T be a linear transformation from V to V. Make V into an $F[x]$-module via T. Then V is a *cyclic* $F[x]$-module (with generator v) if and only if $V = \{p(x)v \mid p(x) \in F[x]\}$, that is, if and only if every element of V can be written as an F-linear combination of elements of the set $\{T^n(v) \mid n \geq 0\}$. This in turn is equivalent to saying $\{v, T(v), T^2(v), \ldots\}$ span V as a vector space over F.

For instance if T is the identity linear transformation from V to V or the zero linear transformation, then for every $v \in V$ and every $p(x) \in F[x]$ we have $p(x)v = \alpha v$ for some $\alpha \in F$. Thus if V has dimension > 1, V cannot be a cyclic $F[x]$-module.

For another example suppose V is affine n-space and T is the "shift operator" described in Section 1. Let e_i be the i^{th} basis vector (as usual) numbered so that T is defined by $T^k(e_n) = e_{n-k}$ for $1 \leq k < n$. Thus V is spanned by the elements $e_n, T(e_n), \ldots, T^{n-1}(e_n)$, that is, V is a cyclic $F[x]$-module with generator e_n. For $n > 1$, V is not, however, a cyclic F-module (i.e., is not a 1-dimensional vector space over F).

Definition. Let $M_1, \ldots, M_k$ be a collection of R-modules. The collection of k-tuples $(m_1, m_2, \ldots, m_k)$ where $m_i \in M_i$ with addition and action of R defined componentwise is called the *direct product* of $M_1, \ldots, M_k$, denoted $M_1 \times \cdots \times M_k$.

It is evident that the direct product of a collection of R-modules is again an R-module. The direct product of $M_1, \ldots, M_k$ is also referred to as the *(external) direct sum* of $M_1, \ldots, M_k$ and denoted $M_1 \oplus \cdots \oplus M_k$. The direct product and direct sum of an infinite number of modules (which are different in general) are defined in Exercise 20.

The next proposition indicates when a module is isomorphic to the direct product of some of its submodules and is the analogue for modules of Theorem 9 in Section 5.4 (which determines when a group is the direct product of two of its subgroups).

Proposition 5. Let $N_1, N_2, \ldots, N_k$ be submodules of the R-module M. Then the following are equivalent:

(1) The map $\pi : N_1 \times N_2 \times \cdots \times N_k \to N_1 + N_2 + \cdots + N_k$ defined by

$$\pi(a_1, a_2, \ldots, a_k) = a_1 + a_2 + \cdots + a_k$$

is an isomorphism (of R-modules): $N_1 + N_2 + \cdots + N_k \cong N_1 \times N_2 \times \cdots \times N_k$.

(2) $N_j \cap (N_1 + N_2 + \cdots + N_{j-1} + N_{j+1} + \cdots + N_k) = 0$ for all $j \in \{1, 2, \ldots, k\}$.

(3) Every $x \in N_1 + \cdots + N_k$ can be written *uniquely* in the form $a_1 + a_2 + \cdots + a_k$ with $a_i \in N_i$.

Proof: To prove (1) implies (2), suppose for some j that (2) fails to hold and let $a_j \in (N_1 + \cdots + N_{j-1} + N_{j+1} + \cdots + N_k) \cap N_j$, with $a_j \neq 0$. Then

$$a_j = a_1 + \cdots + a_{j-1} + a_{j+1} + \cdots + a_k$$

for some $a_i \in N_i$, and $(a_1, \ldots, a_{j-1}, -a_j, a_{j+1}, \ldots, a_k)$ would be a nonzero element of $\ker \pi$, a contradiction.

Assume now that (2) holds. If for some module elements $a_i, b_i \in N_i$ we have

$$a_1 + a_2 + \cdots + a_k = b_1 + b_2 + \cdots + b_k$$

then for each j we have

$$a_j - b_j = (b_1 - a_1) + \cdots + (b_{j-1} - a_{j-1}) + (b_{j+1} - a_{j+1}) + \cdots + (b_k - a_k).$$

The left hand side is in N_j and the right side belongs to $N_1 + \cdots + N_{j-1} + N_{j+1} + \cdots + N_k$. Thus

$$a_j - b_j \in N_j \cap (N_1 + \cdots + N_{j-1} + N_{j+1} + \cdots + N_k) = 0.$$

This shows $a_j = b_j$ for all j, and so (2) implies (3).

Finally, to see that (3) implies (1) observe first that the map π is clearly a surjective R-module homomorphism. Then (3) simply implies π is injective, hence is an isomorphism, completing the proof.

If an R-module $M = N_1 + N_2 + \cdots + N_k$ is the sum of submodules $N_1, N_2, \ldots, N_k$ of M satisfying the equivalent conditions of the proposition above, then M is said to be the *(internal) direct sum* of $N_1, N_2, \ldots, N_k$, written

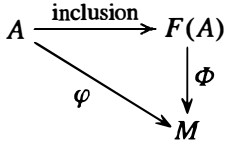
$$M = N_1 \oplus N_2 \oplus \cdots \oplus N_k.$$

By the proposition, this is equivalent to the assertion that every element m of M can be written *uniquely* as a sum of elements $m = n_1 + n_2 + \cdots + n_k$ with $n_i \in N_i$. (Note that part (1) of the proposition is the statement that the internal direct sum of $N_1, N_2, \ldots, N_k$ is isomorphic to their external direct sum, which is the reason we identify them and use the same notation for both.)

Definition. An R-module F is said to be *free* on the subset A of F if for every nonzero element x of F, there exist unique nonzero elements $r_1, r_2, \ldots, r_n$ of R and unique $a_1, a_2, \ldots, a_n$ in A such that $x = r_1a_1 + r_2a_2 + \cdots + r_na_n$, for some $n \in \mathbb{Z}^+$. In this situation we say A is a *basis* or *set of free generators* for F. If R is a commutative ring the cardinality of A is called the *rank* of F (cf. Exercise 27).

One should be careful to note the difference between the uniqueness property of direct sums (Proposition 5(3)) and the uniqueness property of free modules. Namely, in the direct sum of two modules, say $N_1 \oplus N_2$, each element can be written uniquely as $n_1 + n_2$; here the uniqueness refers to the *module elements* n_1 and n_2. In the case of free modules, the uniqueness is on the *ring elements as well as the module elements.* For example, if $R = \mathbb{Z}$ and $N_1 = N_2 = \mathbb{Z}/2\mathbb{Z}$, then each element of $N_1 \oplus N_2$ has a unique representation in the form $n_1 + n_2$ where each $n_i \in N_i$, however n_1 (for instance) can be expressed as n_1 or $3n_1$ or $5n_1 \ldots$ etc., so each element does not have a unique representation in the form $r_1a_1 + r_2a_2$, where $r_1, r_2 \in R$, $a_1 \in N_1$ and $a_2 \in N_2$. Thus $\mathbb{Z}/2\mathbb{Z} \oplus \mathbb{Z}/2\mathbb{Z}$ is not a free $\mathbb{Z}$-module on the set $\{(1, 0), (0, 1)\}$. Similarly, it is not free on any set.

Theorem 6. For any set A there is a free R-module $F(A)$ on the set A and $F(A)$ satisfies the following *universal property:* if M is any R-module and $\varphi : A \to M$ is any map of sets, then there is a unique R-module homomorphism $\Phi : F(A) \to M$ such that $\Phi(a) = \varphi(a)$, for all $a \in A$, that is, the following diagram commutes.



When A is the finite set $\{a_1, a_2, \ldots, a_n\}$, $F(A) = Ra_1 \oplus Ra_2 \oplus \cdots \oplus Ra_n \cong R^n$. (Compare: Section 6.3, free groups.)

Proof: Let $F(A) = \{0\}$ if $A = \emptyset$. If A is nonempty let $F(A)$ be the collection of all set functions $f : A \to R$ such that $f(a) = 0$ for all but finitely many $a \in A$. Make

$F(A)$ into an R-module by pointwise addition of functions and pointwise multiplication of a ring element times a function, i.e.,

$$(f+g)(a) = f(a) + g(a) \quad \text{and}$$
$$(rf)(a) = r(f(a)), \qquad \text{for all } a \in A,\ r \in R \text{ and } f, g \in F(A).$$

It is an easy matter to check that all the R-module axioms hold (the details are omitted). Identify A as a subset of $F(A)$ by $a \mapsto f_a$, where f_a is the function which is 1 at a and zero elsewhere. We can, in this way, think of $F(A)$ as all finite R-linear combinations of elements of A by identifying each function f with the sum $r_1a_1 + r_2a_2 + \cdots + r_na_n$, where f takes on the value r_i at a_i and is zero at all other elements of A. Moreover, each element of $F(A)$ has a unique expression as such a formal sum. To establish the universal property of $F(A)$ suppose $\varphi : A \to M$ is a map of the set A into the R-module M. Define $\Phi : F(A) \to M$ by

$$\Phi : \sum_{i=1}^{n} r_i a_i \mapsto \sum_{i=1}^{n} r_i \varphi(a_i).$$

By the uniqueness of the expression for the elements of $F(A)$ as linear combinations of the a_i we see easily that Φ is a well defined R-module homomorphism (the details are left as an exercise). By definition, the restriction of Φ to A equals φ. Finally, since $F(A)$ is generated by A, once we know the values of an R-module homomorphism on A its values on every element of $F(A)$ are uniquely determined, so Φ is the unique extension of φ to all of $F(A)$.

When A is the finite set $\{a_1, a_2, \ldots, a_n\}$ Proposition 5(3) shows that $F(A) = Ra_1 \oplus Ra_2 \oplus \cdots \oplus Ra_n$. Since $R \cong Ra_i$ for all i (under the map $r \mapsto ra_i$) Proposition 5(1) shows that the direct sum is isomorphic to R^n.

Corollary 7.

(1) If F_1 and F_2 are free modules on the same set A, there is a unique isomorphism between F_1 and F_2 which is the identity map on A.

(2) If F is any free R-module with basis A, then $F \cong F(A)$. In particular, F enjoys the same universal property with respect to A as $F(A)$ does in Theorem 6.

Proof: Exercise.

If F is a free R-module with basis A, we shall often (particularly in the case of vector spaces) define R-module homomorphisms from F into other R-modules simply by specifying their values on the elements of A and then saying "*extend by linearity.*" Corollary 7(2) ensures that this is permissible.

When $R = \mathbb{Z}$, the free module on a set A is called the *free abelian group on A*. If $|A| = n$, $F(A)$ is called the free abelian group of *rank n* and is isomorphic to $\mathbb{Z} \oplus \cdots \oplus \mathbb{Z}$ (n times). These definitions agree with the ones given in Chapter 5.

EXERCISES

In these exercises R is a ring with 1 and M is a left R-module.

1. Prove that if A and B are sets of the same cardinality, then the free modules $F(A)$ and $F(B)$ are isomorphic.
2. Assume R is commutative. Prove that $R^n \cong R^m$ if and only if $n = m$, i.e., two free R-modules of finite rank are isomorphic if and only if they have the same rank. [Apply Exercise 12 of Section 2 with I a maximal ideal of R. You may assume that if F is a field, then $F^n \cong F^m$ if and only if $n = m$, i.e., two finite dimensional vector spaces over F are isomorphic if and only if they have the same dimension — this will be proved later in Section 11.1.]
3. Show that the $F[x]$-modules in Exercises 18 and 19 of Section 1 are both cyclic.
4. An R-module M is called a *torsion* module if for each $m \in M$ there is a nonzero element $r \in R$ such that $rm = 0$, where r may depend on m (i.e., $M = \mathrm{Tor}(M)$ in the notation of Exercise 8 of Section 1). Prove that every finite abelian group is a torsion $\mathbb{Z}$-module. Give an example of an infinite abelian group that is a torsion $\mathbb{Z}$-module.
5. Let R be an integral domain. Prove that every finitely generated torsion R-module has a nonzero annihilator i.e., there is a nonzero element $r \in R$ such that $rm = 0$ for all $m \in M$ — here r does not depend on m (the annihilator of a module was defined in Exercise 9 of Section 1). Give an example of a torsion R-module whose annihilator is the zero ideal.
6. Prove that if M is a finitely generated R-module that is generated by n elements then every quotient of M may be generated by n (or fewer) elements. Deduce that quotients of cyclic modules are cyclic.
7. Let N be a submodule of M. Prove that if both M/N and N are finitely generated then so is M.
8. Let S be the collection of sequences $(a_1, a_2, a_3, \ldots)$ of integers $a_1, a_2, a_3, \ldots$ where all but finitely many of the a_i are 0 (called the *direct sum* of infinitely many copies of $\mathbb{Z}$). Recall that S is a ring under componentwise addition and multiplication and S does not have a multiplicative identity — cf. Exercise 20, Section 7.1. Prove that S is not finitely generated as a module over itself.
9. An R-module M is called *irreducible* if $M \neq 0$ and if 0 and M are the only submodules of M. Show that M is irreducible if and only if $M \neq 0$ and M is a cyclic module with any nonzero element as generator. Determine all the irreducible $\mathbb{Z}$-modules.
10. Assume R is commutative. Show that an R-module M is irreducible if and only if M is isomorphic (as an R-module) to R/I where I is a maximal ideal of R. [By the previous exercise, if M is irreducible there is a natural map $R \to M$ defined by $r \mapsto rm$, where m is any fixed nonzero element of M.]
11. Show that if M_1 and M_2 are irreducible R-modules, then any nonzero R-module homomorphism from M_1 to M_2 is an isomorphism. Deduce that if M is irreducible then $\mathrm{End}_R(M)$ is a division ring (this result is called *Schur's Lemma*). [Consider the kernel and the image.]
12. Let R be a commutative ring and let A, B and M be R-modules. Prove the following isomorphisms of R-modules:
    (a) $\mathrm{Hom}_R(A \times B, M) \cong \mathrm{Hom}_R(A, M) \times \mathrm{Hom}_R(B, M)$
    (b) $\mathrm{Hom}_R(M, A \times B) \cong \mathrm{Hom}_R(M, A) \times \mathrm{Hom}_R(M, B)$.
13. Let R be a commutative ring and let F be a free R-module of finite rank. Prove the following isomorphism of R-modules: $\mathrm{Hom}_R(F, R) \cong F$.

14. Let R be a commutative ring and let F be the free R-module of rank n. Prove that $\mathrm{Hom}_R(F, M) \cong M \times \cdots \times M$ (n times). [Use Exercise 9 in Section 2 and Exercise 12.]

15. An element $e \in R$ is called a *central idempotent* if $e^2 = e$ and $er = re$ for all $r \in R$. If e is a central idempotent in R, prove that $M = eM \oplus (1-e)M$. [Recall Exercise 14 in Section 1.]

The next two exercises establish the Chinese Remainder Theorem for modules (cf. Section 7.6).

16. For any ideal I of R let IM be the submodule defined in Exercise 5 of Section 1. Let $A_1, \ldots, A_k$ be any ideals in the ring R. Prove that the map

$$M \to M/A_1M \times \cdots \times M/A_kM \quad \text{defined by} \quad m \mapsto (m + A_1M, \ldots, m + A_kM)$$

is an R-module homomorphism with kernel $A_1M \cap A_2M \cap \cdots \cap A_kM$.

17. In the notation of the preceding exercise, assume further that the ideals $A_1, \ldots, A_k$ are pairwise comaximal (i.e., $A_i + A_j = R$ for all $i \neq j$). Prove that

$$M/(A_1 \cdots A_k)M \cong M/A_1M \times \cdots \times M/A_kM.$$

[See the proof of the Chinese Remainder Theorem for rings in Section 7.6.]

18. Let R be a Principal Ideal Domain and let M be an R-module that is annihilated by the nonzero, proper ideal (a). Let $a = p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_k^{\alpha_k}$ be the unique factorization of a into distinct prime powers in R. Let M_i be the annihilator of $p_i^{\alpha_i}$ in M, i.e., M_i is the set $\{m \in M \mid p_i^{\alpha_i} m = 0\}$ — called the *p_i-primary component of M*. Prove that

$$M = M_1 \oplus M_2 \oplus \cdots \oplus M_k.$$

19. Show that if M is a finite abelian group of order $a = p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_k^{\alpha_k}$ then, considered as a $\mathbb{Z}$-module, M is annihilated by (a), the p_i-primary component of M is the unique Sylow p_i-subgroup of M and M is isomorphic to the direct product of its Sylow subgroups.

20. Let I be a nonempty index set and for each $i \in I$ let M_i be an R-module. The *direct product* of the modules M_i is defined to be their direct product as abelian groups (cf. Exercise 15 in Section 5.1) with the action of R componentwise multiplication. The *direct sum* of the modules M_i is defined to be the restricted direct product of the abelian groups M_i (cf. Exercise 17 in Section 5.1) with the action of R componentwise multiplication. In other words, the direct sum of the M_i's is the subset of the direct product, $\prod_{i \in I} M_i$, which consists of all elements $\prod_{i \in I} m_i$ such that only finitely many of the components m_i are nonzero; the action of R on the direct product or direct sum is given by $r \prod_{i \in I} m_i = \prod_{i \in I} rm_i$ (cf. Appendix I for the definition of Cartesian products of infinitely many sets). The direct sum will be denoted by $\oplus_{i \in I} M_i$.

- **(a)** Prove that the direct product of the M_i's is an R-module and the direct sum of the M_i's is a submodule of their direct product.
- **(b)** Show that if $R = \mathbb{Z}$, $I = \mathbb{Z}^+$ and M_i is the cyclic group of order i for each i, then the direct sum of the M_i's is not isomorphic to their direct product. [Look at torsion.]

21. Let I be a nonempty index set and for each $i \in I$ let N_i be a submodule of M. Prove that the following are equivalent:

- **(i)** the submodule of M generated by all the N_i's is isomorphic to the direct sum of the N_i's
- **(ii)** if $\{i_1, i_2, \ldots, i_k\}$ is any finite subset of I then $N_{i_1} \cap (N_{i_2} + \cdots + N_{i_k}) = 0$
- **(iii)** if $\{i_1, i_2, \ldots, i_k\}$ is any finite subset of I then $N_1 + \cdots + N_k = N_1 \oplus \cdots \oplus N_k$
- **(iv)** for every element x of the submodule of M generated by the N_i's there are unique elements $a_i \in N_i$ for all $i \in I$ such that all but a finite number of the a_i are zero and x is the (finite) sum of the a_i.

22. Let R be a Principal Ideal Domain, let M be a torsion R-module (cf. Exercise 4) and let p be a prime in R (do not assume M is finitely generated, hence it need not have a nonzero annihilator — cf. Exercise 5). The *p-primary component* of M is the set of all elements of M that are annihilated by some positive power of p.

(a) Prove that the p-primary component is a submodule. [See Exercise 13 in Section 1.]

(b) Prove that this definition of p-primary component agrees with the one given in Exercise 18 when M has a nonzero annihilator.

(c) Prove that M is the (possibly infinite) direct sum of its p-primary components, as p runs over all primes of R.

23. Show that any direct sum of free R-modules is free.

24. *(An arbitrary direct product of free modules need not be free)* For each positive integer i let M_i be the free $\mathbb{Z}$-module $\mathbb{Z}$, and let M be the direct product $\prod_{i\in\mathbb{Z}^+} M_i$ (cf. Exercise 20). Each element of M can be written uniquely in the form $(a_1, a_2, a_3, \dots)$ with $a_i \in \mathbb{Z}$ for all i. Let N be the submodule of M consisting of all such tuples with only finitely many nonzero a_i. Assume M is a free $\mathbb{Z}$-module with basis $\mathcal{B}$.

(a) Show that N is countable.

(b) Show that there is some countable subset $\mathcal{B}_1$ of $\mathcal{B}$ such that N is contained in the submodule, N_1, generated by $\mathcal{B}_1$. Show also that N_1 is countable.

(c) Let $\overline{M} = M/N_1$. Show that $\overline{M}$ is a free $\mathbb{Z}$-module. Deduce that if $\overline{x}$ is any nonzero element of $\overline{M}$ then there are only finitely many distinct positive integers k such that $\overline{x} = k\overline{m}$ for some $m \in M$ (depending on k).

(d) Let $\mathcal{S} = \{(b_1, b_2, b_3, \dots) \mid b_i = \pm i! \text{ for all } i\}$. Prove that $\mathcal{S}$ is uncountable. Deduce that there is some $s \in \mathcal{S}$ with $s \notin N_1$.

(e) Show that the assumption M is free leads to a contradiction: By (d) we may choose $s \in \mathcal{S}$ with $s \notin N_1$. Show that for each positive integer k there is some $m \in M$ with $\overline{s} = k\overline{m}$, contrary to (c). [Use the fact that $N \subseteq N_1$.]

25. In the construction of direct limits, Exercise 8 of Section 7.6, show that if all A_i are R-modules and the maps ρ_{ij} are R-module homomorphisms, then the direct limit $A = \varinjlim A_i$ may be given the structure of an R-module in a natural way such that the maps $\rho_i : A_i \to A$ are all R-module homomorphisms. Verify the corresponding universal property (part (e)) for R-module homomorphisms $\varphi_i : A_i \to C$ commuting with the ρ_{ij}.

26. Carry out the analysis of the preceding exercise corresponding to inverse limits to show that an inverse limit of R-modules is an R-module satisfying the appropriate universal property (cf. Exercise 10 of Section 7.6).

27. *(Free modules over noncommutative rings need not have a unique rank)* Let M be the $\mathbb{Z}$-module $\mathbb{Z} \times \mathbb{Z} \times \cdots$ of Exercise 24 and let R be its endomorphism ring, $R = \mathrm{End}_{\mathbb{Z}}(M)$ (cf. Exercises 29 and 30 in Section 7.1). Define $\varphi_1, \varphi_2 \in R$ by

$$\varphi_1(a_1, a_2, a_3, \dots) = (a_1, a_3, a_5, \dots)$$
$$\varphi_2(a_1, a_2, a_3, \dots) = (a_2, a_4, a_6, \dots)$$

(a) Prove that $\{\varphi_1, \varphi_2\}$ is a free basis of the left R-module R. [Define the maps ψ_1 and ψ_2 by $\psi_1(a_1, a_2, \dots) = (a_1, 0, a_2, 0, \dots)$ and $\psi_2(a_1, a_2, \dots) = (0, a_1, 0, a_2, \dots)$. Verify that $\varphi_i\psi_i = 1$, $\varphi_1\psi_2 = 0 = \varphi_2\psi_1$ and $\psi_1\varphi_1 + \psi_2\varphi_2 = 1$. Use these relations to prove that φ_1, φ_2 are independent and generate R as a left R-module.]

(b) Use (a) to prove that $R \cong R^2$ and deduce that $R \cong R^n$ for all $n \in \mathbb{Z}^+$.

10.4 TENSOR PRODUCTS OF MODULES

In this section we study the tensor product of two modules M and N over a ring (not necessarily commutative) containing 1. Formation of the tensor product is a general construction that, loosely speaking, enables one to form another module in which one can take "products" mn of elements $m \in M$ and $n \in N$. The general construction involves various left- and right- module actions, and it is instructive, by way of motivation, to first consider an important special case: the question of "extending scalars" or "changing the base."

Suppose that the ring R is a subring of the ring S. Throughout this section, we always assume that $1_R = 1_S$ (this ensures that S is a unital R-module).

If N is a left S-module, then N can also be naturally considered as a left R-module since the elements of R (being elements of S) act on N by assumption. The S-module axioms for N include the relations

$$(s_1 + s_2)n = s_1n + s_2n \quad \text{and} \quad s(n_1 + n_2) = sn_1 + sn_2 \tag{10.1}$$

for all $s, s_1, s_2 \in S$ and all $n, n_1, n_2 \in N$, and the relation

$$(s_1s_2)n = s_1(s_2n) \quad \text{for all } s_1, s_2 \in S, \text{ and all } n \in N. \tag{10.2}$$

A particular case of the latter relation is

$$(sr)n = s(rn) \quad \text{for all } s \in S, r \in R \text{ and } n \in N. \tag{10.2'}$$

More generally, if $f : R \to S$ is a ring homomorphism from R into S with $f(1_R) = 1_S$ (for example the injection map if R is a subring of S as above) then it is easy to see that N can be considered as an R-module with $rn = f(r)n$ for $r \in R$ and $n \in N$. In this situation S can be considered as an *extension* of the ring R and the resulting R-module is said to be obtained from N by *restriction of scalars* from S to R.

Suppose now that R is a subring of S and we try to reverse this, namely we start with an R-module N and attempt to define an S-module structure on N that extends the action of R on N to an action of S on N (hence "extending the scalars" from R to S). In general this is impossible, even in the simplest situation: the ring R itself is an R-module but is usually not an S-module for the larger ring S. For example, $\mathbb{Z}$ is a $\mathbb{Z}$-module but it cannot be made into a $\mathbb{Q}$-module (if it could, then $\frac{1}{2} \circ 1 = z$ would be an element of $\mathbb{Z}$ with $z + z = 1$, which is impossible). Although $\mathbb{Z}$ itself cannot be made into a $\mathbb{Q}$-module it is *contained* in a $\mathbb{Q}$-module, namely $\mathbb{Q}$ itself. Put another way, there is an injection (also called an *embedding*) of the $\mathbb{Z}$-module $\mathbb{Z}$ into the $\mathbb{Q}$-module $\mathbb{Q}$ (and similarly the ring R can always be embedded as an R-submodule of the S-module S). This raises the question of whether an arbitrary R-module N can be embedded as an R-submodule of some S-module, or more generally, the question of what R-module homomorphisms exist from N to S-modules. For example, suppose N is a nontrivial *finite* abelian group, say $N = \mathbb{Z}/2\mathbb{Z}$, and consider possible $\mathbb{Z}$-module homomorphisms (i.e., abelian group homomorphisms) of N into some $\mathbb{Q}$-module. A $\mathbb{Q}$-module is just a vector space over $\mathbb{Q}$ and every nonzero element in a vector space over $\mathbb{Q}$ has infinite (additive) order. Since every element of N has finite order, every element of N must map to 0 under such a homomorphism. In other words there are *no* nonzero $\mathbb{Z}$-module homomorphisms from this N to *any* $\mathbb{Q}$-module, much less embeddings of N identifying

N as a submodule of a $\mathbb{Q}$-module. The two $\mathbb{Z}$-modules $\mathbb{Z}$ and $\mathbb{Z}/2\mathbb{Z}$ exhibit extremely different behaviors when we try to "extend scalars" from $\mathbb{Z}$ to $\mathbb{Q}$: the first module maps injectively into some $\mathbb{Q}$-module, the second always maps to 0 in a $\mathbb{Q}$-module.

We now construct for a general R-module N an S-module that is the "best possible" target in which to try to embed N. We shall also see that this module determines *all* of the possible R-module homomorphisms of N into S-modules, in particular determining when N is contained in some S-module (cf. Corollary 9). In the case of $R = \mathbb{Z}$ and $S = \mathbb{Q}$ this construction will give us $\mathbb{Q}$ when applied to the module $N = \mathbb{Z}$, and will give us 0 when applied to the module $N = \mathbb{Z}/2\mathbb{Z}$ (Examples 2 and 3 following Corollary 9).

If the R-module N were already an S-module then of course there is no difficulty in "extending" the scalars from R to S, so we begin the construction by returning to the basic module axioms in order to examine whether we can define "products" of the form sn, for $s \in S$ and $n \in N$. These axioms start with an abelian group N together with a map from $S \times N$ to N, where the image of the pair (s, n) is denoted by sn. It is therefore natural to consider the free $\mathbb{Z}$-module (i.e., , the free abelian group) on the set $S \times N$, i.e., the collection of all finite commuting sums of elements of the form (s_i, n_i) where $s_i \in S$ and $n_i \in N$. This is an abelian group where there are no relations between any distinct pairs (s, n) and (s', n'), i.e., no relations between the "formal products" sn, and in this abelian group the original module N has been thoroughly distinguished from the new "coefficients" from S. To satisfy the relations necessary for an S-module structure imposed in equation (1) and the compatibility relation with the action of R on N in (2′), we must take the quotient of this abelian group by the subgroup H generated by all elements of the form

$$\begin{gathered}(s_1 + s_2, n) - (s_1, n) - (s_2, n), \\ (s, n_1 + n_2) - (s, n_1) - (s, n_2), \text{ and} \\ (sr, n) - (s, rn),\end{gathered} \tag{10.3}$$

for $s, s_1, s_2 \in S$, $n, n_1, n_2 \in N$ and $r \in R$, where rn in the last element refers to the R-module structure already defined on N.

The resulting quotient group is denoted by $S \otimes_R N$ (or just $S \otimes N$ if R is clear from the context) and is called the *tensor product of S and N over R*. If $s \otimes n$ denotes the coset containing (s, n) in $S \otimes_R N$ then by definition of the quotient we have forced the relations

$$\begin{gathered}(s_1 + s_2) \otimes n = s_1 \otimes n + s_2 \otimes n, \\ s \otimes (n_1 + n_2) = s \otimes n_1 + s \otimes n_2, \text{ and} \\ sr \otimes n = s \otimes rn.\end{gathered} \tag{10.4}$$

The elements of $S \otimes_R N$ are called *tensors* and can be written (non-uniquely in general) as finite sums of "simple tensors" of the form $s \otimes n$ with $s \in S$, $n \in N$.

We now show that the tensor product $S \otimes_R N$ is naturally a left S-module under the action defined by

$$s\left(\sum_{\text{finite}} s_i \otimes n_i\right) = \sum_{\text{finite}} (ss_i) \otimes n_i. \tag{10.5}$$

We first check this is well defined, i.e., independent of the representation of the element of $S \otimes_R N$ as a sum of simple tensors. Note first that if s' is any element of S then

$$(s'(s_1 + s_2), n) - (s's_1, n) - (s's_2, n) \quad \big(= (s's_1 + s's_2, n) - (s's_1, n) - (s's_2, n)\big),$$
$$(s's, n_1 + n_2) - (s's, n_1) - (s's, n_2), \text{ and}$$
$$(s'(sr), n) - (s's, rn) \quad \big(= ((s's)r, n) - (s's, rn)\big)$$

each belongs to the set of generators in (3), so in particular each lies in the subgroup H. This shows that multiplying the first entries of the generators in (3) on the left by s' gives another element of H (in fact another generator). Since any element of H is a sum of elements as in (3), it follows that for any element $\sum(s_i, n_i)$ in H also $\sum(s's_i, n_i)$ lies in H. Suppose now that $\sum s_i \otimes n_i = \sum s_i' \otimes n_i'$ are two representations for the same element in $S \otimes_R N$. Then $\sum(s_i, n_i) - \sum(s_i', n_i')$ is an element of H, and by what we have just seen, for any $s \in S$ also $\sum(ss_i, n_i) - \sum(ss_i', n_i')$ is an element of H. But this means that $\sum ss_i \otimes n_i = \sum ss_i' \otimes n_i'$ in $S \otimes_R N$, so the expression in (5) is indeed well defined.

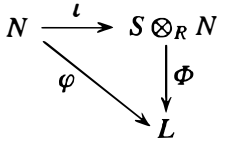
It is now straightforward using the relations in (4) to check that the action defined in (5) makes $S \otimes_R N$ into a left S-module. For example, on the simple tensor $s_i \otimes n_i$,

$$\begin{aligned}
(s + s')(s_i \otimes n_i) &= ((s + s')s_i) \otimes n_i && \text{by definition (5)} \\
&= (ss_i + s's_i) \otimes n_i \\
&= ss_i \otimes n_i + s's_i \otimes n_i && \text{by the first relation in (4)} \\
&= s(s_i \otimes n_i) + s'(s_i \otimes n_i) && \text{by definition (5)}.
\end{aligned}$$

The module $S \otimes_R N$ is called *the (left) S-module obtained by extension of scalars from the (left) R-module N.*

There is a natural map $\iota : N \to S \otimes_R N$ defined by $n \mapsto 1 \otimes n$ (i.e., first map $n \in N$ to the element $(1, n)$ in the free abelian group and then pass to the quotient group). Since $1 \otimes rn = r \otimes n = r(1 \otimes n)$ by (4) and (5), it is easy to check that ι is an R-module homomorphism from N to $S \otimes_R N$. Since we have passed to a quotient group, however, ι is not injective in general. Hence, while there is a natural R-module homomorphism from the original left R-module N to the left S-module $S \otimes_R N$, in general $S \otimes_R N$ need not contain (an isomorphic copy of) N. On the other hand, the relations in equation (3) were the *minimal* relations that we had to impose in order to obtain an S-module, so it is reasonable to expect that the tensor product $S \otimes_R N$ is the "best possible" S-module to serve as target for an R-module homomorphism from N. The next theorem makes this more precise by showing that any other R-module homomorphism from N factors through this one, and is referred to as the *universal property* for the tensor product $S \otimes_R N$. The analogous result for the general tensor product is given in Theorem 10.

Theorem 8. Let R be a subring of S, let N be a left R-module and let $\iota : N \to S \otimes_R N$ be the R-module homomorphism defined by $\iota(n) = 1 \otimes n$. Suppose that L is any left S-module (hence also an R-module) and that $\varphi : N \to L$ is an R-module homomorphism from N to L. Then there is a unique S-module homomorphism $\Phi : S \otimes_R N \to L$ such that φ factors through Φ, i.e., $\varphi = \Phi \circ \iota$ and the diagram



commutes. Conversely, if $\Phi : S \otimes_R N \to L$ is an S-module homomorphism then $\varphi = \Phi \circ \iota$ is an R-module homomorphism from N to L.

Proof: Suppose $\varphi : N \to L$ is an R-module homomorphism to the S-module L. By the universal property of free modules (Theorem 6 in Section 3) there is a $\mathbb{Z}$-module homomorphism from the free $\mathbb{Z}$-module F on the set $S \times N$ to L that sends each generator (s, n) to $s\varphi(n)$. Since φ is an R-module homomorphism, the generators of the subgroup H in equation (3) all map to zero in L. Hence this $\mathbb{Z}$-module homomorphism factors through H, i.e., there is a well defined $\mathbb{Z}$-module homomorphism Φ from $F/H = S \otimes_R N$ to L satisfying $\Phi(s \otimes n) = s\varphi(n)$. Moreover, on simple tensors we have

$$s'\Phi(s \otimes n) = s'(s\varphi(n)) = (s's)\varphi(n) = \Phi((s's) \otimes n) = \Phi(s'(s \otimes n)).$$

for any $s' \in S$. Since Φ is additive it follows that Φ is an S-module homomorphism, which proves the existence statement of the theorem. The module $S \otimes_R N$ is generated as an S-module by elements of the form $1 \otimes n$, so any S-module homomorphism is uniquely determined by its values on these elements. Since $\Phi(1 \otimes n) = \varphi(n)$, it follows that the S-module homomorphism Φ is uniquely determined by φ, which proves the uniqueness statement of the theorem. The converse statement is immediate.

The universal property of $S \otimes_R N$ in Theorem 8 shows that R-module homomorphisms of N into S-modules arise from S-module homomorphisms from $S \otimes_R N$. In particular this determines when it is possible to map N injectively into some S-module:

Corollary 9. Let $\iota : N \to S \otimes_R N$ be the R-module homomorphism in Theorem 8. Then $N/\ker \iota$ is the unique largest quotient of N that can be embedded in any S-module. In particular, N can be embedded as an R-submodule of some left S-module if and only if ι is injective (in which case N is isomorphic to the R-submodule $\iota(N)$ of the S-module $S \otimes_R N$).

Proof: The quotient $N/\ker \iota$ is mapped injectively (by ι) into the S-module $S \otimes_R N$. Suppose now that φ is an R-module homomorphism injecting the quotient $N/\ker \varphi$ of N into an S-module L. Then, by Theorem 8, $\ker \iota$ is mapped to 0 by φ, i.e., $\ker \iota \subseteq \ker \varphi$. Hence $N/\ker \varphi$ is a quotient of $N/\ker \iota$ (namely, the quotient by the submodule $\ker \varphi/\ker \iota$). It follows that $N/\ker \iota$ is the unique largest quotient of N that can be embedded in any S-module. The last statement in the corollary follows immediately.

Examples

(1) For any ring R and any left R-module N we have $R \otimes_R N \cong N$ (so "extending scalars from R to R" does not change the module). This follows by taking φ to be the identity map from N to itself (and $S = R$) in Theorem 8: ι is then an isomorphism with inverse isomorphism given by Φ. In particular, if A is any abelian group (i.e., a $\mathbb{Z}$-module), then $\mathbb{Z} \otimes_{\mathbb{Z}} A = A$.

(2) Let $R = \mathbb{Z}$, $S = \mathbb{Q}$ and let A be a finite abelian group of order n. In this case the $\mathbb{Q}$-module $\mathbb{Q} \otimes_{\mathbb{Z}} A$ obtained by extension of scalars from the $\mathbb{Z}$-module A is 0. To see this, observe first that in any tensor product $1 \otimes 0 = 1 \otimes (0 + 0) = 1 \otimes 0 + 1 \otimes 0$, by the second relation in (4), so

$$1 \otimes 0 = 0.$$

Now, for any simple tensor $q \otimes a$ we can write the rational number q as $(q/n)n$. Then since $na = 0$ in A by Lagrange's Theorem, we have

$$q \otimes a = (\frac{q}{n} \cdot n) \otimes a = \frac{q}{n} \otimes (na) = (q/n) \otimes 0 = (q/n)(1 \otimes 0) = 0.$$

It follows that $\mathbb{Q} \otimes_{\mathbb{Z}} A = 0$. In particular, the map $\iota : A \to S \otimes_R A$ is the zero map. By Theorem 8, we see again that any homomorphism of a finite abelian group into a rational vector space is the zero map. In particular, if A is nontrivial, then the original $\mathbb{Z}$-module A is not contained in the $\mathbb{Q}$-module obtained by extension of scalars.

(3) *Extension of scalars for free modules:* If $N \cong R^n$ is a free module of rank n over R then $S \otimes_R N \cong S^n$ is a free module of rank n over S. We shall prove this shortly (Corollary 18) when we discuss tensor products of direct sums. For example, $\mathbb{Q} \otimes_{\mathbb{Z}} \mathbb{Z}^n \cong \mathbb{Q}^n$. In this case the module obtained by extension of scalars contains (an isomorphic copy of) the original R-module N. For example, $\mathbb{Q} \otimes_{\mathbb{Z}} \mathbb{Z}^n \cong \mathbb{Q}^n$ and $\mathbb{Z}^n$ is a subgroup of the abelian group $\mathbb{Q}^n$.

(4) *Extension of scalars for vector spaces:* As a special case of the previous example, let F be a subfield of the field K and let V be an n-dimensional vector space over F (i.e., $V \cong F^n$). Then $K \otimes_F V \cong K^n$ is a vector space over the larger field K of the same dimension, and the original vector space V is contained in $K \otimes_F V$ as an F-vector subspace.

(5) *Induced modules for finite groups:* Let R be a commutative ring with 1, let G be a finite group and let H be a subgroup of G. As in Section 7.2 we may form the group ring RG and its subring RH. For any RH-module N define the *induced module* $RG \otimes_{RH} N$. In this way we obtain an RG-module for each RH-module N. We shall study properties of induced modules and some of their important applications to group theory in Chapters 17 and 19.

The general tensor product construction follows along the same lines as the extension of scalars above, but before describing it we make two observations from this special case. The first is that the construction of $S \otimes_R N$ as an *abelian group* involved only the elements in equation (3), which in turn only required S to be a *right* R-module and N to be a *left* R-module. In a similar way we shall construct an *abelian group* $M \otimes_R N$ for *any right* R-module M and *any left* R-module N. The second observation is that the *S-module* structure on $S \otimes_R N$ defined by equation (5) required only a *left* S-module structure on S together with a "compatibility relation"

$$s'(sr) = (s's)r \qquad \text{for } s, s' \in S, r \in R,$$

between this left S-module structure and the right R-module structure on S (this was needed in order to deduce that (5) was well defined). We first consider the general construction of $M \otimes_R N$ as an abelian group, after which we shall return to the question of when this abelian group can be given a module structure.

Suppose then that N is a left R-module and that M is a right R-module. The quotient of the free $\mathbb{Z}$-module on the set $M \times N$ by the subgroup generated by all elements of the form

$$\begin{gathered}(m_1 + m_2, n) - (m_1, n) - (m_2, n), \\ (m, n_1 + n_2) - (m, n_1) - (m, n_2), \text{ and} \\ (mr, n) - (m, rn),\end{gathered} \tag{10.6}$$

for $m, m_1, m_2 \in M$, $n, n_1, n_2 \in N$ and $r \in R$ is an abelian group, denoted by $M \otimes_R N$, or simply $M \otimes N$ if the ring R is clear from the context, and is called the *tensor product of M and N over R*. The elements of $M \otimes_R N$ are called *tensors*, and the coset, $m \otimes n$, of (m, n) in $M \otimes_R N$ is called a simple tensor. We have the relations

$$\begin{gathered}(m_1 + m_2) \otimes n = m_1 \otimes n + m_2 \otimes n, \\ m \otimes (n_1 + n_2) = m \otimes n_1 + m \otimes n_2, \text{ and} \\ mr \otimes n = m \otimes rn.\end{gathered} \tag{10.7}$$

Every tensor can be written (non-uniquely in general) as a finite sum of simple tensors.

Remark: We emphasize that care must be taken when working with tensors, since each $m \otimes n$ represents a *coset* in some quotient group, and so we may have $m \otimes n = m' \otimes n'$ where $m \neq m'$ or $n \neq n'$. More generally, an element of $M \otimes N$ may be expressible in many different ways as a sum of simple tensors. In particular, care must be taken when defining maps from $M \otimes_R N$ to another group or module, since a map from $M \otimes N$ which is described on the generators $m \otimes n$ in terms of m and n is not well defined unless it is shown to be independent of the particular choice of $m \otimes n$ as a coset representative.

Another point where care must be exercised is in reference to the element $m \otimes n$ when the modules M and N or the ring R are not clear from the context. The first two examples of extension of scalars give an instance where M is a submodule of a larger module M', and for some $m \in M$ and $n \in N$ we have $m \otimes n = 0$ in $M' \otimes_R N$ but $m \otimes n$ is *nonzero* in $M \otimes_R N$. This is possible because the symbol "$m \otimes n$" represents different cosets, hence possibly different elements, in the two tensor products. In particular, these two examples show that $M \otimes_R N$ need not be a subgroup of $M' \otimes_R N$ even when M is a submodule of M' (cf. also Exercise 2).

Mapping $M \times N$ to the free $\mathbb{Z}$-module on $M \times N$ and then passing to the quotient defines a map $\iota : M \times N \to M \otimes_R N$ with $\iota(m, n) = m \otimes n$. This map is in general not a group homomorphism, but it is additive in both m and n separately and satisfies $\iota(mr, n) = mr \otimes n = m \otimes rn = \iota(m, rn)$. Such maps are given a name:

Definition. Let M be a right R-module, let N be a left R-module and let L be an abelian group (written additively). A map $\varphi : M \times N \to L$ is called *R-balanced* or *middle linear with respect to R* if

$$\varphi(m_1 + m_2, n) = \varphi(m_1, n) + \varphi(m_2, n)$$
$$\varphi(m, n_1 + n_2) = \varphi(m, n_1) + \varphi(m, n_2)$$
$$\varphi(m, rn) = \varphi(mr, n)$$

for all $m, m_1, m_2 \in M$, $n, n_1, n_2 \in N$, and $r \in R$.

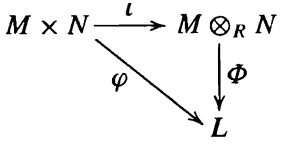
With this terminology, it follows immediately from the relations in (7) that the map $\iota : M \times N \to M \otimes_R N$ is R-balanced. The next theorem proves the extremely useful *universal property of the tensor product* with respect to balanced maps.

Theorem 10. Suppose R is a ring with 1, M is a right R-module, and N is a left R-module. Let $M \otimes_R N$ be the tensor product of M and N over R and let $\iota : M \times N \to M \otimes_R N$ be the R-balanced map defined above.

(1) If $\Phi : M \otimes_R N \to L$ is any group homomorphism from $M \otimes_R N$ to an abelian group L then the composite map $\varphi = \Phi \circ \iota$ is an R-balanced map from $M \times N$ to L.

(2) Conversely, suppose L is an abelian group and $\varphi : M \times N \to L$ is any R-balanced map. Then there is a unique group homomorphism $\Phi : M \otimes_R N \to L$ such that φ factors through ι, i.e., $\varphi = \Phi \circ \iota$ as in (1).

Equivalently, the correspondence $\varphi \leftrightarrow \Phi$ in the commutative diagram



establishes a bijection

$$\left\{ \begin{array}{c} R\text{-balanced maps} \\ \varphi : M \times N \to L \end{array} \right\} \longleftrightarrow \left\{ \begin{array}{c} \text{group homomorphisms} \\ \Phi : M \otimes_R N \to L \end{array} \right\}.$$

Proof: The proof of (1) is immediate from the properties of ι above. For (2), the map φ defines a unique $\mathbb{Z}$-module homomorphism $\tilde{\varphi}$ from the free group on $M \times N$ to L (Theorem 6 in Section 3) such that $\tilde{\varphi}(m, n) = \varphi(m, n) \in L$. Since φ is R-balanced, $\tilde{\varphi}$ maps each of the elements in equation (6) to 0; for example

$$\tilde{\varphi}\left((mr, n) - (m, rn)\right) = \varphi(mr, n) - \varphi(m, rn) = 0.$$

It follows that the kernel of $\tilde{\varphi}$ contains the subgroup generated by these elements, hence $\tilde{\varphi}$ induces a homomorphism Φ on the quotient group $M \otimes_R N$ to L. By definition we then have

$$\Phi(m \otimes n) = \tilde{\varphi}(m, n) = \varphi(m, n),$$

i.e., $\varphi = \Phi \circ \iota$. The homomorphism Φ is uniquely determined by this equation since the elements $m \otimes n$ generate $M \otimes_R N$ as an abelian group. This completes the proof.

Theorem 10 is extremely useful in defining homomorphisms on $M \otimes_R N$ since it replaces the often tedious check that maps defined on simple tensors $m \otimes n$ are well defined with a check that a related map defined on ordered pairs (m, n) is balanced.

The first consequence of the universal property in Theorem 10 is a characterization of the tensor product $M \otimes_R N$ as an abelian group:

Corollary 11. Suppose D is an abelian group and $\iota' : M \times N \to D$ is an R-balanced map such that

(i) the image of ι' generates D as an abelian group, and

(ii) every R-balanced map defined on $M \times N$ factors through ι' as in Theorem 10.

Then there is an isomorphism $f : M \otimes_R N \cong D$ of abelian groups with $\iota' = f \circ \iota$.

Proof: Since $\iota' : M \times N \to D$ is a balanced map, the universal property in (2) of Theorem 10 implies there is a (unique) homomorphism $f : M \otimes_R N \to D$ with $\iota' = f \circ \iota$. In particular $\iota'(m, n) = f(m \otimes n)$ for every $m \in M$, $n \in N$. By the first assumption on ι', these elements generate D as an abelian group, so f is a surjective map. Now, the balanced map $\iota : M \times N \to M \otimes_R N$ together with the second assumption on ι' implies there is a (unique) homomorphism $g : D \to M \otimes_R N$ with $\iota = g \circ \iota'$. Then $m \otimes n = (g \circ f)(m \otimes n)$. Since the simple tensors $m \otimes n$ generate $M \otimes_R N$, it follows that $g \circ f$ is the identity map on $M \otimes_R N$ and so f is injective, hence an isomorphism. This establishes the corollary.

We now return to the question of giving the abelian group $M \otimes_R N$ a *module* structure. As we observed in the special case of extending scalars from R to S for the R-module N, the S-module structure on $S \otimes_R N$ required only a left S-module structure on S together with the compatibility relation $s'(sr) = (s's)r$ for $s, s' \in S$ and $r \in R$. In this special case this relation was simply a consequence of the associative law in the ring S. To obtain an S-module structure on $M \otimes_R N$ more generally we impose a similar structure on M:

Definition. Let R and S be any rings with 1. An abelian group M is called an (S, R)-*bimodule* if M is a left S-module, a right R-module, and $s(mr) = (sm)r$ for all $s \in S$, $r \in R$ and $m \in M$.

Examples

(1) Any ring S is an (S, R)-bimodule for any subring R with $1_R = 1_S$ by the associativity of the multiplication in S. More generally, if $f : R \to S$ is any ring homomorphism with $f(1_R) = 1_S$ then S can be considered as a right R-module with the action $s \cdot r = sf(r)$, and with respect to this action S becomes an (S, R)-bimodule.

(2) Let I be an ideal (two-sided) in the ring R. Then the quotient ring R/I is an $(R/I, R)$-bimodule. This is easy to see directly and is also a special case of the previous example (with respect to the canonical projection homomorphism $R \to R/I$).

(3) Suppose that R is a commutative ring. Then a left (respectively, right) R-module M can always be given the structure of a right (respectively, left) R-module by defining $mr = rm$ (respectively, $rm = mr$), for all $m \in M$ and $r \in R$, and this makes M into

an (R, R)-bimodule. Hence every module (right or left) over a commutative ring R has at least one natural (R, R)-bimodule structure.

(4) Suppose that M is a left S-module and R is a subring contained in the *center* of S (for example, if S is commutative). Then in particular R is commutative so M can be given a right R-module structure as in the previous example. Then for any $s \in S$, $r \in R$ and $m \in M$ by definition of the right action of R we have

$$(sm)r = r(sm) = (rs)m = (sr)m = s(rm) = s(mr)$$

(note that we have used the fact that r commutes with s in the middle equality). Hence M is an (S, R)-bimodule with respect to this definition of the right action of R.

Since the situation in Example 3 occurs so frequently, we give this bimodule structure a name:

Definition. Suppose M is a left (or right) R-module over the commutative ring R. Then the (R, R)-bimodule structure on M defined by letting the left and right R-actions coincide, i.e., $mr = rm$ for all $m \in M$ and $r \in R$, will be called the *standard* R-module structure on M.

Suppose now that N is a left R-module and M is an (S, R)-bimodule. Then just as in the example of extension of scalars the (S, R)-bimodule structure on M implies that

$$s\left(\sum_{\text{finite}} m_i \otimes n_i\right) = \sum_{\text{finite}} (sm_i) \otimes n_i \tag{10.8}$$

gives a well defined action of S under which $M \otimes_R N$ *is a left* S*-module.* Note that Theorem 10 may be used to give an alternate proof that (8) is well defined, replacing the direct calculations on the relations defining the tensor product with the easier check that a map is R-balanced, as follows. It is very easy to see that for each fixed $s \in S$ the map $(m, n) \mapsto sm \otimes n$ is an R-balanced map from $M \times N$ to $M \otimes_R N$. By Theorem 10 there is a well defined group homomorphism λ_s from $M \otimes_R N$ to itself such that $\lambda_s(m \otimes n) = sm \otimes n$. Since the right side of (8) is then $\lambda_s(\sum m_i \otimes n_i)$, the fact that λ_s is well defined shows that this expression is indeed independent of the representation of the tensor $\sum m_i \otimes n_i$ as a sum of simple tensors. Because λ_s is additive, equation (8) holds.

By a completely parallel argument, if M is a right R-module and N is an (R, S)-bimodule then the tensor product $M \otimes_R N$ has the structure of a *right* S-module, where $(\sum m_i \otimes n_i)\, s = \sum m_i \otimes (n_i s)$.

Before giving some more examples of tensor products it is worthwhile to highlight one frequently encountered special case of the previous discussion, namely the case when M and N are two left modules over a *commutative* ring R and $S = R$ (in some works on tensor products this is the only case considered). Then the standard R-module structure on M defined previously gives M the structure of an (R, R)-bimodule, so in this case the tensor product $M \otimes_R N$ always has the structure of a left R-module.

The corresponding map $\iota : M \times N \to M \otimes_R N$ maps $M \times N$ into an R-module and is additive in each factor. Since $r(m \otimes n) = rm \otimes n = mr \otimes n = m \otimes rn$ it also satisfies

$$r\iota(m, n) = \iota(rm, n) = \iota(m, rn).$$

Such maps are given a name:

Definition. Let R be a commutative ring with 1 and let M, N, and L be left R-modules. The map $\varphi : M \times N \to L$ is called *R-bilinear* if it is R-linear in each factor, i.e., if

$$\varphi(r_1m_1 + r_2m_2, n) = r_1\varphi(m_1, n) + r_2\varphi(m_2, n), \quad \text{and}$$
$$\varphi(m, r_1n_1 + r_2n_2) = r_1\varphi(m, n_1) + r_2\varphi(m, n_2)$$

for all $m, m_1, m_2 \in M$, $n, n_1, n_2 \in N$ and $r_1, r_2 \in R$.

With this terminology Theorem 10 gives

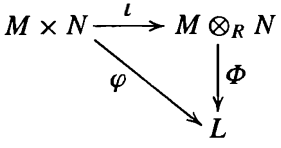
Corollary 12. Suppose R is a commutative ring. Let M and N be two left R-modules and let $M \otimes_R N$ be the tensor product of M and N over R, where M is given the standard R-module structure. Then $M \otimes_R N$ is a left R-module with

$$r(m \otimes n) = (rm) \otimes n = (mr) \otimes n = m \otimes (rn),$$

and the map $\iota : M \times N \to M \otimes_R N$ with $\iota(m, n) = m \otimes n$ is an R-bilinear map. If L is any left R-module then there is a bijection

$$\left\{ \begin{array}{c} R\text{-bilinear maps} \\ \varphi : M \times N \to L \end{array} \right\} \longleftrightarrow \left\{ \begin{array}{c} R\text{-module homomorphisms} \\ \Phi : M \otimes_R N \to L \end{array} \right\}$$

where the correspondence between φ and Φ is given by the commutative diagram



Proof: We have shown $M \otimes_R N$ is an R-module and that ι is bilinear. It remains only to check that in the bijective correspondence in Theorem 10 the bilinear maps correspond with the R-module homomorphisms. If $\varphi : M \times N \to L$ is bilinear then it is an R-balanced map, so the corresponding $\Phi : M \otimes_R N$ is a group homomorphism. Moreover, on simple tensors $\Phi((rm) \otimes n) = \varphi(rm, n) = r\varphi(m, n) = r\Phi(m \otimes n)$, where the middle equality holds because φ is R-linear in the first variable. Since Φ is additive this extends to sums of simple tensors to show Φ is an R-module homomorphism. Conversely, if Φ is an R-module homomorphism it is an exercise to see that the corresponding balanced map φ is bilinear.

Examples

(1) In any tensor product $M \otimes_R N$ we have $m \otimes 0 = m \otimes (0 + 0) = (m \otimes 0) + (m \otimes 0)$, so $m \otimes 0 = 0$. Likewise $0 \otimes n = 0$.

(2) We have $\mathbb{Z}/2\mathbb{Z} \otimes_{\mathbb{Z}} \mathbb{Z}/3\mathbb{Z} = 0$, since $3a = a$ for $a \in \mathbb{Z}/2\mathbb{Z}$ so that

$$a \otimes b = 3a \otimes b = a \otimes 3b = a \otimes 0 = 0$$

and every simple tensor is reduced to 0. In particular $1 \otimes 1 = 0$. It follows that there are no nonzero balanced (or bilinear) maps from $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/3\mathbb{Z}$ to any abelian group.

On the other hand, consider the tensor product $\mathbb{Z}/2\mathbb{Z} \otimes_{\mathbb{Z}} \mathbb{Z}/2\mathbb{Z}$, which is generated as an abelian group by the elements $0 \otimes 0 = 1 \otimes 0 = 0 \otimes 1 = 0$ and $1 \otimes 1$. In this case $1 \otimes 1 \neq 0$ since, for example, the map $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z} \to \mathbb{Z}/2\mathbb{Z}$ defined by $(a, b) \mapsto ab$ is clearly nonzero and linear in both a and b. Since $2(1 \otimes 1) = 2 \otimes 1 = 0 \otimes 1 = 0$, the element $1 \otimes 1$ is of order 2. Hence $\mathbb{Z}/2\mathbb{Z} \otimes_{\mathbb{Z}} \mathbb{Z}/2\mathbb{Z} \cong \mathbb{Z}/2\mathbb{Z}$.

(3) In general,

$$\mathbb{Z}/m\mathbb{Z} \otimes_{\mathbb{Z}} \mathbb{Z}/n\mathbb{Z} \cong \mathbb{Z}/d\mathbb{Z},$$

where d is the g.c.d. of the integers m and n. To see this, observe first that

$$a \otimes b = a \otimes (b \cdot 1) = (ab) \otimes 1 = ab(1 \otimes 1),$$

from which it follows that $\mathbb{Z}/m\mathbb{Z} \otimes_{\mathbb{Z}} \mathbb{Z}/n\mathbb{Z}$ is a cyclic group with $1 \otimes 1$ as generator. Since $m(1 \otimes 1) = m \otimes 1 = 0 \otimes 1 = 0$ and similarly $n(1 \otimes 1) = 1 \otimes n = 0$, we have $d(1 \otimes 1) = 0$, so the cyclic group has order dividing d. The map $\varphi : \mathbb{Z}/m\mathbb{Z} \times \mathbb{Z}/n\mathbb{Z} \to \mathbb{Z}/d\mathbb{Z}$ defined by $\varphi(a \bmod m, b \bmod n) = ab \bmod d$ is well defined since d divides both m and n. It is clearly $\mathbb{Z}$-bilinear. The induced map $\Phi : \mathbb{Z}/m\mathbb{Z} \otimes_{\mathbb{Z}} \mathbb{Z}/n\mathbb{Z} \to \mathbb{Z}/d\mathbb{Z}$ from Corollary 12 maps $1 \otimes 1$ to the element $1 \in \mathbb{Z}/d\mathbb{Z}$, which is an element of order d. In particular $\mathbb{Z}/m\mathbb{Z} \otimes_{\mathbb{Z}} \mathbb{Z}/n\mathbb{Z}$ has order at least d. Hence $1 \otimes 1$ is an element of order d and Φ gives an isomorphism $\mathbb{Z}/m\mathbb{Z} \otimes_{\mathbb{Z}} \mathbb{Z}/n\mathbb{Z} \cong \mathbb{Z}/d\mathbb{Z}$.

(4) In $\mathbb{Q}/\mathbb{Z} \otimes_{\mathbb{Z}} \mathbb{Q}/\mathbb{Z}$ a simple tensor has the form $(a/b \bmod \mathbb{Z}) \otimes (c/d \bmod \mathbb{Z})$ for some rational numbers a/b and c/d. Then

$$\begin{aligned}(\tfrac{a}{b} \bmod \mathbb{Z}) \otimes (\tfrac{c}{d} \bmod \mathbb{Z}) &= d(\tfrac{a}{bd} \bmod \mathbb{Z}) \otimes (\tfrac{c}{d} \bmod \mathbb{Z}) \\ &= (\tfrac{a}{bd} \bmod \mathbb{Z}) \otimes d(\tfrac{c}{d} \bmod \mathbb{Z}) = (\tfrac{a}{bd} \bmod \mathbb{Z}) \otimes 0 = 0\end{aligned}$$

and so

$$\mathbb{Q}/\mathbb{Z} \otimes_{\mathbb{Z}} \mathbb{Q}/\mathbb{Z} = 0.$$

In a similar way, $A \otimes_{\mathbb{Z}} B = 0$ for any *divisible* abelian group A and *torsion* abelian group B (an abelian group in which every element has finite order). For example

$$\mathbb{Q} \otimes_{\mathbb{Z}} \mathbb{Q}/\mathbb{Z} = 0.$$

(5) The structure of a tensor product can vary considerably depending on the ring over which the tensors are taken. For example $\mathbb{Q} \otimes_{\mathbb{Q}} \mathbb{Q}$ and $\mathbb{Q} \otimes_{\mathbb{Z}} \mathbb{Q}$ are isomorphic as left $\mathbb{Q}$-modules (both are one dimensional vector spaces over $\mathbb{Q}$) — cf. the exercises. On the other hand we shall see at the end of this section that $\mathbb{C} \otimes_{\mathbb{C}} \mathbb{C}$ and $\mathbb{C} \otimes_{\mathbb{R}} \mathbb{C}$ are not isomorphic $\mathbb{C}$-modules (the former is a 1-dimensional vector space over $\mathbb{C}$ and the latter is 2-dimensional over $\mathbb{C}$).

(6) *General extension of scalars or change of base:* Let $f : R \to S$ be a ring homomorphism with $f(1_R) = 1_S$. Then $s \cdot r = sf(r)$ gives S the structure of a right R-module with respect to which S is an (S, R)-bimodule. Then for any left R-module N, the resulting tensor product $S \otimes_R N$ is a left S-module obtained by *changing the base* from R to S. This gives a slight generalization of the notion of extension of scalars (where R was a subring of S).

(7) Let $f : R \to S$ be a ring homomorphism as in the preceding example. Then we have $S \otimes_R R \cong S$ as left S-modules, as follows. The map $\varphi : S \times R \to S$ defined by $(s, r) \mapsto sr$ (where $sr = sf(r)$ by definition of the right R-action on S), is an R-balanced map, as is easily checked. For example,

$$\varphi(s_1 + s_2, r) = (s_1 + s_2)r = s_1r + s_2r = \varphi(s_1, r) + \varphi(s_2, r)$$

and

$$\varphi(sr, r') = (sr)r' = s(rr') = \varphi(s, rr').$$

By Theorem 10 we have an associated group homomorphism $\Phi : S \otimes_R R \to S$ with $\Phi(s \otimes r) = sr$. Since $\Phi(s'(s \otimes r)) = \Phi(s's \otimes r) = s'sr = s'\Phi(s \otimes r)$, it follows that Φ is also an S-module homomorphism. The map $\Phi' : S \to S \otimes_R R$ with $s \mapsto s \otimes 1$ is an S-module homomorphism that is inverse to Φ because $\Phi \circ \Phi'(s) = \Phi(s \otimes 1) = s$ gives $\Phi\Phi' = 1$, and

$$\Phi' \circ \Phi(s \otimes r) = \Phi'(sr) = sr \otimes 1 = s \otimes r$$

shows that $\Phi'\Phi$ is the identity on simple tensors, hence $\Phi'\Phi = 1$.

(8) Let R be a ring (not necessarily commutative), let I be a two sided ideal in R, and let N be a left R-module. Then as previously mentioned, R/I is an $(R/I, R)$-bimodule, so the tensor product $R/I \otimes_R N$ is a left R/I-module. This is an example of "extension of scalars" with respect to the natural projection homomorphism $R \to R/I$.

Define

$$IN = \left\{ \sum_{\text{finite}} a_i n_i \mid a_i \in I, n_i \in N \right\},$$

which is easily seen to be a left R-submodule of N (cf. Exercise 5, Section 1). Then

$$(R/I) \otimes_R N \cong N/IN,$$

as left R-modules, as follows. The tensor product is generated as an abelian group by the simple tensors $(r \bmod I) \otimes n = r(1 \otimes n)$ for $r \in R$ and $n \in N$ (viewing the R/I-module tensor product as an R-module on which I acts trivially). Hence the elements $1 \otimes n$ generate $(R/I) \otimes_R N$ as an R/I-module. The map $N \to (R/I) \otimes_R N$ defined by $n \mapsto 1 \otimes n$ is a left R-module homomorphism and, by the previous observation, is surjective. Under this map $a_i n_i$ with $a_i \in I$ and $n_i \in N$ maps to $1 \otimes a_i n_i = a_i \otimes n_i = 0$, and so IN is contained in the kernel. This induces a surjective R-module homomorphism $f : N/IN \to (R/I) \otimes_R N$ with $f(n \bmod I) = 1 \otimes n$. We show f is an isomorphism by exhibiting its inverse. The map $(R/I) \times N \to N/IN$ defined by mapping $(r \bmod I, n)$ to $(rn \bmod IN)$ is well defined and easily checked to be R-balanced. It follows by Theorem 10 that there is an associated group homomorphism $g : (R/I) \otimes N \to N/IN$ with $g((r \bmod I) \otimes n) = rn \bmod IN$. As usual, $fg = 1$ and $gf = 1$, so f is a bijection and $(R/I) \otimes_R N \cong N/IN$, as claimed.

As an example, let $R = \mathbb{Z}$ with ideal $I = m\mathbb{Z}$ and let N be the $\mathbb{Z}$-module $\mathbb{Z}/n\mathbb{Z}$. Then $IN = m(\mathbb{Z}/n\mathbb{Z}) = (m\mathbb{Z} + n\mathbb{Z})/n\mathbb{Z} = d\mathbb{Z}/n\mathbb{Z}$ where d is the g.c.d. of m and n. Then $N/IN \cong \mathbb{Z}/d\mathbb{Z}$ and we recover the isomorphism $\mathbb{Z}/m\mathbb{Z} \otimes_{\mathbb{Z}} \mathbb{Z}/n\mathbb{Z} \cong \mathbb{Z}/d\mathbb{Z}$ of Example 3 above.

We now establish some of the basic properties of tensor products. Note the frequent application of Theorem 10 to establish the existence of homomorphisms.

Theorem 13. *(The "Tensor Product" of Two Homomorphisms)* Let M, M' be right R-modules, let N, N' be left R-modules, and suppose $\varphi : M \to M'$ and $\psi : N \to N'$ are R-module homomorphisms.

(1) There is a unique group homomorphism, denoted by $\varphi \otimes \psi$, mapping $M \otimes_R N$ into $M' \otimes_R N'$ such that $(\varphi \otimes \psi)(m \otimes n) = \varphi(m) \otimes \psi(n)$ for all $m \in M$ and $n \in N$.

(2) If M, M' are also (S, R)-bimodules for some ring S and φ is also an S-module homomorphism, then $\varphi \otimes \psi$ is a homomorphism of left S-modules. In particular, if R is commutative then $\varphi \otimes \psi$ is always an R-module homomorphism for the standard R-module structures.

(3) If $\lambda : M' \to M''$ and $\mu : N' \to N''$ are R-module homomorphisms then $(\lambda \otimes \mu) \circ (\varphi \otimes \psi) = (\lambda \circ \varphi) \otimes (\mu \circ \psi)$.

Proof: The map $(m, n) \mapsto \varphi(m) \otimes \psi(n)$ from $M \times N$ to $M' \otimes_R N'$ is clearly R-balanced, so (1) follows immediately from Theorem 10.

In (2) the definition of the (left) action of S on M together with the assumption that φ is an S-module homomorphism imply that on simple tensors

$$(\varphi \otimes \psi)(s(m \otimes n)) = (\varphi \otimes \psi)(sm \otimes n) = \varphi(sm) \otimes \psi(n) = s\varphi(m) \otimes \psi(n).$$

Since $\varphi \otimes \psi$ is additive, this extends to sums of simple tensors to show that $\varphi \otimes \psi$ is an S-module homomorphism. This gives (2).

The uniqueness condition in Theorem 10 implies (3), which completes the proof.

The next result shows that we may write $M \otimes N \otimes L$, or more generally, an n-fold tensor product $M_1 \otimes M_2 \otimes \cdots \otimes M_n$, unambiguously whenever it is defined.

Theorem 14. *(Associativity of the Tensor Product)* Suppose M is a right R-module, N is an (R, T)-bimodule, and L is a left T-module. Then there is a unique isomorphism

$$(M \otimes_R N) \otimes_T L \cong M \otimes_R (N \otimes_T L)$$

of abelian groups such that $(m \otimes n) \otimes l \mapsto m \otimes (n \otimes l)$. If M is an (S, R)-bimodule, then this is an isomorphism of S-modules.

Proof: Note first that the (R, T)-bimodule structure on N makes $M \otimes_R N$ into a right T-module and $N \otimes_T L$ into a left R-module, so both sides of the isomorphism are well defined. For each fixed $l \in L$, the mapping $(m, n) \mapsto m \otimes (n \otimes l)$ is R-balanced, so by Theorem 10 there is a homomorphism $M \otimes_R N \to M \otimes_R (N \otimes_T L)$ with $m \otimes n \mapsto m \otimes (n \otimes l)$. This shows that the map from $(M \otimes_R N) \times L$ to $M \otimes_R (N \otimes_T L)$ given by $(m \otimes n, l) \mapsto m \otimes (n \otimes l)$ is well defined. Since it is easily seen to be T-balanced, another application of Theorem 10 implies that it induces a homomorphism $(M \otimes_R N) \otimes_T L \to M \otimes_R (N \otimes_T L)$ such that $(m \otimes n) \otimes l \mapsto m \otimes (n \otimes l)$. In a similar way we can construct a homomorphism in the opposite direction that is inverse to this one. This proves the group isomorphism.

Assume in addition M is an (S, R)-bimodule. Then for $s \in S$ and $t \in T$ we have

$$s((m \otimes n)t) = s(m \otimes nt) = sm \otimes nt = (sm \otimes n)t = (s(m \otimes n))t$$

so that $M \otimes_R N$ is an (S, T)-bimodule. Hence $(M \otimes_R N) \otimes_T L$ is a left S-module. Since $N \otimes_T L$ is a left R-module, also $M \otimes_R (N \otimes_T L)$ is a left S-module. The group isomorphism just established is easily seen to be a homomorphism of left S-modules by the same arguments used in previous proofs: it is additive and is S-linear on simple tensors since $s((m \otimes n) \otimes l) = s(m \otimes n) \otimes l = (sm \otimes n) \otimes l$ maps to the element $sm \otimes (n \otimes l) = s(m \otimes (n \otimes l))$. The proof is complete.

Corollary 15. Suppose R is commutative and M, N, and L are left R-modules. Then

$$(M \otimes N) \otimes L \cong M \otimes (N \otimes L)$$

as R-modules for the standard R-module structures on M, N and L.

There is a natural extension of the notion of a bilinear map:

Definition. Let R be a commutative ring with 1 and let $M_1, M_2, \ldots, M_n$ and L be R-modules with the standard R-module structures. A map $\varphi : M_1 \times \cdots \times M_n \to L$ is called *n-multilinear over R* (or simply *multilinear* if n and R are clear from the context) if it is an R-module homomorphism in each component when the other component entries are kept constant, i.e., for each i

$$\varphi(m_1, \ldots, m_{i-1}, rm_i + r'm_i', m_{i+1}, \ldots, m_n)$$
$$= r\varphi(m_1, \ldots, m_i, \ldots, m_n) + r'\varphi(m_1, \ldots, m_i', \ldots, m_n)$$

for all $m_i, m_i' \in M_i$ and $r, r' \in R$. When $n = 2$ (respectively, 3) one says φ is *bilinear* (respectively *trilinear*) rather than 2-multilinear (or 3-multilinear).

One may construct the n-fold tensor product $M_1 \otimes M_2 \otimes \cdots \otimes M_n$ from first principles and prove its analogous universal property with respect to multilinear maps from $M_1 \times \cdots \times M_n$ to L. By the previous theorem and corollary, however, an n-fold tensor product may be obtained unambiguously by iterating the tensor product of pairs of modules since any bracketing of $M_1 \otimes \cdots \otimes M_n$ into tensor products of pairs gives an isomorphic R-module. The universal property of the tensor product of a pair of modules in Theorem 10 and Corollary 12 then implies that multilinear maps factor uniquely through the R-module $M_1 \otimes \cdots \otimes M_n$, i.e., this tensor product is the universal object with respect to multilinear functions:

Corollary 16. Let R be a commutative ring and let $M_1, \ldots, M_n, L$ be R-modules. Let $M_1 \otimes M_2 \otimes \cdots \otimes M_n$ denote any bracketing of the tensor product of these modules and let

$$\iota : M_1 \times \cdots \times M_n \to M_1 \otimes \cdots \otimes M_n$$

be the map defined by $\iota(m_1, \ldots, m_n) = m_1 \otimes \cdots \otimes m_n$. Then

(1) for every R-module homomorphism $\Phi : M_1 \otimes \cdots \otimes M_n \to L$ the map $\varphi = \Phi \circ \iota$ is n-multilinear from $M_1 \times \cdots \times M_n$ to L, and

(2) if $\varphi : M_1 \times \cdots \times M_n \to L$ is an n-multilinear map then there is a unique R-module homomorphism $\Phi : M_1 \otimes \cdots \otimes M_n \to L$ such that $\varphi = \Phi \circ \iota$.

Hence there is a bijection

$$\left\{ \begin{array}{c} n\text{-multilinear maps} \\ \varphi : M_1 \times \cdots \times M_n \to L \end{array} \right\} \longleftrightarrow \left\{ \begin{array}{c} R\text{-module homomorphisms} \\ \Phi : M_1 \otimes \cdots \otimes M_n \to L \end{array} \right\}$$

with respect to which the following diagram commutes:

$$\begin{array}{ccc} M \times \cdots \times M_n & \xrightarrow{\iota} & M \otimes \cdots \otimes M_n \\ & \searrow^{\varphi} & \downarrow \Phi \\ & & L \end{array}$$

We have already seen examples where $M_1 \otimes_R N$ is not contained in $M \otimes_R N$ even when M_1 is an R-submodule of M. The next result shows in particular that (an isomorphic copy of) $M_1 \otimes_R N$ *is* contained in $M \otimes_R N$ if M_1 is an R-module *direct summand* of M.

Theorem 17. *(Tensor Products of Direct Sums)* Let M, M' be right R-modules and let N, N' be left R-modules. Then there are unique group isomorphisms

$$(M \oplus M') \otimes_R N \cong (M \otimes_R N) \oplus (M' \otimes_R N)$$
$$M \otimes_R (N \oplus N') \cong (M \otimes_R N) \oplus (M \otimes_R N')$$

such that $(m, m') \otimes n \mapsto (m \otimes n, m' \otimes n)$ and $m \otimes (n, n') \mapsto (m \otimes n, m \otimes n')$ respectively. If M, M' are also (S, R)-bimodules, then these are isomorphisms of left S-modules. In particular, if R is commutative, these are isomorphisms of R-modules.

Proof: The map $(M \oplus M') \times N \to (M \otimes_R N) \oplus (M' \otimes_R N)$ defined by $((m, m'), n) \mapsto (m \otimes n, m' \otimes n)$ is well defined since m and m' in $M \oplus M'$ are uniquely defined in the direct sum. The map is clearly R-balanced, so induces a homomorphism f from $(M \oplus M') \otimes N$ to $(M \otimes_R N) \oplus (M' \otimes_R N)$ with

$$f((m, m') \otimes n) = (m \otimes n, m' \otimes n).$$

In the other direction, the R-balanced maps $M \times N \to (M \oplus M') \otimes_R N$ and $M' \times N \to (M \oplus M') \otimes_R N$ given by $(m, n) \mapsto (m, 0) \otimes n$ and $(m', n) \mapsto (0, m') \otimes n$, respectively, define homomorphisms from $M \otimes_R N$ and $M' \otimes_R N$ to $(M \oplus M') \otimes_R N$. These in turn give a homomorphism g from the direct sum $(M \otimes_R N) \oplus (M' \otimes_R N)$ to $(M \oplus M') \otimes_R N$ with

$$g((m \otimes n_1, m' \otimes n_2)) = (m, 0) \otimes n_1 + (0, m') \otimes n_2.$$

An easy check shows that f and g are inverse homomorphisms and are S-module isomorphisms when M and M' are (S, R)-bimodules. This completes the proof.

The previous theorem clearly extends by induction to any finite direct sum of R-modules. The corresponding result is also true for arbitrary direct sums. For example

$$M \otimes (\oplus_{i \in I} N_i) \cong \oplus_{i \in I} (M \otimes N_i),$$

where I is any index set (cf. the exercises). This result is referred to by saying that *tensor products commute with direct sums.*

Corollary 18. *(Extension of Scalars for Free Modules)* The module obtained from the free R-module $N \cong R^n$ by extension of scalars from R to S is the free S-module S^n, i.e.,

$$S \otimes_R R^n \cong S^n$$

as left S-modules.

Proof: This follows immediately from Theorem 17 and the isomorphism $S \otimes_R R \cong S$ proved in Example 7 previously.

Corollary 19. Let R be a commutative ring and let $M \cong R^s$ and $N \cong R^t$ be free R-modules with bases $m_1, \ldots, m_s$ and $n_1, \ldots, n_t$, respectively. Then $M \otimes_R N$ is a free R-module of rank st, with basis $m_i \otimes n_j$, $1 \leq i \leq s$ and $1 \leq j \leq t$, i.e.,

$$R^s \otimes_R R^t \cong R^{st}.$$

Remark: More generally, the tensor product of two free modules of arbitrary rank over a commutative ring is free (cf. the exercises).

Proof: This follows easily from Theorem 17 and the first example following Corollary 9.

Proposition 20. Suppose R is a commutative ring and M, N are left R-modules, considered with the standard R-module structures. Then there is a unique R-module isomorphism

$$M \otimes_R N \cong N \otimes_R M$$

mapping $m \otimes n$ to $n \otimes m$.

Proof: The map $M \times N \to N \otimes M$ defined by $(m, n) \mapsto n \otimes m$ is R-balanced. Hence it induces a unique homomorphism f from $M \otimes N$ to $N \otimes M$ with $f(m \otimes n) = n \otimes m$. Similarly, we have a unique homomorphism g from $N \otimes M$ to $M \otimes N$ with $g(n \otimes m) = m \otimes n$ giving the inverse of f, and both maps are easily seen to be R-module isomorphisms.

Remark: When $M = N$ it is not in general true that $a \otimes b = b \otimes a$ for $a, b \in M$. We shall study "symmetric tensors" in Section 11.6.

We end this section by showing that the tensor product of R-algebras is again an R-algebra.

Proposition 21. Let R be a commutative ring and let A and B be R-algebras. Then the multiplication $(a \otimes b)(a' \otimes b') = aa' \otimes bb'$ is well defined and makes $A \otimes_R B$ into an R-algebra.

Proof: Note first that the definition of an R-algebra shows that

$$r(a \otimes b) = ra \otimes b = ar \otimes b = a \otimes rb = a \otimes br = (a \otimes b)r$$

for every $r \in R$, $a \in A$ and $b \in B$. To show that $A \otimes B$ is an R-algebra the main task is, as usual, showing that the specified multiplication is well defined. One way to proceed is to use two applications of Corollary 16, as follows. The map $\varphi : A \times B \times A \times B \to A \otimes B$ defined by $f(a, b, a', b') = aa' \otimes bb'$ is multilinear over R. For example,

$$\begin{aligned} f(a, r_1 b_1 + r_2 b_2, a', b') &= aa' \otimes (r_1 b_1 + r_2 b_2) b' \\ &= aa' \otimes r_1 b_1 b' + aa' \otimes r_2 b_2 b' \\ &= r_1 f(a, b_1, a', b') + r_2 f(a, b_2, a', b'). \end{aligned}$$

By Corollary 16, there is a corresponding R-module homomorphism Φ from $A \otimes B \otimes A \otimes B$ to $A \otimes B$ with $\Phi(a \otimes b \otimes a' \otimes b') = aa' \otimes bb'$. Viewing $A \otimes B \otimes A \otimes B$ as $(A \otimes B) \otimes (A \otimes B)$, we can apply Corollary 16 once more to obtain a well defined R-bilinear mapping φ' from $(A \otimes B) \times (A \otimes B)$ to $A \otimes B$ with $\varphi'(a \otimes b, a' \otimes b') = aa' \otimes bb'$. This shows that the multiplication is indeed well defined (and also that it satisfies the distributive laws). It is now a simple matter (left to the exercises) to check that with this multiplication $A \otimes B$ is an R-algebra.

Example

The tensor product $\mathbb{C} \otimes_{\mathbb{R}} \mathbb{C}$ is free of rank 4 as a module over $\mathbb{R}$ with basis given by $e_1 = 1 \otimes 1$, $e_2 = 1 \otimes i$, $e_3 = i \otimes 1$, and $e_4 = i \otimes i$ (by Corollary 19). By Proposition 21, this tensor product is also a (commutative) ring with $e_1 = 1$, and, for example,

$$e_4^2 = (i \otimes i)(i \otimes i) = i^2 \otimes i^2 = (-1) \otimes (-1) = (-1)(-1) \otimes 1 = 1.$$

Then $(e_4 - 1)(e_4 + 1) = 0$, so $\mathbb{C} \otimes_{\mathbb{R}} \mathbb{C}$ is not an integral domain.

The ring $\mathbb{C} \otimes_{\mathbb{R}} \mathbb{C}$ is an $\mathbb{R}$-algebra and the left and right $\mathbb{R}$-actions are the same: $xr = rx$ for every $r \in \mathbb{R}$ and $x \in \mathbb{C} \otimes_{\mathbb{R}} \mathbb{C}$. The ring $\mathbb{C} \otimes_{\mathbb{R}} \mathbb{C}$ has a structure of a left $\mathbb{C}$-module because the first $\mathbb{C}$ is a $(\mathbb{C}, \mathbb{R})$-bimodule. It also has a right $\mathbb{C}$-module structure because the second $\mathbb{C}$ is an $(\mathbb{R}, \mathbb{C})$-bimodule. For example,

$$i \cdot e_1 = i \cdot (1 \otimes 1) = (i \cdot 1) \otimes 1 = i \otimes 1 = e_3$$

and

$$e_1 \cdot i = (1 \otimes 1) \cdot i = 1 \otimes (1 \cdot i) = 1 \otimes i = e_2.$$

This example also shows that even when the rings involved are commutative there may be natural left and right module structures (over some ring) that are not the same.

EXERCISES

Let R be a ring with 1.

1. Let $f : R \to S$ be a ring homomorphism from the ring R to the ring S with $f(1_R) = 1_S$. Verify the details that $sr = sf(r)$ defines a right R-action on S under which S is an (S, R)-bimodule.
2. Show that the element "$2 \otimes 1$" is 0 in $\mathbb{Z} \otimes_{\mathbb{Z}} \mathbb{Z}/2\mathbb{Z}$ but is nonzero in $2\mathbb{Z} \otimes_{\mathbb{Z}} \mathbb{Z}/2\mathbb{Z}$.
3. Show that $\mathbb{C} \otimes_{\mathbb{R}} \mathbb{C}$ and $\mathbb{C} \otimes_{\mathbb{C}} \mathbb{C}$ are both left $\mathbb{R}$-modules but are not isomorphic as $\mathbb{R}$-modules.
4. Show that $\mathbb{Q} \otimes_{\mathbb{Z}} \mathbb{Q}$ and $\mathbb{Q} \otimes_{\mathbb{Q}} \mathbb{Q}$ are isomorphic left $\mathbb{Q}$-modules. [Show they are both 1-dimensional vector spaces over $\mathbb{Q}$.]
5. Let A be a finite abelian group of order n and let p^k be the largest power of the prime p dividing n. Prove that $\mathbb{Z}/p^k\mathbb{Z} \otimes_{\mathbb{Z}} A$ is isomorphic to the Sylow p-subgroup of A.
6. If R is any integral domain with quotient field Q, prove that $(Q/R) \otimes_R (Q/R) = 0$.
7. If R is any integral domain with quotient field Q and N is a left R-module, prove that every element of the tensor product $Q \otimes_R N$ can be written as a simple tensor of the form $(1/d) \otimes n$ for some nonzero $d \in R$ and some $n \in N$.
8. Suppose R is an integral domain with quotient field Q and let N be any R-module. Let $U = R^\times$ be the set of nonzero elements in R and define $U^{-1}N$ to be the set of equivalence classes of ordered pairs of elements (u, n) with $u \in U$ and $n \in N$ under the equivalence relation $(u, n) \sim (u', n)$ if and only if $u'n = un'$ in N.

(a) Prove that $U^{-1}N$ is an abelian group under the addition defined by $\overline{(u_1, n_1)} + \overline{(u_2, n_2)} = \overline{(u_1u_2, u_2n_1 + u_1n_2)}$. Prove that $r\overline{(u, n)} = \overline{(u, rn)}$ defines an action of R on $U^{-1}N$ making it into an R-module. [This is an example of *localization* considered in general in Section 4 of Chapter 15, cf. also Section 5 in Chapter 7.]

(b) Show that the map from $Q \times N$ to $U^{-1}N$ defined by sending $(a/b, n)$ to $\overline{(b, an)}$ for $a \in R$, $b \in U$, $n \in N$, is an R-balanced map, so induces a homomorphism f from $Q \otimes_R N$ to $U^{-1}N$. Show that the map g from $U^{-1}N$ to $Q \otimes_R N$ defined by $g(\overline{(u, n)}) = (1/u) \otimes n$ is well defined and is an inverse homomorphism to f. Conclude that $Q \otimes_R N \cong U^{-1}N$ as R-modules.

(c) Conclude from (b) that $(1/d) \otimes n$ is 0 in $Q \otimes_R N$ if and only if $rn = 0$ for some nonzero $r \in R$.

(d) If A is an abelian group, show that $\mathbb{Q} \otimes_{\mathbb{Z}} A = 0$ if and only if A is a torsion abelian group (i.e., every element of A has finite order).

9. Suppose R is an integral domain with quotient field Q and let N be any R-module. Let $Q \otimes_R N$ be the module obtained from N by extension of scalars from R to Q. Prove that the kernel of the R-module homomorphism $\iota : N \to Q \otimes_R N$ is the torsion submodule of N (cf. Exercise 8 in Section 1). [Use the previous exercise.]

10. Suppose R is commutative and $N \cong R^n$ is a free R-module of rank n with R-module basis $e_1, \ldots, e_n$.

(a) For any nonzero R-module M show that every element of $M \otimes N$ can be written uniquely in the form $\sum_{i=1}^{n} m_i \otimes e_i$ where $m_i \in M$. Deduce that if $\sum_{i=1}^{n} m_i \otimes e_i = 0$ in $M \otimes N$ then $m_i = 0$ for $i = 1, \ldots, n$.

(b) Show that if $\sum m_i \otimes n_i = 0$ in $M \otimes N$ where the n_i are merely assumed to be R-linearly independent then it is not necessarily true that all the m_i are 0. [Consider $R = \mathbb{Z}$, $n = 1$, $M = \mathbb{Z}/2\mathbb{Z}$, and the element $1 \otimes 2$.]

11. Let $\{e_1, e_2\}$ be a basis of $V = \mathbb{R}^2$. Show that the element $e_1 \otimes e_2 + e_2 \otimes e_1$ in $V \otimes_{\mathbb{R}} V$ cannot be written as a simple tensor $v \otimes w$ for any $v, w \in \mathbb{R}^2$.

12. Let V be a vector space over the field F and let v, v' be nonzero elements of V. Prove that $v \otimes v' = v' \otimes v$ in $V \otimes_F V$ if and only if $v = av'$ for some $a \in F$.

13. Prove that the usual dot product of vectors defined by letting $(a_1, \ldots, a_n) \cdot (b_1, \ldots, b_n)$ be $a_1b_1 + \cdots + a_nb_n$ is a bilinear map from $\mathbb{R}^n \times \mathbb{R}^n$ to $\mathbb{R}$.

14. Let I be an arbitrary nonempty index set and for each $i \in I$ let N_i be a left R-module. Let M be a right R-module. Prove the group isomorphism: $M \otimes (\oplus_{i \in I} N_i) \cong \oplus_{i \in I} (M \otimes N_i)$, where the direct sum of an arbitrary collection of modules is defined in Exercise 20, Section 3. [Use the same argument as for the direct sum of two modules, taking care to note where the direct *sum* hypothesis is needed — cf. the next exercise.]

15. Show that tensor products do not commute with direct products in general. [Consider the extension of scalars from $\mathbb{Z}$ to $\mathbb{Q}$ of the direct product of the modules $M_i = \mathbb{Z}/2^i\mathbb{Z}$, $i = 1, 2, \ldots$]

16. Suppose R is commutative and let I and J be ideals of R, so R/I and R/J are naturally R-modules.

(a) Prove that every element of $R/I \otimes_R R/J$ can be written as a simple tensor of the form $(1 \bmod I) \otimes (r \bmod J)$.

(b) Prove that there is an R-module isomorphism $R/I \otimes_R R/J \cong R/(I + J)$ mapping $(r \bmod I) \otimes (r' \bmod J)$ to $rr' \bmod (I + J)$.

17. Let $I = (2, x)$ be the ideal generated by 2 and x in the ring $R = \mathbb{Z}[x]$. The ring $\mathbb{Z}/2\mathbb{Z} = R/I$ is naturally an R-module annihilated by both 2 and x.

(a) Show that the map $\varphi : I \times I \to \mathbb{Z}/2\mathbb{Z}$ defined by

$$\varphi(a_0 + a_1x + \cdots + a_nx^n, b_0 + b_1x + \cdots + b_mx^m) = \frac{a_0}{2}b_1 \bmod 2$$

is R-bilinear.

(b) Show that there is an R-module homomorphism from $I \otimes_R I \to \mathbb{Z}/2\mathbb{Z}$ mapping $p(x) \otimes q(x)$ to $\frac{p(0)}{2}q'(0)$ where q' denotes the usual polynomial derivative of q.

(c) Show that $2 \otimes x \neq x \otimes 2$ in $I \otimes_R I$.

18. Suppose I is a principal ideal in the integral domain R. Prove that the R-module $I \otimes_R I$ has no nonzero torsion elements (i.e., $rm = 0$ with $0 \neq r \in R$ and $m \in I \otimes_R I$ implies that $m = 0$).

19. Let $I = (2, x)$ be the ideal generated by 2 and x in the ring $R = \mathbb{Z}[x]$ as in Exercise 17. Show that the nonzero element $2 \otimes x - x \otimes 2$ in $I \otimes_R I$ is a torsion element. Show in fact that $2 \otimes x - x \otimes 2$ is annihilated by both 2 and x and that the submodule of $I \otimes_R I$ generated by $2 \otimes x - x \otimes 2$ is isomorphic to R/I.

20. Let $I = (2, x)$ be the ideal generated by 2 and x in the ring $R = \mathbb{Z}[x]$. Show that the element $2 \otimes 2 + x \otimes x$ in $I \otimes_R I$ is not a simple tensor, i.e., cannot be written as $a \otimes b$ for some $a, b \in I$.

21. Suppose R is commutative and let I and J be ideals of R.

(a) Show there is a surjective R-module homomorphism from $I \otimes_R J$ to the product ideal IJ mapping $i \otimes j$ to the element ij.

(b) Give an example to show that the map in (a) need not be injective (cf. Exercise 17).

22. Suppose that M is a left and a right R-module such that $rm = mr$ for all $r \in R$ and $m \in M$. Show that the elements r_1r_2 and r_2r_1 act the same on M for every $r_1, r_2 \in R$. (This explains why the assumption that R is commutative in the definition of an R-algebra is a fairly natural one.)

23. Verify the details that the multiplication in Proposition 19 makes $A \otimes_R B$ into an R-algebra.

24. Prove that the extension of scalars from $\mathbb{Z}$ to the Gaussian integers $\mathbb{Z}[i]$ of the ring $\mathbb{R}$ is isomorphic to $\mathbb{C}$ as a ring: $\mathbb{Z}[i] \otimes_{\mathbb{Z}} \mathbb{R} \cong \mathbb{C}$ as rings.

25. Let R be a subring of the commutative ring S and let x be an indeterminate over S. Prove that $S[x]$ and $S \otimes_R R[x]$ are isomorphic as S-algebras.

26. Let S be a commutative ring containing R (with $1_S = 1_R$) and let $x_1, \ldots, x_n$ be independent indeterminates over the ring S. Show that for every ideal I in the polynomial ring $R[x_1, \ldots, x_n]$ that $S \otimes_R (R[x_1, \ldots, x_n]/I) \cong S[x_1, \ldots, x_n]/IS[x_1, \ldots, x_n]$ as S-algebras.

The next exercise shows the ring $\mathbb{C} \otimes_{\mathbb{R}} \mathbb{C}$ introduced at the end of this section is isomorphic to $\mathbb{C} \times \mathbb{C}$. One may also prove this via Exercise 26 and Proposition 16 in Section 9.5, since $\mathbb{C} \cong \mathbb{R}[x]/(x^2 + 1)$. The ring $\mathbb{C} \times \mathbb{C}$ is also discussed in Exercise 23 of Section 1.

27. **(a)** Write down a formula for the multiplication of two elements $a \cdot 1 + b \cdot e_2 + c \cdot e_3 + d \cdot e_4$ and $a' \cdot 1 + b' \cdot e_2 + c' \cdot e_3 + d' \cdot e_4$ in the example $A = \mathbb{C} \otimes_{\mathbb{R}} \mathbb{C}$ following Proposition 21 (where $1 = 1 \otimes 1$ is the identity of A).

(b) Let $\epsilon_1 = \frac{1}{2}(1 \otimes 1 + i \otimes i)$ and $\epsilon_2 = \frac{1}{2}(1 \otimes 1 - i \otimes i)$. Show that $\epsilon_1\epsilon_2 = 0, \epsilon_1 + \epsilon_2 = 1$, and $\epsilon_j^2 = \epsilon_j$ for $j = 1, 2$ (ϵ_1 and ϵ_2 are called *orthogonal idempotents* in A). Deduce that A is isomorphic as a ring to the direct product of two principal ideals: $A \cong A\epsilon_1 \times A\epsilon_2$ (cf. Exercise 1, Section 7.6).

(c) Prove that the map $\varphi : \mathbb{C} \times \mathbb{C} \to \mathbb{C} \times \mathbb{C}$ by $\varphi(z_1, z_2) = (z_1z_2, z_1\overline{z_2})$, where $\overline{z_2}$ denotes the complex conjugate of z_2, is an $\mathbb{R}$-bilinear map.

(d) Let Φ be the $\mathbb{R}$-module homomorphism from A to $\mathbb{C} \times \mathbb{C}$ obtained from φ in (c). Show that $\Phi(\epsilon_1) = (0, 1)$ and $\Phi(\epsilon_2) = (1, 0)$. Show also that Φ is $\mathbb{C}$-linear, where the action of $\mathbb{C}$ is on the left tensor factor in A and on both factors in $\mathbb{C} \times \mathbb{C}$. Deduce that Φ is surjective. Show that Φ is a $\mathbb{C}$-algebra isomorphism.

10.5 EXACT SEQUENCES—PROJECTIVE, INJECTIVE, AND FLAT MODULES

One of the fundamental results for studying the structure of an algebraic object B (e.g., a group, a ring, or a module) is the First Isomorphism Theorem, which relates the subobjects of B (the normal subgroups, the ideals, or the submodules, respectively) with the possible homomorphic images of B. We have already seen many examples applying this theorem to understand the structure of B from an understanding of its "smaller" constituents—for example in analyzing the structure of the dihedral group D_8 by determining its center and the resulting quotient by the center.

In most of these examples we began *first* with a given B and then determined some of its basic properties by constructing a homomorphism φ (often given implicitly by the specification of $\ker \varphi \subseteq B$) and examining both $\ker \varphi$ and the resulting quotient $B/\ker \varphi$. We now consider in some greater detail the reverse situation, namely whether we may *first* specify the "smaller constituents." More precisely, we consider whether, given two modules A and C, there exists a module B containing (an isomorphic copy of) A such that the resulting quotient module B/A is isomorphic to C—in which case B is said to be an *extension of C by A*. It is then natural to ask how many such B exist for a given A and C, and the extent to which properties of B are determined by the corresponding properties of A and C. There are, of course, analogous problems in the contexts of groups and rings. This is the *extension problem* first discussed (for groups) in Section 3.4; in this section we shall be primarily concerned with left modules over a ring R, making note where necessary of the modifications required for some other structures, notably noncommutative groups. As in the previous section, throughout this section all rings contain a 1.

We first introduce a very convenient notation. To say that A is isomorphic to a submodule of B, is to say that there is an injective homomorphism $\psi : A \to B$ (so then $A \cong \psi(A) \subseteq B$). To say that C is isomorphic to the resulting quotient is to say that there is a surjective homomorphism $\varphi : B \to C$ with $\ker \varphi = \psi(A)$. In particular this gives us a pair of homomorphisms:

$$A \xrightarrow{\psi} B \xrightarrow{\varphi} C$$

with image $\psi = \ker \varphi$. A pair of homomorphisms with this property is given a name:

Definition.

(1) The pair of homomorphisms $X \xrightarrow{\alpha} Y \xrightarrow{\beta} Z$ is said to be *exact* (at Y) if image $\alpha = \ker \beta$.

(2) A sequence $\cdots \to X_{n-1} \to X_n \to X_{n+1} \to \cdots$ of homomorphisms is said to be an *exact sequence* if it is exact at every X_n between a pair of homomorphisms.

With this terminology, the pair of homomorphisms $A \xrightarrow{\psi} B \xrightarrow{\varphi} C$ above is exact at B. We can also use this terminology to express the fact that for these maps ψ is injective and φ is surjective:

Proposition 22. Let A, B and C be R-modules over some ring R. Then

(1) The sequence $0 \to A \xrightarrow{\psi} B$ is exact (at A) if and only if ψ is injective.

(2) The sequence $B \xrightarrow{\varphi} C \to 0$ is exact (at C) if and only if φ is surjective.

Proof: The (uniquely defined) homomorphism $0 \to A$ has image 0 in A. This will be the kernel of ψ if and only if ψ is injective. Similarly, the kernel of the (uniquely defined) zero homomorphism $C \to 0$ is all of C, which is the image of φ if and only if φ is surjective.

Corollary 23. The sequence $0 \to A \xrightarrow{\psi} B \xrightarrow{\varphi} C \to 0$ is exact if and only if ψ is injective, φ is surjective, and image $\psi = \ker \varphi$, i.e., B is an extension of C by A.

Definition. The exact sequence $0 \to A \xrightarrow{\psi} B \xrightarrow{\varphi} C \to 0$ is called a *short exact sequence*.

In terms of this notation, the extension problem can be stated succinctly as follows: given modules A and C, determine all the short exact sequences

$$0 \longrightarrow A \xrightarrow{\psi} B \xrightarrow{\varphi} C \longrightarrow 0. \tag{10.9}$$

We shall see below that the exact sequence notation is also extremely convenient for analyzing the extent to which properties of A and C determine the corresponding properties of B. If A, B and C are groups written multiplicatively, the sequence (9) will be written

$$1 \longrightarrow A \xrightarrow{\psi} B \xrightarrow{\varphi} C \longrightarrow 1 \tag{10.9'}$$

where 1 denotes the trivial group. Both Proposition 22 and Corollary 23 are valid with the obvious notational changes.

Note that any exact sequence can be written as a succession of short exact sequences since to say $X \xrightarrow{\alpha} Y \xrightarrow{\beta} Z$ is exact at Y is the same as saying that the sequence $0 \to \alpha(X) \to Y \to Y/\ker \beta \to 0$ is a short exact sequence.

Examples

(1) Given modules A and C we can always form their direct sum $B = A \oplus C$ and the sequence

$$0 \to A \xrightarrow{\iota} A \oplus C \xrightarrow{\pi} C \to 0$$

where $\iota(a) = (a, 0)$ and $\pi(a, c) = c$ is a short exact sequence. In particular, it follows that there always exists at least one extension of C by A.

(2) As a special case of the previous example, consider the two $\mathbb{Z}$-modules $A = \mathbb{Z}$ and $C = \mathbb{Z}/n\mathbb{Z}$:

$$0 \longrightarrow \mathbb{Z} \xrightarrow{\iota} \mathbb{Z} \oplus (\mathbb{Z}/n\mathbb{Z}) \xrightarrow{\varphi} \mathbb{Z}/n\mathbb{Z} \longrightarrow 0,$$

giving one extension of $\mathbb{Z}/n\mathbb{Z}$ by $\mathbb{Z}$.

Another extension of $\mathbb{Z}/n\mathbb{Z}$ by $\mathbb{Z}$ is given by the short exact sequence

$$0 \to \mathbb{Z} \xrightarrow{n} \mathbb{Z} \xrightarrow{\pi} \mathbb{Z}/n\mathbb{Z} \to 0$$

where n denotes the map $x \mapsto nx$ given by multiplication by n, and π denotes the natural projection. Note that the modules in the middle of the previous two exact sequences are not isomorphic even though the respective "A" and "C" terms are isomorphic. Thus there are (at least) two "essentially different" or "inequivalent" ways of extending $\mathbb{Z}/n\mathbb{Z}$ by $\mathbb{Z}$.

(3) If $\varphi : B \to C$ is any homomorphism we may form an exact sequence:

$$0 \longrightarrow \ker\varphi \xrightarrow{\iota} B \xrightarrow{\varphi} \text{image}\,\varphi \longrightarrow 0$$

where ι is the inclusion map. In particular, if φ is surjective, the sequence $\varphi : B \to C$ may be extended to a short exact sequence with $A = \ker\varphi$.

(4) One particularly important instance of the preceding example is when M is an R-module and S is a set of generators for M. Let $F(S)$ be the free R-module on S. Then

$$0 \longrightarrow K \xrightarrow{\iota} F(S) \xrightarrow{\varphi} M \longrightarrow 0$$

is the short exact sequence where φ is the unique R-module homomorphism which is the identity on S (cf. Theorem 6) and $K = \ker\varphi$.

More generally, when M is any group (possibly non-abelian) the above short exact sequence (with 1's at the ends, if M is written multiplicatively) describes a *presentation* of M, where K is the normal subgroup of $F(S)$ generated by the *relations* defining M (cf. Section 6.3).

(5) Two "inequivalent" extensions G of the Klein 4-group by the cyclic group Z_2 of order two are

$$1 \longrightarrow Z_2 \xrightarrow{\iota} D_8 \xrightarrow{\varphi} Z_2 \times Z_2 \longrightarrow 1, \text{ and}$$
$$1 \longrightarrow Z_2 \xrightarrow{\iota} Q_8 \xrightarrow{\varphi} Z_2 \times Z_2 \longrightarrow 1,$$

where in each case ι maps Z_2 injectively into the center of G (recall that both D_8 and Q_8 have centers of order two), and φ is the natural projection of G onto $G/Z(G)$.

Two other inequivalent extensions G of the Klein 4-group by Z_2 occur when G is either of the abelian groups $Z_2 \times Z_2 \times Z_2$ or $Z_2 \times Z_4$ for appropriate maps ι and φ.

Examples 2 and 5 above show that, for a fixed A and C, in general there may be several extensions of C by A. To distinguish different extensions we define the notion of a homomorphism (and isomorphism) between two exact sequences. Recall first that a diagram involving various homomorphisms is said to *commute* if any compositions of homomorphisms with the same starting and ending points are equal, i.e., the composite map defined by following a path of homomorphisms in the diagram depends only on the starting and ending points and not on the choice of the path taken.

Definition. Let $0 \to A \to B \to C \to 0$ and $0 \to A' \to B' \to C' \to 0$ be two short exact sequences of modules.

(1) A *homomorphism of short exact sequences* is a triple α, β, γ of module homomorphisms such that the following diagram commutes:

$$\begin{array}{ccccccccc} 0 & \longrightarrow & A & \longrightarrow & B & \longrightarrow & C & \longrightarrow & 0 \\ & & \downarrow{\scriptstyle \alpha} & & \downarrow{\scriptstyle \beta} & & \downarrow{\scriptstyle \gamma} & & \\ 0 & \longrightarrow & A' & \longrightarrow & B' & \longrightarrow & C' & \longrightarrow & 0 \end{array}$$

The homomorphism is an *isomorphism of short exact sequences* if α, β, γ are all isomorphisms, in which case the extensions B and B' are said to be *isomorphic extensions.*

(2) The two exact sequences are called *equivalent* if $A = A'$, $C = C'$, and there is an isomorphism between them as in (1) that is the identity maps on A and C (i.e., α and γ are the identity). In this case the corresponding extensions B and B' are said to be *equivalent* extensions.

If B and B' are isomorphic extensions then in particular B and B' are isomorphic as R-modules, but more is true: there is an R-module isomorphism between B and B' that restricts to an isomorphism from A to A' and induces an isomorphism on the quotients C and C'. For a given A and C the condition that two extensions B and B' of C by A are equivalent is stronger still: there must exist an R-module isomorphism between B and B' that restricts to the *identity* map on A and induces the *identity* map on C. The notion of isomorphic extensions measures how many different extensions of C by A there are, allowing for C and A to be changed by an isomorphism. The notion of equivalent extensions measures how many different extensions of C by A there are when A and C are rigidly fixed.

Homomorphisms and isomorphisms between short exact sequences of multiplicative groups (9′) are defined similarly.

It is an easy exercise to see that the composition of homomorphisms of short exact sequences is also a homomorphism. Likewise, if the triple α, β, γ is an isomorphism (or equivalence) then $\alpha^{-1}, \beta^{-1}, \gamma^{-1}$ is an isomorphism (equivalence, respectively) in the reverse direction. It follows that "isomorphism" (or equivalence) is an equivalence relation on any set of short exact sequences.

Examples

(1) Let m and n be integers greater than 1. Assume n divides m and let $k = m/n$. Define a map from the exact sequence of $\mathbb{Z}$-modules in Example 2 of the preceding set of examples:

$$\begin{array}{ccccccccc} 0 & \longrightarrow & \mathbb{Z} & \xrightarrow{n} & \mathbb{Z} & \xrightarrow{\pi} & \mathbb{Z}/n\mathbb{Z} & \longrightarrow & 0 \\ & & \downarrow{\scriptstyle \alpha} & & \downarrow{\scriptstyle \beta} & & \downarrow{\scriptstyle \gamma} & & \\ 0 & \longrightarrow & \mathbb{Z}/k\mathbb{Z} & \xrightarrow{\iota} & \mathbb{Z}/m\mathbb{Z} & \xrightarrow{\pi'} & \mathbb{Z}/n\mathbb{Z} & \longrightarrow & 0 \end{array}$$

where α and β are the natural projections, γ is the identity map, ι maps $a \bmod k$ to $na \bmod m$, and π' is the natural projection of $\mathbb{Z}/m\mathbb{Z}$ onto its quotient $(\mathbb{Z}/m\mathbb{Z})/(n\mathbb{Z}/m\mathbb{Z})$

(which is isomorphic to $\mathbb{Z}/n\mathbb{Z}$). One easily checks that this is a homomorphism of short exact sequences.

(2) If again $0 \to \mathbb{Z} \xrightarrow{n} \mathbb{Z} \xrightarrow{\pi} \mathbb{Z}/n\mathbb{Z} \to 0$ is the short exact sequence of $\mathbb{Z}$-modules defined previously, map each module to itself by $x \mapsto -x$. This triple of homomorphisms gives an isomorphism of the exact sequence with itself. This isomorphism is not an equivalence of sequences since it is not the identity on the first $\mathbb{Z}$.

(3) The short exact sequences in Examples 1 and 2 following Corollary 23 are not isomorphic—the extension modules are not isomorphic $\mathbb{Z}$-modules (abelian groups). Likewise the two extensions, D_8 and Q_8, in Example 5 of the same set are not isomorphic (hence not equivalent), even though the two end terms "A" and "C" are the same for both sequences.

(4) Consider the maps

$$\begin{array}{ccccccccc} 0 & \longrightarrow & \mathbb{Z}/2\mathbb{Z} & \xrightarrow{\psi} & \mathbb{Z}/2\mathbb{Z} \oplus \mathbb{Z}/2\mathbb{Z} & \xrightarrow{\varphi} & \mathbb{Z}/2\mathbb{Z} & \longrightarrow & 0 \\ & & \downarrow \mathrm{id} & & \downarrow \beta & & \downarrow \mathrm{id} & & \\ 0 & \longrightarrow & \mathbb{Z}/2\mathbb{Z} & \xrightarrow{\psi'} & \mathbb{Z}/2\mathbb{Z} \oplus \mathbb{Z}/2\mathbb{Z} & \xrightarrow{\varphi'} & \mathbb{Z}/2\mathbb{Z} & \longrightarrow & 0 \end{array}$$

where ψ maps $\mathbb{Z}/2\mathbb{Z}$ injectively into the first component of the direct sum and φ projects the direct sum onto its second component. Also ψ' embeds $\mathbb{Z}/2\mathbb{Z}$ into the *second* component of the direct sum and φ' projects the direct sum onto its *first* component. If β maps the direct sum $\mathbb{Z}/2\mathbb{Z} \oplus \mathbb{Z}/2\mathbb{Z}$ to itself by interchanging the two factors, then this diagram is seen to commute, hence giving an equivalence of the two exact sequences that is not the identity isomorphism.

(5) We exhibit two isomorphic but inequivalent $\mathbb{Z}$-module extensions. For $i = 1, 2$ define

$$0 \longrightarrow \mathbb{Z}/2\mathbb{Z} \xrightarrow{\psi} \mathbb{Z}/4\mathbb{Z} \oplus \mathbb{Z}/2\mathbb{Z} \xrightarrow{\varphi_i} \mathbb{Z}/2\mathbb{Z} \oplus \mathbb{Z}/2\mathbb{Z} \longrightarrow 0$$

where $\psi : 1 \mapsto (2, 0)$ in both sequences, φ_1 is defined by $\varphi_1(a \bmod 4, b \bmod 2) = (a \bmod 2, b \bmod 2)$, and $\varphi_2(a \bmod 4, b \bmod 2) = (b \bmod 2, a \bmod 2)$. It is easy to see that the resulting two sequences are both short exact sequences.

An evident isomorphism between these two exact sequences is provided by the triple of maps id, id, γ, where $\gamma : \mathbb{Z}/2\mathbb{Z} \oplus \mathbb{Z}/2\mathbb{Z} \to \mathbb{Z}/2\mathbb{Z} \oplus \mathbb{Z}/2\mathbb{Z}$ is the map $\gamma((c, d)) = (d, c)$ that interchanges the two direct factors.

We now check that these two isomorphic sequences are *not equivalent*, as follows. Since $\varphi_1(0, 1) = (0, 1)$, any equivalence, id, β, id, from the first sequence to the second must map $(0, 1) \in \mathbb{Z}/4\mathbb{Z} \oplus \mathbb{Z}/2\mathbb{Z}$ to either $(1, 0)$ or $(3, 0)$ in $\mathbb{Z}/4\mathbb{Z} \oplus \mathbb{Z}/2\mathbb{Z}$, since these are the two possible elements mapping to $(0, 1)$ by φ_2. This is impossible, however, since the isomorphism β cannot send an element of order 2 to an element of order 4.

Put another way, equivalences involving the same extension module B are automorphisms of B that restrict to the identity on both $\psi(A)$ and $B/\psi(A)$. Any such automorphism of $B = \mathbb{Z}/4\mathbb{Z} \oplus \mathbb{Z}/2\mathbb{Z}$ must fix the coset $(0, 1) + \psi(A)$ since this is the unique nonidentity coset containing elements of order 2. Thus maps which send this coset to different elements in C give inequivalent extensions. In particular, there is yet a third inequivalent extension involving the same modules $A = \mathbb{Z}/2\mathbb{Z}$, $B = \mathbb{Z}/4\mathbb{Z} \oplus \mathbb{Z}/2\mathbb{Z}$ and $C = \mathbb{Z}/2\mathbb{Z} \oplus \mathbb{Z}/2\mathbb{Z}$, that maps the coset $(0, 1) + \psi(A)$ to the element $(1, 1) \in \mathbb{Z}/2\mathbb{Z} \oplus \mathbb{Z}/2\mathbb{Z}$.

By similar reasoning there are three inequivalent but isomorphic group extensions of $Z_2 \times Z_2$ by Z_2 with $B \cong D_8$ (cf. the exercises).

The homomorphisms α, β, γ in a homomorphism of short exact sequences are not independent. The next result gives some relations among these three homomorphisms.

Proposition 24. *(The Short Five Lemma)* Let α, β, γ be a homomorphism of short exact sequences

$$\begin{array}{ccccccccc} 0 & \longrightarrow & A & \longrightarrow & B & \longrightarrow & C & \longrightarrow & 0 \\ & & \downarrow{\scriptstyle \alpha} & & \downarrow{\scriptstyle \beta} & & \downarrow{\scriptstyle \gamma} & & \\ 0 & \longrightarrow & A' & \longrightarrow & B' & \longrightarrow & C' & \longrightarrow & 0 \end{array}$$

(1) If α and γ are injective then so is β.
(2) If α and γ are surjective then so is β.
(3) If α and γ are isomorphisms then so is β (and then the two sequences are isomorphic).

Remark: These results hold also for short exact sequences of (possibly non-abelian) groups (as the proof demonstrates).

Proof: We shall prove (1), leaving the proof of (2) as an exercise (and (3) follows immediately from (1) and (2)). Suppose then that α and γ are injective and suppose $b \in B$ with $\beta(b) = 0$. Let $\psi : A \to B$ and $\varphi : B \to C$ denote the homomorphisms in the first short exact sequence. Since $\beta(b) = 0$, it follows in particular that the image of $\beta(b)$ in the quotient C' is also 0. By the commutativity of the diagram this implies that $\gamma(\varphi(b)) = 0$, and since γ is assumed injective, we obtain $\varphi(b) = 0$, i.e., b is in the kernel of φ. By the exactness of the first sequence, this means that b is in the image of ψ, i.e., $b = \psi(a)$ for some $a \in A$. Then, again by the commutativity of the diagram, the image of $\alpha(a)$ in B' is the same as $\beta(\psi(a)) = \beta(b) = 0$. But α and the map from A' to B' are injective by assumption, and it follows that $a = 0$. Finally, $b = \psi(a) = \psi(0) = 0$ and we see that β is indeed injective.

We have already seen that there is always at least one extension of a module C by A, namely the direct sum $B = A \oplus C$. In this case the module B contains a submodule C' isomorphic to C (namely $C' = 0 \oplus C$) as well as the submodule A, and this submodule complement to A "splits" B into a direct sum. In the case of groups the existence of a subgroup complement C' to a normal subgroup in B implies that B is a semidirect product (cf. Section 5 in Chapter 5). The fact that B is a direct sum in the context of modules is a reflection of the fact that the underlying group structure in this case is *abelian*; for abelian groups semidirect products are direct products. In either case the corresponding short exact sequence is said to "split":

Definition.

(1) Let R be a ring and let $0 \to A \xrightarrow{\psi} B \xrightarrow{\varphi} C \to 0$ be a short exact sequence of R-modules. The sequence is said to be *split* if there is an R-module complement to $\psi(A)$ in B. In this case, up to isomorphism, $B = A \oplus C$ (more precisely, $B = \psi(A) \oplus C'$ for some submodule C', and C' is mapped isomorphically onto C by φ: $\varphi(C') \cong C$).

(2) If $1 \to A \xrightarrow{\psi} B \xrightarrow{\varphi} C \to 1$ is a short exact sequence of groups, then the sequence is said to be *split* if there is a subgroup complement to $\psi(A)$ in B. In this case, up to isomorphism, $B = A \rtimes C$ (more precisely, $B = \psi(A) \rtimes C'$ for some subgroup C', and C' is mapped isomorphically onto C by φ: $\varphi(C') \cong C$).

In either case the extension B is said to be a *split extension* of C by A.

The question of whether an extension splits is the question of the existence of a complement to $\psi(A)$ in B isomorphic (by φ) to C, so the notion of a split extension may equivalently be phrased in the language of homomorphisms:

Proposition 25. The short exact sequence $0 \to A \xrightarrow{\psi} B \xrightarrow{\varphi} C \to 0$ of R-modules is split if and only if there is an R-module homomorphism $\mu : C \to B$ such that $\varphi \circ \mu$ is the identity map on C. Similarly, the short exact sequence $1 \to A \xrightarrow{\psi} B \xrightarrow{\varphi} C \to 1$ of groups is split if and only if there is a group homomorphism $\mu : C \to B$ such that $\varphi \circ \mu$ is the identity map on C.

Proof: This follows directly from the definitions: if μ is given define $C' = \mu(C) \subseteq B$ and if C' is given define $\mu = \varphi^{-1} : C \cong C' \subseteq B$.

Definition. With notation as in Proposition 25, any set map $\mu : C \to B$ such that $\varphi \circ \mu = \mathrm{id}$ is called a *section* of φ. If μ is a *homomorphism* as in Proposition 25 then μ is called a *splitting homomorphism* for the sequence.

Note that a section of φ is nothing more than a choice of coset representatives in B for the quotient $B/\ker\varphi \cong C$. A section is a (splitting) homomorphism if this set of coset representatives forms a *submodule* (respectively, *subgroup*) in B, in which case this submodule (respectively, subgroup) gives a complement to $\psi(A)$ in B.

Examples

(1) The split short exact sequence $0 \to A \xrightarrow{\iota} A \oplus C \xrightarrow{\pi} C \to 0$ has the evident splitting homomorphism $\mu(c) = (0, c)$.

(2) The extension $0 \to \mathbb{Z} \xrightarrow{\iota} \mathbb{Z} \oplus (\mathbb{Z}/n\mathbb{Z}) \xrightarrow{\varphi} \mathbb{Z}/n\mathbb{Z} \to 0$, of $\mathbb{Z}/n\mathbb{Z}$ by $\mathbb{Z}$ is split (with splitting homomorphism μ mapping $\mathbb{Z}/n\mathbb{Z}$ isomorphically onto the second factor of the direct sum). On the other hand, the exact sequence of $\mathbb{Z}$-modules $0 \to \mathbb{Z} \xrightarrow{n} \mathbb{Z} \xrightarrow{\pi} \mathbb{Z}/n\mathbb{Z} \to 0$ is not split since there is no nonzero homomorphism of $\mathbb{Z}/n\mathbb{Z}$ into $\mathbb{Z}$.

(3) Neither D_8 nor Q_8 is a split extension of $Z_2 \times Z_2$ by Z_2 because in neither group is there a subgroup complement to the center (Section 2.5 gives the subgroup structures of these groups).

(4) The group D_8 *is* a split extension of Z_2 by Z_4, i.e., there is a split short exact sequence

$$1 \longrightarrow Z_4 \xrightarrow{\iota} D_8 \xrightarrow{\pi} Z_2 \longrightarrow 1,$$

namely,

$$1 \longrightarrow \langle r \rangle \xrightarrow{\iota} D_8 \xrightarrow{\pi} \langle \bar{s} \rangle \longrightarrow 1,$$

using our usual set of generators for D_8. Here ι is the inclusion map and $\pi : r^a s^b \mapsto \bar{s}^b$ is the projection onto the quotient $D_8/\langle r \rangle \cong Z_2$. The splitting homomorphism μ

maps $\langle \bar{s} \rangle$ isomorphically onto the complement $\langle s \rangle$ for $\langle r \rangle$ in D_8. Equivalently, D_8 is the semidirect product of the normal subgroup $\langle r \rangle$ (isomorphic to Z_4) with $\langle s \rangle$ (isomorphic to Z_2).

On the other hand, while Q_8 is also an extension of Z_2 by Z_4 (for example, $\langle i \rangle \cong Z_4$ has quotient isomorphic to Z_2), Q_8 is *not* a split extension of Z_2 by Z_4: no cyclic subgroup of Q_8 of order 4 has a complement in Q_8.

Section 5.5 contains many more examples of split extensions of groups.

Proposition 25 shows that an extension B of C by A is a split extension if and only if there is a splitting homomorphism μ of the projection map $\varphi : B \to C$ from B to the quotient C. The next proposition shows in particular that for modules this is equivalent to the existence of a splitting homomorphism for ψ at the other end of the sequence.

Proposition 26. Let $0 \to A \xrightarrow{\psi} B \xrightarrow{\varphi} C \to 0$ be a short exact sequence of modules (respectively, $1 \to A \xrightarrow{\psi} B \xrightarrow{\varphi} C \to 1$ a short exact sequence of groups). Then $B = \psi(A) \oplus C'$ for some submodule C' of B with $\varphi(C') \cong C$ (respectively, $B = \psi(A) \times C'$ for some subgroup C' of B with $\varphi(C') \cong C$) if and only if there is a homomorphism $\lambda : B \to A$ such that $\lambda \circ \psi$ is the identity map on A.

Proof: This is similar to the proof of Proposition 25. If λ is given, define $C' = \ker \lambda \subseteq B$ and if C' is given define $\lambda : B = \psi(A) \oplus C' \to A$ by $\lambda((\psi(a), c') = a$. Note that in this case $C' = \ker \lambda$ is *normal* in B, so that C' is a *normal* complement to $\psi(A)$ in B, which in turn implies that B is the *direct sum* of $\psi(A)$ and C' (cf. Theorem 9 of Section 5.4).

Proposition 26 shows that for general group extensions, the existence of a splitting homomorphism λ on the *left* end of the sequence is stronger than the condition that the extension splits: in this case the extension group is a *direct* product, and not just a *semidirect* product. The fact that these two notions are equivalent in the context of modules is again a reflection of the abelian nature of the underlying groups, where semidirect products are always direct products.

Modules and $Hom_R(D, \_\_)$

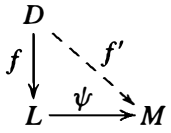
Let R be a ring with 1 and suppose the R-module M is an extension of N by L, with

$$0 \longrightarrow L \xrightarrow{\psi} M \xrightarrow{\varphi} N \longrightarrow 0$$

the corresponding short exact sequence of R-modules. It is natural to ask whether properties for L and N imply related properties for the extension M. The first situation we shall consider is whether an R-module homomorphism from some fixed R-module D to either L or N implies there is also an R-module homomorphism from D to M.

The question of obtaining a homomorphism from D to M given a homomorphism from D to L is easily disposed of: if $f \in \mathrm{Hom}_R(D, L)$ is an R-module homomorphism from D to L then the composite $f' = \psi \circ f$ is an R-module homomorphism from D to

M. The relation between these maps can be indicated pictorially by the commutative diagram



Put another way, composition with ψ induces a map

$$\psi' : \mathrm{Hom}_R(D, L) \longrightarrow \mathrm{Hom}_R(D, M)$$
$$f \longmapsto f' = \psi \circ f.$$

Recall that, by Proposition 2, $\mathrm{Hom}_R(D, L)$ and $\mathrm{Hom}_R(D, M)$ are abelian groups.

Proposition 27. Let D, L and M be R-modules and let $\psi : L \to M$ be an R-module homomorphism. Then the map

$$\psi' : \mathrm{Hom}_R(D, L) \longrightarrow \mathrm{Hom}_R(D, M)$$
$$f \longmapsto f' = \psi \circ f$$

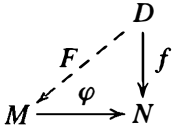
is a homomorphism of abelian groups. If ψ is injective, then ψ' is also injective, i.e.,

$$\text{if} \quad 0 \longrightarrow L \xrightarrow{\psi} M \quad \text{is exact,}$$

$$\text{then} \quad 0 \longrightarrow \mathrm{Hom}_R(D, L) \xrightarrow{\psi'} \mathrm{Hom}_R(D, M) \quad \text{is also exact.}$$

Proof: The fact that ψ' is a homomorphism is immediate. If ψ is injective, then distinct homomorphisms f and g from D into L give distinct homomorphisms $\psi \circ f$ and $\psi \circ g$ from D into M, which is to say that ψ' is also injective.

While obtaining homomorphisms into M from homomorphisms into the submodule L is straightforward, the situation for homomorphisms into the quotient N is much less evident. More precisely, given an R-module homomorphism $f : D \to N$ the question is whether there exists an R-module homomorphism $F : D \to M$ that *extends* or *lifts* f to M, i.e., that makes the following diagram commute:



As before, composition with the homomorphism φ induces a homomorphism of abelian groups

$$\varphi' : \mathrm{Hom}_R(D, M) \longrightarrow \mathrm{Hom}_R(D, N)$$
$$F \longmapsto F' = \varphi \circ F.$$

In terms of φ', the homomorphism f to N lifts to a homomorphism to M if and only if f is in the image of φ' (namely, f is the image of the lift F).

In general it may not be possible to lift a homomorphism f from D to N to a homomorphism from D to M. For example, consider the nonsplit exact sequence $0 \to \mathbb{Z} \xrightarrow{2} \mathbb{Z} \xrightarrow{\pi} \mathbb{Z}/2\mathbb{Z} \to 0$ from the previous set of examples. Let $D = \mathbb{Z}/2\mathbb{Z}$ and let f be the identity map from D into N. Any homomorphism F of D into $M = \mathbb{Z}$ must map D to 0 (since $\mathbb{Z}$ has no elements of order 2), hence $\pi \circ F$ maps D to 0 in N, and in particular, $\pi \circ F \neq f$. Phrased in terms of the map φ', this shows that

$$\text{if} \quad M \xrightarrow{\varphi} N \longrightarrow 0 \quad \text{is exact,}$$

$$\text{then} \quad \mathrm{Hom}_R(D, M) \xrightarrow{\varphi'} \mathrm{Hom}_R(D, N) \longrightarrow 0 \quad \text{is } \textit{not necessarily} \text{ exact.}$$

These results relating the homomorphisms into L and N to the homomorphisms into M can be neatly summarized as part of the following theorem.

Theorem 28. Let D, L, M, and N be R-modules. If

$$0 \longrightarrow L \xrightarrow{\psi} M \xrightarrow{\varphi} N \longrightarrow 0 \quad \text{is exact,}$$

then the associated sequence

$$0 \to \mathrm{Hom}_R(D, L) \xrightarrow{\psi'} \mathrm{Hom}_R(D, M) \xrightarrow{\varphi'} \mathrm{Hom}_R(D, N) \quad \text{is exact.} \tag{10.10}$$

A homomorphism $f : D \to N$ lifts to a homomorphism $F : D \to M$ if and only if $f \in \mathrm{Hom}_R(D, N)$ is in the image of φ'. In general $\varphi' : \mathrm{Hom}_R(D, M) \to \mathrm{Hom}_R(D, N)$ need not be surjective; the map φ' is surjective if and only if every homomorphism from D to N lifts to a homomorphism from D to M, in which case the sequence (10) can be extended to a short exact sequence.

The sequence (10) is exact for *all* R-modules D if and only if the sequence

$$0 \to L \xrightarrow{\psi} M \xrightarrow{\varphi} N \quad \text{is exact.}$$

Proof: The only item in the first statement that has not already been proved is the exactness of (10) at $\mathrm{Hom}_R(D, M)$, i.e., $\ker \varphi' = \text{image } \psi'$. Suppose $F : D \to M$ is an element of $\mathrm{Hom}_R(D, M)$ lying in the kernel of φ', i.e., with $\varphi \circ F = 0$ as homomorphisms from D to N. If $d \in D$ is any element of D, this implies that $\varphi(F(d)) = 0$ and $F(d) \in \ker \varphi$. By the exactness of the sequence defining the extension M we have $\ker \varphi = \text{image } \psi$, so there is some element $l \in L$ with $F(d) = \psi(l)$. Since ψ is injective, the element l is unique, so this gives a well defined map $F' : D \to L$ given by $F'(d) = l$. It is an easy check to verify that F' is a homomorphism, i.e., $F' \in \mathrm{Hom}_R(D, L)$. Since $\psi \circ F'(d) = \psi(l) = F(d)$, we have $F = \psi'(F')$ which shows that F is in the image of ψ', proving that $\ker \varphi' \subseteq \text{image } \psi'$. Conversely, if F is in the image of ψ' then $F = \psi'(F')$ for some $F' \in \mathrm{Hom}_R(D, L)$ and so $\varphi(F(d)) = \varphi(\psi(F'(d)))$ for any $d \in D$. Since $\ker \varphi = \text{image } \psi$ we have $\varphi \circ \psi = 0$, and it follows that $\varphi(F(d)) = 0$ for any $d \in D$, i.e., $\varphi'(F) = 0$. Hence F is in the kernel of φ', proving the reverse containment: $\text{image } \psi' \subseteq \ker \varphi'$.

For the last statement in the theorem, note first that the surjectivity of φ was not required for the proof that (10) is exact, so the "if" portion of the statement has already

been proved. For the converse, suppose that the sequence (10) is exact for all R-modules D. In general, $\mathrm{Hom}_R(R, X) \cong X$ for any left R-module X, the isomorphism being given by mapping a homomorphism to its value on the element $1 \in R$ (cf. Exercise 10(b)). Taking $D = R$ in (10), the exactness of the sequence $0 \to L \xrightarrow{\psi} M \xrightarrow{\varphi} N$ follows easily.

By Theorem 28, the sequence

$$0 \longrightarrow \mathrm{Hom}_R(D, L) \xrightarrow{\psi'} \mathrm{Hom}_R(D, M) \xrightarrow{\varphi'} \mathrm{Hom}_R(D, N) \longrightarrow 0 \qquad (10.11)$$

is in general *not* a short exact sequence since the homomorphism φ' need not be surjective. The question of whether this sequence is exact precisely measures the extent to which the homomorphisms from D into M are uniquely determined by pairs of homomorphisms from D into L and D into N. More precisely, this sequence is exact if and only if there is a bijection $F \leftrightarrow (g, f)$ between homomorphisms $F : D \to M$ and pairs of homomorphisms $g : D \to L$ and $f : D \to N$ given by $F|_{\psi(L)} = \psi'(g)$ and $f = \varphi'(F)$.

One situation in which the sequence (11) is exact occurs when the original sequence $0 \to L \to M \to N \to 0$ is a *split* exact sequence, i.e., when $M = L \oplus N$. In this case the sequence (11) is also a split exact sequence, as the first part of the following proposition shows.

Proposition 29. Let D, L and N be R-modules. Then
(1) $\mathrm{Hom}_R(D, L \oplus N) \cong \mathrm{Hom}_R(D, L) \oplus \mathrm{Hom}_R(D, N)$, and
(2) $\mathrm{Hom}_R(L \oplus N, D) \cong \mathrm{Hom}_R(L, D) \oplus \mathrm{Hom}_R(N, D)$.

Proof: Let $\pi_1 : L \oplus N \to L$ be the natural projection from $L \oplus N$ to L and similarly let π_2 be the natural projection to N. If $f \in \mathrm{Hom}_R(D, L \oplus N)$ then the compositions $\pi_1 \circ f$ and $\pi_2 \circ f$ give elements in $\mathrm{Hom}_R(D, L)$ and $\mathrm{Hom}_R(D, N)$, respectively. This defines a map from $\mathrm{Hom}_R(D, L \oplus N)$ to $\mathrm{Hom}_R(D, L) \oplus \mathrm{Hom}_R(D, N)$ which is easily seen to be a homomorphism. Conversely, given $f_1 \in \mathrm{Hom}_R(D, L)$ and $f_2 \in \mathrm{Hom}_R(D, N)$, define the map $f \in \mathrm{Hom}_R(D, L \oplus N)$ by $f(d) = (f_1(d), f_2(d))$. This defines a map from $\mathrm{Hom}_R(D, L) \oplus \mathrm{Hom}_R(D, N)$ to $\mathrm{Hom}_R(D, L \oplus N)$ that is easily checked to be a homomorphism inverse to the map above, proving the isomorphism in (1). The proof of (2) is similar and is left as an exercise.

The results in Proposition 29 extend immediately by induction to any finite direct sum of R-modules. These results are referred to by saying that Hom *commutes with finite direct sums in either variable* (compare to Theorem 17 for a corresponding result for tensor products). For infinite direct sums the situation is more complicated. Part (1) remains true if $L \oplus N$ is replaced by an arbitrary direct sum and the direct sum on the right hand side is replaced by a direct product (Exercise 13 shows that the direct product is necessary). Part (2) remains true if the direct sums on both sides are replaced by direct products.

This proposition shows that if the sequence

$$0 \longrightarrow L \xrightarrow{\psi} M \xrightarrow{\varphi} N \longrightarrow 0$$

is a split short exact sequence of R-modules, then

$$0 \longrightarrow \mathrm{Hom}_R(D, L) \xrightarrow{\psi'} \mathrm{Hom}_R(D, M) \xrightarrow{\varphi'} \mathrm{Hom}_R(D, N) \longrightarrow 0$$

is also a split short exact sequence of abelian groups for every R-module D. Exercise 14 shows that a converse holds: if $0 \to \mathrm{Hom}_R(D, L) \xrightarrow{\psi'} \mathrm{Hom}_R(D, M) \xrightarrow{\varphi'} \mathrm{Hom}_R(D, N) \to 0$ is exact for every R-module D then $0 \to L \xrightarrow{\psi} M \xrightarrow{\varphi} N \to 0$ is a split short exact sequence (which then implies that if the original Hom sequence is exact for every D, then in fact it is split exact for every D).

Proposition 29 identifies a situation in which the sequence (11) is exact in terms of the modules L, M, and N. The next result adopts a slightly different perspective, characterizing instead the modules D having the property that the sequence (10) in Theorem 28 can *always* be extended to a short exact sequence:

Proposition 30. Let P be an R-module. Then the following are equivalent:

(1) For any R-modules L, M, and N, if

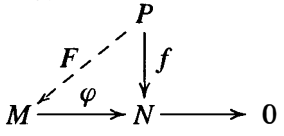
$$0 \longrightarrow L \xrightarrow{\psi} M \xrightarrow{\varphi} N \longrightarrow 0$$

is a short exact sequence, then

$$0 \longrightarrow \mathrm{Hom}_R(P, L) \xrightarrow{\psi'} \mathrm{Hom}_R(P, M) \xrightarrow{\varphi'} \mathrm{Hom}_R(P, N) \longrightarrow 0$$

is also a short exact sequence.

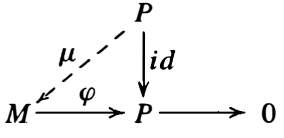
(2) For any R-modules M and N, if $M \xrightarrow{\varphi} N \to 0$ is exact, then every R-module homomorphism from P into N lifts to an R-module homomorphism into M, i.e., given $f \in \mathrm{Hom}_R(P, N)$ there is a lift $F \in \mathrm{Hom}_R(P, M)$ making the following diagram commute:



(3) If P is a quotient of the R-module M then P is isomorphic to a direct summand of M, i.e., every short exact sequence $0 \to L \to M \to P \to 0$ splits.

(4) P is a direct summand of a free R-module.

Proof: The equivalence of (1) and (2) is a restatement of a result in Theorem 28. Suppose now that (2) is satisfied, and let $0 \to L \xrightarrow{\psi} M \xrightarrow{\varphi} P \to 0$ be exact. By (2), the identity map from P to P lifts to a homomorphism μ making the following diagram commute:

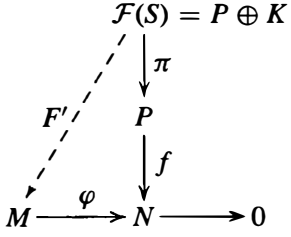


Then $\varphi \circ \mu = 1$, so μ is a splitting homomorphism for the sequence, which proves (3).

Every module P is the quotient of a free module (for example, the free module on the

set of elements in P), so there is always an exact sequence $0 \to \ker \varphi \to \mathcal{F} \xrightarrow{\varphi} P \to 0$ where $\mathcal{F}$ is a free R-module (cf. Example 4 following Corollary 23). If (3) is satisfied, then this sequence splits, so $\mathcal{F}$ is isomorphic to the direct sum of $\ker \varphi$ and P, which proves (4).

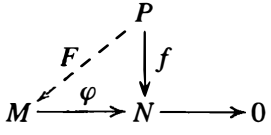
Finally, to prove (4) implies (2), suppose that P is a direct summand of a free R-module on some set S, say $\mathcal{F}(S) = P \oplus K$, and that we are given a homomorphism f from P to N as in (2). Let π denote the natural projection from $\mathcal{F}(S)$ to P, so that $f \circ \pi$ is a homomorphism from $\mathcal{F}(S)$ to N. For any $s \in S$ define $n_s = f \circ \pi(s) \in N$ and let $m_s \in M$ be any element of M with $\varphi(m_s) = n_s$ (which exists because φ is surjective). By the universal property for free modules (Theorem 6 of Section 3), there is a unique R-module homomorphism F' from $\mathcal{F}(S)$ to M with $F'(s) = m_s$. The diagram is the following:



By definition of the homomorphism F' we have $\varphi \circ F'(s) = \varphi(m_s) = n_s = f \circ \pi(s)$, from which it follows that $\varphi \circ F' = f \circ \pi$ on $\mathcal{F}(S)$, i.e., the diagram above is commutative. Now define a map $F : P \to M$ by $F(d) = F'((d, 0))$. Since F is the composite of the injection $P \to \mathcal{F}(S)$ with the homomorphism F', it follows that F is an R-module homomorphism. Then

$$\varphi \circ F(d) = \varphi \circ F'((d, 0)) = f \circ \pi((d, 0)) = f(d)$$

i.e., $\varphi \circ F = f$, so the diagram



commutes, which proves that (4) implies (2) and completes the proof.

Definition. An R-module P is called *projective* if it satisfies any of the equivalent conditions of Proposition 30.

The third statement in Proposition 30 can be rephrased as saying that any module M that projects onto P has (an isomorphic copy of) P as a direct summand, which explains the terminology.

The following result is immediate from Proposition 30 (and its proof):

Corollary 31. Free modules are projective. A finitely generated module is projective if and only if it is a direct summand of a finitely generated free module. Every module is a quotient of a projective module.

If D is fixed, then given any R-module X we have an associated abelian group $\mathrm{Hom}_R(D, X)$. Further, an R-module homomorphism $\alpha : X \to Y$ induces an abelian group homomorphism $\alpha' : \mathrm{Hom}_R(D, X) \to \mathrm{Hom}_R(D, Y)$, defined by $\alpha'(f) = \alpha \circ f$. Put another way, the map $\mathrm{Hom}_R(D, \_\_)$ is a *covariant functor* from the category of R-modules to the category of abelian groups (cf. Appendix II). Theorem 28 shows that applying this functor to the terms in the exact sequence

$$0 \longrightarrow L \xrightarrow{\psi} M \xrightarrow{\varphi} N \longrightarrow 0$$

produces an exact sequence

$$0 \to \mathrm{Hom}_R(D, L) \xrightarrow{\psi'} \mathrm{Hom}_R(D, M) \xrightarrow{\varphi'} \mathrm{Hom}_R(D, N).$$

This is referred to by saying that $\mathrm{Hom}_R(D, \_\_)$ is a *left exact* functor. By Proposition 30, the functor $\mathrm{Hom}_R(D, \_\_)$ is *exact*, i.e., always takes short exact sequences to short exact sequences, if and only if D is projective. We summarize this as

Corollary 32. If D is an R-module, then the functor $\mathrm{Hom}_R(D, \_\_)$ from the category of R-modules to the category of abelian groups is left exact. It is exact if and only if D is a projective R-module.

Note that if $\mathrm{Hom}_R(D, \_\_)$ takes short exact sequences to short exact sequences, then it takes exact sequences of any length to exact sequences since any exact sequence can be broken up into a succession of short exact sequences.

As we have seen, the functor $\mathrm{Hom}_R(D, \_\_)$ is in general not exact on the right. Measuring the extent to which functors such as $\mathrm{Hom}_R(D, \_\_)$ fail to be exact leads to the notions of "homological algebra," considered in Chapter 17.

Examples

(1) We shall see in Section 11.1 that if $R = F$ is a field then every F-module is projective (although we only prove this for finitely generated modules).

(2) By Corollary 31, $\mathbb{Z}$ is a projective $\mathbb{Z}$-module. This can be seen directly as follows: suppose f is a map from $\mathbb{Z}$ to N and $M \xrightarrow{\varphi} N \to 0$ is exact. The homomorphism f is uniquely determined by the value $n = f(1)$. Then f can be lifted to a homomorphism $F : \mathbb{Z} \to M$ by first defining $F(1) = m$, where m is any element in M mapped to n by φ, and then extending F to all of $\mathbb{Z}$ by additivity.

By the first statement in Proposition 30, since $\mathbb{Z}$ is projective, if

$$0 \longrightarrow L \xrightarrow{\psi} M \xrightarrow{\varphi} N \longrightarrow 0$$

is an exact sequence of $\mathbb{Z}$-modules, then

$$0 \longrightarrow \mathrm{Hom}_{\mathbb{Z}}(\mathbb{Z}, L) \xrightarrow{\psi'} \mathrm{Hom}_{\mathbb{Z}}(\mathbb{Z}, M) \xrightarrow{\varphi'} \mathrm{Hom}_{\mathbb{Z}}(\mathbb{Z}, N) \longrightarrow 0$$

is also an exact sequence. This can also be seen directly using the isomorphism $\mathrm{Hom}_{\mathbb{Z}}(\mathbb{Z}, M) \cong M$ of abelian groups, which shows that the two exact sequences above are essentially the same.

(3) Free $\mathbb{Z}$-modules have no nonzero elements of finite order so no nonzero finite abelian group can be isomorphic to a submodule of a free module. By Corollary 31 it follows that no nonzero finite abelian group is a projective $\mathbb{Z}$-module.

(4) As a particular case of the preceding example, we see that for $n \geq 2$ the $\mathbb{Z}$-module $\mathbb{Z}/n\mathbb{Z}$ is not projective. By Theorem 28 it must be possible to find a short exact sequence which after applying the functor $\mathrm{Hom}_{\mathbb{Z}}(\mathbb{Z}/n\mathbb{Z}, \_\_)$ is no longer exact on the right. One such sequence is the exact sequence of Example 2 following Corollary 23:

$$0 \longrightarrow \mathbb{Z} \xrightarrow{n} \mathbb{Z} \xrightarrow{\pi} \mathbb{Z}/n\mathbb{Z} \longrightarrow 0,$$

for $n \geq 2$. Note first that $\mathrm{Hom}_{\mathbb{Z}}(\mathbb{Z}/n\mathbb{Z}, \mathbb{Z}) = 0$ since there are no nonzero $\mathbb{Z}$-module homomorphisms from $\mathbb{Z}/n\mathbb{Z}$ to $\mathbb{Z}$. It is also easy to see that $\mathrm{Hom}_{\mathbb{Z}}(\mathbb{Z}/n\mathbb{Z}, \mathbb{Z}/n\mathbb{Z}) \cong \mathbb{Z}/n\mathbb{Z}$, as follows. Every homomorphism f is uniquely determined by $f(1) = a \in \mathbb{Z}/n\mathbb{Z}$, and given any $a \in \mathbb{Z}/n\mathbb{Z}$ there is a unique homomorphism f_a with $f_a(1) = a$; the map $f_a \mapsto a$ is easily checked to be an isomorphism from $\mathrm{Hom}_{\mathbb{Z}}(\mathbb{Z}/n\mathbb{Z}, \mathbb{Z}/n\mathbb{Z})$ to $\mathbb{Z}/n\mathbb{Z}$.

Applying $\mathrm{Hom}_{\mathbb{Z}}(\mathbb{Z}/n\mathbb{Z}, \_\_)$ to the short exact sequence above thus gives the sequence

$$0 \longrightarrow 0 \xrightarrow{n'} 0 \xrightarrow{\pi'} \mathbb{Z}/n\mathbb{Z} \longrightarrow 0$$

which is not exact at its only nonzero term.

(5) Since $\mathbb{Q}/\mathbb{Z}$ is a torsion $\mathbb{Z}$-module it is not a submodule of a free $\mathbb{Z}$-module, hence is not projective. Note also that the exact sequence $0 \to \mathbb{Z} \to \mathbb{Q} \xrightarrow{\pi} \mathbb{Q}/\mathbb{Z} \to 0$ does not split since $\mathbb{Q}$ contains no submodule isomorphic to $\mathbb{Q}/\mathbb{Z}$.

(6) The $\mathbb{Z}$-module $\mathbb{Q}$ is not projective (cf. the exercises).

(7) We shall see in Chapter 12 that a finitely generated $\mathbb{Z}$-module is projective if and only if it is free.

(8) Let R be the commutative ring $\mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$ under componentwise addition and multiplication. If P_1 and P_2 are the principal ideals generated by $(1, 0)$ and $(0, 1)$ respectively then $R = P_1 \oplus P_2$, hence both P_1 and P_2 are projective R-modules by Proposition 30. Neither P_1 nor P_2 is free, since any free module has order a multiple of four.

(9) The direct sum of two projective modules is again projective (cf. Exercise 3).

(10) We shall see in Part VI that if F is any field and $n \in \mathbb{Z}^+$ then the ring $R = M_n(F)$ of all $n \times n$ matrices with entries from F has the property that every R-module is projective. We shall also see that if G is a finite group of order n and $n \neq 0$ in the field F then the group ring FG also has the property that every module is projective.

Injective Modules and $Hom_R(\_\_, D)$

If $0 \longrightarrow L \xrightarrow{\psi} M \xrightarrow{\varphi} N \longrightarrow 0$ is a short exact sequence of R-modules then, instead of considering maps *from* an R-module D into L or N and the extent to which these determine maps from D into M, we can consider the "dual" question of maps from L or N *to* D. In this case, it is easy to dispose of the situation of a map from N to D: an R-module map from N to D immediately gives a map from M to D simply by composing with φ. It is easy to check that this defines an injective homomorphism of abelian groups

$$\varphi' : \mathrm{Hom}_R(N, D) \longrightarrow \mathrm{Hom}_R(M, D)$$
$$f \longmapsto f' = f \circ \varphi,$$

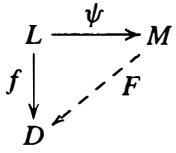
or, put another way,

$$\text{if} \quad M \xrightarrow{\varphi} N \to 0 \quad \text{is exact,}$$

$$\text{then} \quad 0 \to \mathrm{Hom}_R(N, D) \xrightarrow{\varphi'} \mathrm{Hom}_R(M, D) \quad \text{is exact.}$$

(Note that the associated maps on the homomorphism groups are in the reverse direction from the original maps.)

On the other hand, given an R-module homomorphism f from L to D it may not be possible to extend f to a map F from M to D, i.e., given f it may not be possible to find a map F making the following diagram commute:



For example, consider the exact sequence $0 \longrightarrow \mathbb{Z} \xrightarrow{\psi} \mathbb{Z} \xrightarrow{\varphi} \mathbb{Z}/2\mathbb{Z} \longrightarrow 0$ of $\mathbb{Z}$-modules, where ψ is multiplication by 2 and φ is the natural projection. Take $D = \mathbb{Z}/2\mathbb{Z}$ and let $f : \mathbb{Z} \to \mathbb{Z}/2\mathbb{Z}$ be reduction modulo 2 on the first $\mathbb{Z}$ in the sequence. There is only one nonzero homomorphism F from the second $\mathbb{Z}$ in the sequence to $\mathbb{Z}/2\mathbb{Z}$ (namely, reduction modulo 2), but this F does not lift the map f since $F \circ \psi(\mathbb{Z}) = F(2\mathbb{Z}) = 0$, so $F \circ \psi \neq f$.

Composition with ψ induces an abelian group homomorphism ψ' from $\mathrm{Hom}_R(M, D)$ to $\mathrm{Hom}_R(L, D)$, and in terms of the map ψ', the homomorphism $f \in \mathrm{Hom}_R(L, D)$ can be lifted to a homomorphism from M to D if and only if f is in the image of ψ'. The example above shows that

$$\text{if} \quad 0 \longrightarrow L \xrightarrow{\psi} M \quad \text{is exact,}$$

$$\text{then} \quad \mathrm{Hom}_R(M, D) \xrightarrow{\psi'} \mathrm{Hom}_R(L, D) \longrightarrow 0 \quad \text{is } \textit{not necessarily} \text{ exact.}$$

We can summarize these results in the following dual version of Theorem 28:

Theorem 33. Let D, L, M, and N be R-modules. If

$$0 \longrightarrow L \xrightarrow{\psi} M \xrightarrow{\varphi} N \longrightarrow 0 \quad \text{is exact,}$$

then the associated sequence

$$0 \to \mathrm{Hom}_R(N, D) \xrightarrow{\varphi'} \mathrm{Hom}_R(M, D) \xrightarrow{\psi'} \mathrm{Hom}_R(L, D) \quad \text{is exact.} \qquad (10.12)$$

A homomorphism $f : L \to D$ lifts to a homomorphism $F : M \to D$ if and only if $f \in \mathrm{Hom}_R(L, D)$ is in the image of ψ'. In general $\psi' : \mathrm{Hom}_R(M, D) \to \mathrm{Hom}_R(L, D)$ need not be surjective; the map ψ' is surjective if and only if every homomorphism from L to D lifts to a homomorphism from M to D, in which case the sequence (12) can be extended to a short exact sequence.

The sequence (12) is exact for *all* R-modules D if and only if the sequence

$$L \xrightarrow{\psi} M \xrightarrow{\varphi} N \to 0 \quad \text{is exact.}$$

Proof: The only item remaining to be proved in the first statement is the exactness of (12) at $\mathrm{Hom}_R(M, D)$. The proof of this statement is very similar to the proof of the corresponding result in Theorem 28 and is left as an exercise. Note also that the injectivity of ψ is not required, which proves the "if" portion of the final statement of the theorem.

Suppose now that the sequence (12) is exact for all R-modules D. We first show that $\varphi : M \to N$ is a surjection. Take $D = N/\varphi(M)$. If $\pi_1 : N \to N/\varphi(M)$ is the natural projection homomorphism, then $\pi_1 \circ \varphi(M) = 0$ by definition of π_1. Since $\pi_1 \circ \varphi = \varphi'(\pi_1)$, this means that the element $\pi_1 \in \mathrm{Hom}_R(N, N/\varphi(M))$ is mapped to 0 by φ'. Since φ' is assumed to be injective for all modules D, this means π_1 is the zero map, i.e., $N = \varphi(M)$ and so φ is a surjection. We next show that $\varphi \circ \psi = 0$, which will imply that image $\psi \subseteq \ker \varphi$. For this we take $D = N$ and observe that the identity map id_N on N is contained in $\mathrm{Hom}_R(N, N)$, hence $\varphi'(id_N) \in \mathrm{Hom}_R(M, N)$. Then the exactness of (12) for $D = N$ implies that $\varphi'(id_N) \in \ker \psi'$, so $\psi'(\varphi'(id_N)) = 0$. Then $id_N \circ \psi \circ \varphi = 0$, i.e., $\psi \circ \varphi = 0$, as claimed. Finally, we show that $\ker \varphi \subseteq$ image ψ. Let $D = M/\psi(L)$ and let $\pi_2 : M \to M/\psi(L)$ be the natural projection. Then $\psi'(\pi_2) = 0$ since $\pi_2(\psi(L)) = 0$ by definition of π_2. The exactness of (12) for this D then implies that π_2 is in the image of φ', say $\pi_2 = \varphi'(f)$ for some homomorphism $f \in \mathrm{Hom}_R(N, M/\psi(L))$, i.e., $\pi_2 = f \circ \varphi$. If $m \in \ker \varphi$ then $\pi_2(m) = f(\varphi(m)) = 0$, which means that $m \in \psi(L)$ since π_2 is just the projection from M into the quotient $M/\psi(L)$. Hence $\ker \varphi \subseteq$ image ψ, completing the proof.

By Theorem 33, the sequence

$$0 \longrightarrow \mathrm{Hom}_R(N, D) \xrightarrow{\varphi'} \mathrm{Hom}_R(M, D) \xrightarrow{\psi'} \mathrm{Hom}_R(L, D) \longrightarrow 0$$

is in general *not* a short exact sequence since ψ' need not be surjective, and the question of whether this sequence is exact precisely measures the extent to which homomorphisms from M to D are uniquely determined by pairs of homomorphisms from L and N to D.

The second statement in Proposition 29 shows that this sequence is exact when the original exact sequence $0 \to L \to M \to N \to 0$ is a *split* exact sequence. In fact in this case the sequence $0 \to \mathrm{Hom}_R(N, D) \xrightarrow{\varphi'} \mathrm{Hom}_R(M, D) \xrightarrow{\psi'} \mathrm{Hom}_R(L, D) \to 0$ is also a split exact sequence of abelian groups for every R-module D. Exercise 14 shows that a converse holds: if $0 \to \mathrm{Hom}_R(N, D) \xrightarrow{\varphi'} \mathrm{Hom}_R(M, D) \xrightarrow{\psi'} \mathrm{Hom}_R(L, D) \to 0$ is exact for every R-module D then $0 \to L \xrightarrow{\psi} M \xrightarrow{\varphi} N \to 0$ is a split short exact sequence (which then implies that if the Hom sequence is exact for every D, then in fact it is split exact for every D).

There is also a dual version of the first three parts of Proposition 30, which describes the R-modules D having the property that the sequence (12) in Theorem 33 can *always* be extended to a short exact sequence:

Proposition 34. Let Q be an R-module. Then the following are equivalent:

(1) For any R-modules L, M, and N, if

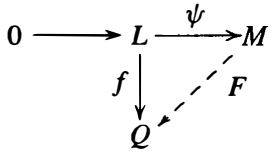
$$0 \longrightarrow L \xrightarrow{\psi} M \xrightarrow{\varphi} N \longrightarrow 0$$

is a short exact sequence, then

$$0 \longrightarrow \mathrm{Hom}_R(N, Q) \xrightarrow{\varphi'} \mathrm{Hom}_R(M, Q) \xrightarrow{\psi'} \mathrm{Hom}_R(L, Q) \longrightarrow 0$$

is also a short exact sequence.

(2) For any R-modules L and M, if $0 \to L \xrightarrow{\psi} M$ is exact, then every R-module homomorphism from L into Q lifts to an R-module homomorphism of M into Q, i.e., given $f \in \mathrm{Hom}_R(L, Q)$ there is a lift $F \in \mathrm{Hom}_R(M, Q)$ making the following diagram commute:



(3) If Q is a submodule of the R-module M then Q is a direct summand of M, i.e., every short exact sequence $0 \to Q \to M \to N \to 0$ splits.

Proof: The equivalence of (1) and (2) is part of Theorem 33. Suppose now that (2) is satisfied and let $0 \to Q \xrightarrow{\psi} M \xrightarrow{\varphi} N \to 0$ be exact. Taking $L = Q$ and f the identity map from Q to itself, it follows by (2) that there is a homomorphism $F : M \to Q$ with $F \circ \psi = 1$, so F is a splitting homomorphism for the sequence, which proves (3). The proof that (3) implies (2) is outlined in the exercises.

Definition. An R-module Q is called *injective* if it satisfies any of the equivalent conditions of Proposition 34.

The third statement in Proposition 34 can be rephrased as saying that any module M into which Q injects has (an isomorphic copy of) Q as a direct summand, which explains the terminology.

If D is fixed, then given any R-module X we have an associated abelian group $\mathrm{Hom}_R(X, D)$. Further, an R-module homomorphism $\alpha : X \to Y$ induces an abelian group homomorphism $\alpha' : \mathrm{Hom}_R(Y, D) \to \mathrm{Hom}_R(X, D)$, defined by $\alpha'(f) = f \circ \alpha$, that "reverses" the direction of the arrow. Put another way, the map $\mathrm{Hom}_R(D, \_\_)$ is a *contravariant functor* from the category of R-modules to the category of abelian groups (cf. Appendix II). Theorem 33 shows that applying this functor to the terms in the exact sequence

$$0 \longrightarrow L \xrightarrow{\psi} M \xrightarrow{\varphi} N \longrightarrow 0$$

produces an exact sequence

$$0 \to \mathrm{Hom}_R(N, D) \xrightarrow{\varphi'} \mathrm{Hom}_R(M, D) \xrightarrow{\psi'} \mathrm{Hom}_R(L, D).$$

This is referred to by saying that $\mathrm{Hom}_R(\_\_, D)$ is a *left exact* (contravariant) functor. Note that the functor $\mathrm{Hom}_R(\_\_, D)$ and the functor $\mathrm{Hom}_R(D, \_\_)$ considered earlier

are both left exact; the former reverses the directions of the maps in the original short exact sequence, the latter maintains the directions of the maps.

By Proposition 34, the functor $\mathrm{Hom}_R(\_, D)$ is *exact*, i.e., always takes short exact sequences to short exact sequences (and hence exact sequences of any length to exact sequences), if and only if D is injective. We summarize this in the following proposition, which is dual to the covariant result of Corollary 32.

Corollary 35. If D is an R-module, then the functor $\mathrm{Hom}_R(\_, D)$ from the category of R-modules to the category of abelian groups is left exact. It is exact if and only if D is an injective R-module.

We have seen that an R-module is projective if and only if it is a direct summand of a free R-module. Providing such a simple characterization of injective R-modules is not so easy. The next result gives a criterion for Q to be an injective R-module (a result due to Baer, who introduced the notion of injective modules around 1940), and using it we can give a characterization of injective modules when $R = \mathbb{Z}$ (or, more generally, when R is a P.I.D.). Recall that a $\mathbb{Z}$-module A (i.e., an abelian group, written additively) is said to be *divisible* if $A = nA$ for all nonzero integers n. For example, both $\mathbb{Q}$ and $\mathbb{Q}/\mathbb{Z}$ are divisible (cf. Exercises 18 and 19 in Section 2.4 and Exercise 15 in Section 3.1).

Proposition 36. Let Q be an R-module.

(1) *(Baer's Criterion)* The module Q is injective if and only if for every left ideal I of R any R-module homomorphism $g : I \to Q$ can be extended to an R-module homomorphism $G : R \to Q$.

(2) If R is a P.I.D. then Q is injective if and only if $rQ = Q$ for every nonzero $r \in R$. In particular, a $\mathbb{Z}$-module is injective if and only if it is divisible. When R is a P.I.D., quotient modules of injective R-modules are again injective.

Proof: If Q is injective and $g : I \to Q$ is an R-module homomorphism from the nonzero ideal I of R into Q, then g can be extended to an R-module homomorphism from R into Q by Proposition 34(2) applied to the exact sequence $0 \to I \to R$, which proves the "only if" portion of (1). Suppose conversely that every homomorphism $g : I \to Q$ can be lifted to a homomorphism $G : R \to Q$. To show that Q is injective we must show that if $0 \to L \to M$ is exact and $f : L \to Q$ is an R-module homomorphism then there is a lift $F : M \to Q$ extending f. If $\mathcal{S}$ is the collection (f', L') of lifts $f' : L' \to Q$ of f to a submodule L' of M containing L, then the ordering $(f', L') \leq (f'', L'')$ if $L' \subseteq L''$ and $f'' = f'$ on L' partially orders $\mathcal{S}$. Since $\mathcal{S} \neq \emptyset$, by Zorn's Lemma there is a maximal element (F, M') in $\mathcal{S}$. The map $F : M' \to Q$ is a lift of f and it suffices to show that $M' = M$. Suppose that there is some element $m \in M$ not contained in M' and let $I = \{r \in R \mid rm \in M'\}$. It is easy to check that I is a left ideal in R, and the map $g : I \to Q$ defined by $g(x) = F(xm)$ is an R-module homomorphism from I to Q. By hypothesis, there is a lift $G : R \to Q$ of g. Consider the submodule $M' + Rm$ of M, and define the map $F' : M' + Rm \to Q$ by $F'(m' + rm) = F(m') + G(r)$. If $m_1 + r_1 m = m_2 + r_2 m$ then $(r_1 - r_2)m = m_2 - m_1$

shows that $r_1 - r_2 \in I$, so that

$$G(r_1 - r_2) = g(r_1 - r_2) = F((r_1 - r_2)m) = F(m_2 - m_1),$$

and so $F(m_1) + G(r_1) = F(m_2) + G(r_2)$. Hence F' is well defined and it is then immediate that F' is an R-module homomorphism extending f to $M' + Rm$. This contradicts the maximality of M', so that $M' = M$, which completes the proof of (1).

To prove (2), suppose R is a P.I.D. Any nonzero ideal I of R is of the form $I = (r)$ for some nonzero element r of R. An R-module homomorphism $f : I \to Q$ is completely determined by the image $f(r) = q$ in Q. This homomorphism can be extended to a homomorphism $F : R \to Q$ if and only if there is an element q' in Q with $F(1) = q'$ satisfying $q = f(r) = F(r) = rq'$. It follows that Baer's criterion for Q is satisfied if and only if $rQ = Q$, which proves the first two statements in (2). The final statement follows since a quotient of a module Q with $rQ = Q$ for all $r \neq 0$ in R has the same property.

Examples

(1) Since $\mathbb{Z}$ is not divisible, $\mathbb{Z}$ is not an injective $\mathbb{Z}$-module. This also follows from the fact that the exact sequence $0 \longrightarrow \mathbb{Z} \xrightarrow{2} \mathbb{Z} \longrightarrow \mathbb{Z}/2\mathbb{Z} \longrightarrow 0$ corresponding to multiplication by 2 does not split.

(2) The rational numbers $\mathbb{Q}$ is an injective $\mathbb{Z}$-module.

(3) The quotient $\mathbb{Q}/\mathbb{Z}$ of the injective $\mathbb{Z}$-module $\mathbb{Q}$ is an injective $\mathbb{Z}$-module.

(4) It is immediate that a direct sum of divisible $\mathbb{Z}$-modules is again divisible, hence a direct sum of injective $\mathbb{Z}$-modules is again injective. For example, $\mathbb{Q} \oplus \mathbb{Q}/\mathbb{Z}$ is an injective $\mathbb{Z}$-module. (See also Exercise 4).

(5) We shall see in Chapter 12 that no nonzero finitely generated $\mathbb{Z}$-module is injective.

(6) Suppose that the ring R is an integral domain. An R-module A is said to be a *divisible* R-module if $rA = A$ for every nonzero $r \in R$. The proof of Proposition 36 shows that in this case an injective R-module is divisible.

(7) We shall see in Section 11.1 that if $R = F$ is a field then every F-module is injective.

(8) We shall see in Part VI that if F is any field and $n \in \mathbb{Z}^+$ then the ring $R = M_n(F)$ of all $n \times n$ matrices with entries from F has the property that every R-module is injective (and also projective). We shall also see that if G is a finite group of order n and $n \neq 0$ in the field F then the group ring FG also has the property that every module is injective (and also projective).

Corollary 37. Every $\mathbb{Z}$-module is a submodule of an injective $\mathbb{Z}$-module.

Proof: Let M be a $\mathbb{Z}$-module and let A be any set of $\mathbb{Z}$-module generators of M. Let $\mathcal{F} = F(A)$ be the free $\mathbb{Z}$-module on the set A. Then by Theorem 6 there is a surjective $\mathbb{Z}$-module homomorphism from $\mathcal{F}$ to M and if $\mathcal{K}$ denotes the kernel of this homomorphism then $\mathcal{K}$ is a $\mathbb{Z}$-submodule of $\mathcal{F}$ and we can identify $M = \mathcal{F}/\mathcal{K}$. Let $\mathcal{Q}$ be the free $\mathbb{Q}$-module on the set A. Then $\mathcal{Q}$ is a direct sum of a number of copies of $\mathbb{Q}$, so is a divisible, hence (by Proposition 36) injective, $\mathbb{Z}$-module containing $\mathcal{F}$. Then $\mathcal{K}$ is also a $\mathbb{Z}$-submodule of $\mathcal{Q}$, so the quotient $\mathcal{Q}/\mathcal{K}$ is injective, again by Proposition 36. Since $M = \mathcal{F}/\mathcal{K} \subseteq \mathcal{Q}/\mathcal{K}$, it follows that M is contained in an injective $\mathbb{Z}$-module.

Corollary 37 can be used to prove the following more general version valid for arbitrary R-modules. This theorem is the injective analogue of the results in Theorem 6 and Corollary 31 showing that every R-module is a quotient of a projective R-module.

Theorem 38. Let R be a ring with 1 and let M be an R-module. Then M is contained in an injective R-module.

Proof: A proof is outlined in Exercises 15 to 17.

It is possible to prove a sharper result than Theorem 38, namely that there is a *minimal* injective R-module H containing M in the sense that any injective map of M into an injective R-module Q factors through H. More precisely, if $M \subseteq Q$ for an injective R-module Q then there is an injection $\iota : H \hookrightarrow Q$ that restricts to the identity map on M; using ι to identify H as a subset of Q we have $M \subseteq H \subseteq Q$. (cf. Theorem 57.13 in *Representation Theory of Finite Groups and Associative Algebras* by C. Curtis and I. Reiner, John Wiley & Sons, 1966). This module H is called the *injective hull* or *injective envelope* of M. The universal property of the injective hull of M with respect to inclusions of M into injective R-modules should be compared to the universal property with respect to homomorphisms of M of the free module $F(A)$ on a set of generators A for M in Theorem 6. For example, the injective hull of $\mathbb{Z}$ is $\mathbb{Q}$, and the injective hull of any field is itself (cf. the exercises).

Flat Modules and $D \otimes_R \_\_$

We now consider the behavior of extensions $0 \longrightarrow L \xrightarrow{\psi} M \xrightarrow{\varphi} N \longrightarrow 0$ of R-modules with respect to tensor products.

Suppose that D is a *right* R-module. For any homomorphism $f : X \to Y$ of left R-modules we obtain a homomorphism $1 \otimes f : D \otimes_R X \to D \otimes_R Y$ of abelian groups (Theorem 13). If in addition D is an (S, R)-bimodule (for example, when $S = R$ is commutative and D is given the standard (R, R)-bimodule structure as in Section 4), then $1 \otimes f$ is a homomorphism of left S-modules. Put another way,

$$D \otimes_R \_\_ : X \longrightarrow D \otimes_R X$$

is a *covariant functor* from the category of left R-modules to the category of abelian groups (respectively, to the category of left S-modules when D is an (S, R)-bimodule), cf. Appendix II. In a similar way, if D is a left R-module then $\_\_ \otimes_R D$ is a covariant functor from the category of right R-modules to the category of abelian groups (respectively, to the category of right S-modules when D is an (R, S)-bimodule). Note that, unlike Hom, the tensor product is covariant in both variables, and we shall therefore concentrate on $D \otimes_R \_\_$, leaving as an exercise the minor alterations necessary for $\_\_ \otimes_R D$.

We have already seen examples where the map $1 \otimes \psi : D \otimes_R L \to D \otimes_R M$ induced by an injective map $\psi : L \hookrightarrow M$ is no longer injective (for example the injection $\mathbb{Z} \hookrightarrow \mathbb{Q}$ of $\mathbb{Z}$-modules induces the zero map from $\mathbb{Z}/2\mathbb{Z} \otimes_{\mathbb{Z}} \mathbb{Z} = \mathbb{Z}/2\mathbb{Z}$ to $\mathbb{Z}/2\mathbb{Z} \otimes_{\mathbb{Z}} \mathbb{Q} = 0$). On the other hand, suppose that $\varphi : M \to N$ is a surjective R-module homomorphism. The tensor product $D \otimes_R N$ is generated as an abelian group by the simple tensors $d \otimes n$ for $d \in D$ and $n \in N$. The surjectivity of φ implies that $n = \varphi(m)$ for some $m \in M$, and then $1 \otimes \varphi(d \otimes m) = d \otimes \varphi(m) = d \otimes n$ shows that $1 \otimes \varphi$ is a surjective homomorphism of abelian groups from $D \otimes_R M$ to $D \otimes_R N$. This proves most of the following theorem.

Theorem 39. Suppose that D is a right R-module and that L, M and N are left R-modules. If

$$0 \longrightarrow L \xrightarrow{\psi} M \xrightarrow{\varphi} N \longrightarrow 0 \quad \text{is exact,}$$

then the associated sequence of abelian groups

$$D \otimes_R L \xrightarrow{1\otimes\psi} D \otimes_R M \xrightarrow{1\otimes\varphi} D \otimes_R N \longrightarrow 0 \quad \text{is exact.} \tag{10.13}$$

If D is an (S, R)-bimodule then (13) is an exact sequence of left S-modules. In particular, if $S = R$ is a commutative ring, then (13) is an exact sequence of R-modules with respect to the standard R-module structures. The map $1 \otimes \varphi$ is not in general injective, i.e., the sequence (13) cannot in general be extended to a short exact sequence.

The sequence (13) is exact for *all* right R-modules D if and only if

$$L \xrightarrow{\psi} M \xrightarrow{\varphi} N \to 0 \quad \text{is exact.}$$

Proof: For the first statement it remains to prove the exactness of (13) at $D \otimes_R M$. Since $\varphi \circ \psi = 0$, we have

$$(1 \otimes \varphi)\left(\sum d_i \otimes \psi(l_i)\right) = \sum d_i \otimes (\varphi \circ \psi(l_i)) = 0$$

and it follows that $\text{image}(1\otimes\psi) \subseteq \ker(1\otimes\varphi)$. In particular, there is a natural projection $\pi : (D \otimes_R M)/\text{ image}(1 \otimes \psi) \to (D \otimes_R M)/\ker(1 \otimes \varphi) = D \otimes_R N$. The composite of the two projection homomorphisms

$$D \otimes_R M \to (D \otimes_R M)/\text{ image}(1 \otimes \psi) \xrightarrow{\pi} D \otimes_R N$$

is the quotient of $D \otimes_R M$ by $\ker(1 \otimes \varphi)$, so is just the map $1 \otimes \varphi$. We shall show that π is an isomorphism, which will show that the kernel of $1 \otimes \varphi$ is just the kernel of the first projection above, i.e., $\text{image}(1 \otimes \psi)$, giving the exactness of (13) at $D \otimes_R M$. To see that π is an isomorphism we define an inverse map. First define $\pi' : D \times N \to (D \otimes_R M)/\text{ image}(1 \otimes \psi)$ by $\pi'((d, n)) = d \otimes m$ for any $m \in M$ with $\varphi(m) = n$. Note that this is well defined: any other element $m' \in M$ mapping to n differs from m by an element in $\ker\varphi = \text{image}\,\psi$, i.e., $m' = m + \psi(l)$ for some $l \in L$, and $d \otimes \psi(l) \in \text{image}(1 \otimes \psi)$. It is easy to check that π' is a balanced map, so induces a homomorphism $\tilde{\pi} : D \times N \to (D \otimes_R M)/\text{ image}(1 \otimes \psi)$ with $\tilde{\pi}(d \otimes n) = d \otimes m$. Then $\tilde{\pi} \circ \pi(d \otimes m) = \tilde{\pi}(d \otimes \varphi(m)) = d \otimes m$ shows that $\tilde{\pi} \circ \pi = 1$. Similarly, $\pi \circ \tilde{\pi} = 1$, so that π and $\tilde{\pi}$ are inverse isomorphisms, completing the proof that (13) is exact. Note also that the injectivity of ψ was not required for the proof.

Finally, suppose (13) is exact for every right R-module D. In general, $R \otimes_R X \cong X$ for any left R-module X (Example 1 following Corollary 9). Taking $D = R$ the exactness of the sequence $L \xrightarrow{\psi} M \xrightarrow{\varphi} N \to 0$ follows.

By Theorem 39, the sequence

$$0 \longrightarrow D \otimes_R L \xrightarrow{1\otimes\psi} D \otimes_R M \xrightarrow{1\otimes\varphi} D \otimes_R N \longrightarrow 0$$

is not in general exact since $1 \otimes \psi$ need not be injective. If $0 \to L \xrightarrow{\psi} M \xrightarrow{\varphi} N \to 0$ is a *split* short exact sequence, however, then since tensor products commute with direct sums by Theorem 17, it follows that

$$0 \longrightarrow D \otimes_R L \xrightarrow{1\otimes\psi} D \otimes_R M \xrightarrow{1\otimes\varphi} D \otimes_R N \longrightarrow 0$$

is also a split short exact sequence.

The following result relating to modules D having the property that (13) can always be extended to a short exact sequence is immediate from Theorem 39:

Proposition 40. Let A be a right R-module. Then the following are equivalent:

(1) For any left R-modules L, M, and N, if

$$0 \longrightarrow L \xrightarrow{\psi} M \xrightarrow{\varphi} N \longrightarrow 0$$

is a short exact sequence, then

$$0 \longrightarrow A \otimes_R L \xrightarrow{1\otimes\psi} A \otimes_R M \xrightarrow{1\otimes\varphi} A \otimes_R N \longrightarrow 0$$

is also a short exact sequence.

(2) For any left R-modules L and M, if $0 \to L \xrightarrow{\psi} M$ is an exact sequence of left R-modules (i.e., $\psi : L \to M$ is injective) then $0 \to A \otimes_R L \xrightarrow{1\otimes\psi} A \otimes_R M$ is an exact sequence of abelian groups (i.e., $1 \otimes \psi : A \otimes_R L \to A \otimes_R M$ is injective).

Definition. A right R-module A is called *flat* if it satisfies either of the two equivalent conditions of Proposition 40.

For a fixed right R-module D, the first part of Theorem 39 is referred to by saying that the functor $D \otimes_R \_\_$ is *right exact.*

Corollary 41. If D is a right R-module, then the functor $D \otimes_R \_\_$ from the category of left R-modules to the category of abelian groups is right exact. If D is an (S, R)-bimodule (for example when $S = R$ is commutative and D is given the standard R-module structure), then $D \otimes_R \_\_$ is a right exact functor from the category of left R-modules to the category of left S-modules. The functor is exact if and only if D is a flat R-module.

We have already seen some flat modules:

Corollary 42. Free modules are flat; more generally, projective modules are flat.

Proof: To show that the free R-module F is flat it suffices to show that for any injective map $\psi : L \to M$ of R-modules L and M the induced map $1 \otimes \psi : F \otimes_R L \to F \otimes_R M$ is also injective. Suppose first that $F \cong R^n$ is a finitely generated free R-module. In this case $F \otimes_R L = R^n \otimes_R L \cong L^n$ since $R \otimes_R L \cong L$ and tensor products commute with direct sums. Similarly $F \otimes_R M \cong M^n$ and under these isomorphisms

the map $1 \otimes \psi : F \otimes_R L \to F \otimes_R M$ is just the natural map of L^n to M^n induced by the inclusion ψ in each component. In particular, $1 \otimes \psi$ is injective and it follows that any finitely generated free module is flat. Suppose now that F is an arbitrary free module and that the element $\sum f_i \otimes l_i \in F \otimes_R L$ is mapped to 0 by $1 \otimes \psi$. This means that the element $\sum (f_i, \psi(l_i))$ can be written as a sum of generators as in equation (6) in the previous section in the free group on $F \times M$. Since this sum of elements is finite, all of the first coordinates of the resulting equation lie in some finitely generated free submodule F' of F. Then this equation implies that $\sum f_i \otimes l_i \in F' \otimes_R L$ is mapped to 0 in $F' \otimes_R M$. Since F' is a finitely generated free module, the injectivity we proved above shows that $\sum f_i \otimes l_i$ is 0 in $F' \otimes_R L$ and so also in $F \otimes_R L$. It follows that $1 \otimes \psi$ is injective and hence that F is flat.

Suppose now that P is a projective module. Then P is a direct summand of a free module F (Proposition 30), say $F = P \oplus P'$. If $\psi : L \to M$ is injective then $1 \otimes \psi : F \otimes_R L \to F \otimes_R M$ is also injective by what we have already shown. Since $F = P \oplus P'$ and tensor products commute with direct sums, this shows that

$$1 \otimes \psi : (P \otimes_R L) \oplus (P' \otimes_R L) \to (P \otimes_R M) \oplus (P' \otimes_R M)$$

is injective. Hence $1 \otimes \psi : P \otimes_R L \to P \otimes_R M$ is injective, proving that P is flat.

Examples

(1) Since $\mathbb{Z}$ is a projective $\mathbb{Z}$-module it is flat. The example before Theorem 39 shows that $\mathbb{Z}/2\mathbb{Z}$ not a flat $\mathbb{Z}$-module.

(2) The $\mathbb{Z}$-module $\mathbb{Q}$ is a flat $\mathbb{Z}$-module, as follows. Suppose $\psi : L \to M$ is an injective map of $\mathbb{Z}$-modules. Every element of $\mathbb{Q} \otimes_{\mathbb{Z}} L$ can be written in the form $(1/d) \otimes l$ for some nonzero integer d and some $l \in L$ (Exercise 7 in Section 4). If $(1/d) \otimes l$ is in the kernel of $1 \otimes \psi$ then $(1/d) \otimes \psi(l)$ is 0 in $\mathbb{Q} \otimes_{\mathbb{Z}} M$. By Exercise 8 in Section 4 this means $c\psi(l) = 0$ in M for some nonzero integer c. Then $\psi(c \cdot l) = 0$, and the injectivity of ψ implies $c \cdot l = 0$ in L. But this implies that $(1/d) \otimes l = (1/cd) \otimes (c \cdot l) = 0$ in L, which shows that $1 \otimes \psi$ is injective.

(3) The $\mathbb{Z}$-module $\mathbb{Q}/\mathbb{Z}$ is injective (by Proposition 36), but is not flat: the injective map $\psi(z) = 2z$ from $\mathbb{Z}$ to $\mathbb{Z}$ does not remain injective after tensoring with $\mathbb{Q}/\mathbb{Z}$ ($1 \otimes \psi : \mathbb{Q}/\mathbb{Z} \otimes_{\mathbb{Z}} \mathbb{Z} \to \mathbb{Q}/\mathbb{Z} \otimes \mathbb{Z}$ has the nonzero element $(\frac{1}{2} + \mathbb{Z}) \otimes 1$ in its kernel — identifying $\mathbb{Q}/\mathbb{Z} = \mathbb{Q}/\mathbb{Z} \otimes_{\mathbb{Z}} \mathbb{Z}$ this is the statement that multiplication by 2 has the element 1/2 in its kernel).

(4) The direct sum of flat modules is flat (Exercise 5). In particular, $\mathbb{Q} \oplus \mathbb{Z}$ is flat. This module is neither projective nor injective (since $\mathbb{Q}$ is not projective by Exercise 8 and $\mathbb{Z}$ is not injective by Proposition 36 (cf. Exercises 3 and 4).

We close this section with an important relation between Hom and tensor products:

Theorem 43. *(Adjoint Associativity)* Let R and S be rings, let A be a right R-module, let B be an (R, S)-bimodule and let C be a right S-module. Then there is an isomorphism of abelian groups:

$$\mathrm{Hom}_S(A \otimes_R B, C) \cong \mathrm{Hom}_R(A, \mathrm{Hom}_S(B, C))$$

(the homomorphism groups are right module homomorphisms—note that $\mathrm{Hom}_S(B, C)$ has the structure of a right R-module, cf. the exercises). If $R = S$ is commutative this is an isomorphism of R-modules with the standard R-module structures.

Proof: Suppose $\varphi : A \otimes_R B \to C$ is a homomorphism. For any fixed $a \in A$ define the map $\Phi(a)$ from B to C by $\Phi(a)(b) = \varphi(a \otimes b)$. It is easy to check that $\Phi(a)$ is a homomorphism of right S-modules and that the map Φ from A to $\mathrm{Hom}_S(B, C)$ given by mapping a to $\Phi(a)$ is a homomorphism of right R-modules. Then $f(\varphi) = \Phi$ defines a group homomorphism from $\mathrm{Hom}_S(A \otimes_R B, C)$ to $\mathrm{Hom}_R(A, \mathrm{Hom}_S(B, C))$. Conversely, suppose $\Phi : A \to \mathrm{Hom}_S(B, C)$ is a homomorphism. The map from $A \times B$ to C defined by mapping (a, b) to $\Phi(a)(c)$ is an R-balanced map, so induces a homomorphism φ from $A \otimes_R B$ to C. Then $g(\Phi) = \varphi$ defines a group homomorphism inverse to f and gives the isomorphism in the theorem.

As a first application of Theorem 43 we give an alternate proof of the first result in Theorem 39 that the tensor product is right exact in the case where $S = R$ is a commutative ring. If $0 \longrightarrow L \longrightarrow M \longrightarrow N \longrightarrow 0$ is exact, then by Theorem 33 the sequence

$$0 \longrightarrow \mathrm{Hom}_R(N, E) \longrightarrow \mathrm{Hom}_R(M, E) \longrightarrow \mathrm{Hom}_R(L, E)$$

is exact for every R-module E. Then by Theorem 28, the sequence

$$0 \to \mathrm{Hom}_R(D, \mathrm{Hom}_R(N, E)) \to \mathrm{Hom}_R(D, \mathrm{Hom}_R(M, E)) \to \mathrm{Hom}_R(D, \mathrm{Hom}_R(L, E))$$

is exact for all D and all E. By adjoint associativity, this means the sequence

$$0 \longrightarrow \mathrm{Hom}_R(D \otimes_R N, E) \longrightarrow \mathrm{Hom}_R(D \otimes_R M, E) \longrightarrow \mathrm{Hom}_R(D \otimes_R L, E)$$

is exact for any D and all E. Then, by the second part of Theorem 33, it follows that the sequence

$$D \otimes_R L \longrightarrow D \otimes_R M \longrightarrow D \otimes_R N \longrightarrow 0$$

is exact for all D, which is the right exactness of the tensor product.

As a second application of Theorem 43 we prove that the tensor product of two projective modules over a commutative ring R is again projective (see also Exercise 9 for a more direct proof).

Corollary 44. If R is commutative then the tensor product of two projective R-modules is projective.

Proof: Let P_1 and P_2 be projective modules. Then by Corollary 32, $\mathrm{Hom}_R(P_2, \_\_)$ is an exact functor from the category of R-modules to the category of R-modules. Then the composition $\mathrm{Hom}_R(P_1, \mathrm{Hom}_R(P_2, \_\_))$ is an exact functor by the same corollary. By Theorem 43 this means that $\mathrm{Hom}_R(P_1 \otimes_R P_2, \_\_)$ is an exact functor on R-modules. It follows again from Corollary 32 that $P_1 \otimes_R P_2$ is projective.

Summary

Each of the functors $\mathrm{Hom}_R(A, \_\_)$, $\mathrm{Hom}_R(\_\_, A)$, and $A \otimes_R \_\_$, map left R-modules to abelian groups; the functor $\_\_ \otimes_R A$ maps right R-modules to abelian groups. When R is commutative all four functors map R-modules to R-modules.

(1) Let A be a left R-module. The functor $\mathrm{Hom}_R(A, \_\_)$ is covariant and left exact; the module A is projective if and only if $\mathrm{Hom}_R(A, \_\_)$ is exact (i.e., is also right exact).

(2) Let A be a left R-module. The functor $\mathrm{Hom}_R(\_\_, A)$ is contravariant and left exact; the module A is injective if and only if $\mathrm{Hom}_R(\_\_, A)$ is exact.

(3) Let A be a right R-module. The functor $A \otimes_R \_\_$ is covariant and right exact; the module A is flat if and only if $A \otimes_R \_\_$ is exact (i.e., is also left exact).

(4) Let A be a left R-module. The functor $\_\_ \otimes_R A$ is covariant and right exact; the module A is flat if and only if $\_\_ \otimes_R A$ is exact.

(5) Projective modules are flat. The $\mathbb{Z}$-module $\mathbb{Q}/\mathbb{Z}$ is injective but not flat. The $\mathbb{Z}$-module $\mathbb{Z} \oplus \mathbb{Q}$ is flat but neither projective nor injective.

EXERCISES

Let R be a ring with 1.

1. Suppose that

$$\begin{array}{ccccc} A & \xrightarrow{\psi} & B & \xrightarrow{\varphi} & C \\ \downarrow{\scriptstyle\alpha} & & \downarrow{\scriptstyle\beta} & & \downarrow{\scriptstyle\gamma} \\ A' & \xrightarrow{\psi'} & B' & \xrightarrow{\varphi'} & C' \end{array}$$

is a commutative diagram of groups and that the rows are exact. Prove that

(a) if φ and α are surjective, and β is injective then γ is injective. [If $c \in \ker \gamma$, show there is a $b \in B$ with $\varphi(b) = c$. Show that $\varphi'(\beta(b)) = 0$ and deduce that $\beta(b) = \psi'(a')$ for some $a' \in A'$. Show there is an $a \in A$ with $\alpha(a) = a'$ and that $\beta(\psi(a)) = \beta(b)$. Conclude that $b = \psi(a)$ and hence $c = \varphi(b) = 0$.]

(b) if ψ', α, and γ are injective, then β is injective,

(c) if φ, α, and γ are surjective, then β is surjective,

(d) if β is injective, α and γ are surjective, then γ is injective,

(e) if β is surjective, γ and ψ' are injective, then α is surjective.

2. Suppose that

$$\begin{array}{ccccccc} A & \longrightarrow & B & \longrightarrow & C & \longrightarrow & D \\ \downarrow{\scriptstyle\alpha} & & \downarrow{\scriptstyle\beta} & & \downarrow{\scriptstyle\gamma} & & \downarrow{\scriptstyle\delta} \\ A' & \longrightarrow & B' & \longrightarrow & C' & \longrightarrow & D' \end{array}$$

is a commutative diagram of groups, and that the rows are exact. Prove that

(a) if α is surjective, and β, δ are injective, then γ is injective.

(b) if δ is injective, and α, γ are surjective, then β is surjective.

3. Let P_1 and P_2 be R-modules. Prove that $P_1 \oplus P_2$ is a projective R-module if and only if both P_1 and P_2 are projective.

4. Let Q_1 and Q_2 be R-modules. Prove that $Q_1 \oplus Q_2$ is an injective R-module if and only if both Q_1 and Q_2 are injective.

5. Let A_1 and A_2 be R-modules. Prove that $A_1 \oplus A_2$ is a flat R-module if and only if both A_1 and A_2 are flat. More generally, prove that an arbitrary direct sum $\sum A_i$ of R-modules is flat if and only if each A_i is flat. [Use the fact that tensor product commutes with arbitrary direct sums.]

6. Prove that the following are equivalent for a ring R:

(i) Every R-module is projective.

(ii) Every R-module is injective.

7. Let A be a nonzero finite abelian group.
   (a) Prove that A is not a projective $\mathbb{Z}$-module.
   (b) Prove that A is not an injective $\mathbb{Z}$-module.

8. Let Q be a nonzero divisible $\mathbb{Z}$-module. Prove that Q is not a projective $\mathbb{Z}$-module. Deduce that the rational numbers $\mathbb{Q}$ is not a projective $\mathbb{Z}$-module. [Show first that if F is any free module then $\cap_{n=1}^{\infty} nF = 0$ (use a basis of F to prove this). Now suppose to the contrary that Q is projective and derive a contradiction from Proposition 30(4).]

9. Assume R is commutative with 1.
   (a) Prove that the tensor product of two free R-modules is free. [Use the fact that tensor products commute with direct sums.]
   (b) Use (a) to prove that the tensor product of two projective R-modules is projective.

10. Let R and S be rings with 1 and let M and N be left R-modules. Assume also that M is an (R, S)-bimodule.
   (a) For $s \in S$ and for $\varphi \in \mathrm{Hom}_R(M, N)$ define $(s\varphi) : M \to N$ by $(s\varphi)(m) = \varphi(ms)$. Prove that $s\varphi$ is a homomorphism of left R-modules, and that this action of S on $\mathrm{Hom}_R(M, N)$ makes it into a *left* S-module.
   (b) Let $S = R$ and let $M = R$ (considered as an (R, R)-bimodule by left and right ring multiplication on itself). For each $n \in N$ define $\varphi_n : R \to N$ by $\varphi_n(r) = rn$, i.e., φ_n is the unique R-module homomorphism mapping 1_R to n. Show that $\varphi_n \in \mathrm{Hom}_R(R, N)$. Use part (a) to show that the map $n \mapsto \varphi_n$ is an isomorphism of left R-modules: $N \cong \mathrm{Hom}_R(R, N)$.
   (c) Deduce that if N is a free (respectively, projective, injective, flat) left R-module, then $\mathrm{Hom}_R(R, N)$ is also a free (respectively, projective, injective, flat) left R-module.

11. Let R and S be rings with 1 and let M and N be left R-modules. Assume also that N is an (R, S)-bimodule.
   (a) For $s \in S$ and for $\varphi \in \mathrm{Hom}_R(M, N)$ define $(\varphi s) : M \to N$ by $(\varphi s)(m) = \varphi(m)s$. Prove that φs is a homomorphism of left R-modules, and that this action of S on $\mathrm{Hom}_R(M, N)$ makes it into a *right* S-module. Deduce that $\mathrm{Hom}_R(M, R)$ is a right R-module, for any R-module M—called the *dual module* to M.
   (b) Let $N = R$ be considered as an (R, R)-bimodule as usual. Under the action defined in part (a) show that the map $r \mapsto \varphi_r$ is an isomorphism of right R-modules: $\mathrm{Hom}_R(R, R) \cong R$, where φ_r is the homomorphism that maps 1_R to r. Deduce that if M is a finitely generated free left R-module, then $\mathrm{Hom}_R(M, R)$ is a free right R-module of the same rank. (cf. also Exercise 13.)
   (c) Show that if M is a finitely generated projective R-module then its dual module $\mathrm{Hom}_R(M, R)$ is also projective.

12. Let A be an R-module, let I be any nonempty index set and for each $i \in I$ let B_i be an R-module. Prove the following isomorphisms of abelian groups; when R is commutative prove also that these are R-module isomorphisms. (Arbitrary direct sums and direct products of modules are introduced in Exercise 20 of Section 3.)
   (a) $\mathrm{Hom}_R(\bigoplus_{i \in I} B_i, A) \cong \prod_{i \in I} \mathrm{Hom}_R(B_i, A)$
   (b) $\mathrm{Hom}_R(A, \prod_{i \in I} B_i) \cong \prod_{i \in I} \mathrm{Hom}_R(A, B_i)$.

13. (a) Show that the dual of the free $\mathbb{Z}$-module with countable basis is not free. [Use the preceding exercise and Exercise 24, Section 3.] (See also Exercise 5 in Section 11.3.)
   (b) Show that the dual of the free $\mathbb{Z}$-module with countable basis is also not projective. [You may use the fact that any submodule of a free $\mathbb{Z}$-module is free.]

14. Let $0 \longrightarrow L \xrightarrow{\psi} M \xrightarrow{\varphi} N \longrightarrow 0$ be a sequence of R-modules.

(a) Prove that the associated sequence

$$0 \longrightarrow \mathrm{Hom}_R(D, L) \xrightarrow{\psi'} \mathrm{Hom}_R(D, M) \xrightarrow{\varphi'} \mathrm{Hom}_R(D, N) \longrightarrow 0$$

is a short exact sequence of abelian groups for all R-modules D if and only if the original sequence is a split short exact sequence. [To show the sequence splits, take $D = N$ and show the lift of the identity map in $\mathrm{Hom}_R(N, N)$ to $\mathrm{Hom}_R(N, M)$ is a splitting homomorphism for φ.]

(b) Prove that the associated sequence

$$0 \longrightarrow \mathrm{Hom}_R(N, D) \xrightarrow{\varphi'} \mathrm{Hom}_R(M, D) \xrightarrow{\psi'} \mathrm{Hom}_R(L, D) \longrightarrow 0$$

is a short exact sequence of abelian groups for all R-modules D if and only if the original sequence is a split short exact sequence.

15. Let M be a left R-module where R is a ring with 1.

(a) Show that $\mathrm{Hom}_{\mathbb{Z}}(R, M)$ is a left R-module under the action $(r\varphi)(r') = \varphi(r'r)$ (see Exercise 10).

(b) Suppose that $0 \to A \xrightarrow{\psi} B$ is an exact sequence of R-modules. Prove that if every homomorphism f from A to M lifts to a homomorphism F from B to M with $f = F \circ \psi$, then every homomorphism f' from A to $\mathrm{Hom}_{\mathbb{Z}}(R, M)$ lifts to a homomorphism F' from B to $\mathrm{Hom}_{\mathbb{Z}}(R, M)$ with $f' = F' \circ \psi$. [Given f', show that $f(a) = f'(a)(1_R)$ defines a homomorphism of A to M. If F is the associated lift of f to B, show that $F'(b)(r) = F(rb)$ defines a homomorphism from B to $\mathrm{Hom}_{\mathbb{Z}}(R, M)$ that lifts f'.]

(c) Prove that if Q is an injective R-module then $\mathrm{Hom}_{\mathbb{Z}}(R, Q)$ is also an injective R-module.

16. This exercise proves Theorem 38 that every left R-module M is contained in an injective left R-module.

(a) Show that M is contained in an injective $\mathbb{Z}$-module Q. [M is a $\mathbb{Z}$-module—use Corollary 37.]

(b) Show that $\mathrm{Hom}_R(R, M) \subseteq \mathrm{Hom}_{\mathbb{Z}}(R, M) \subseteq \mathrm{Hom}_{\mathbb{Z}}(R, Q)$.

(c) Use the R-module isomorphism $M \cong \mathrm{Hom}_R(R, M)$ (Exercise 10) and the previous exercise to conclude that M is contained in an injective module.

17. This exercise completes the proof of Proposition 34. Suppose that Q is an R-module with the property that every short exact sequence $0 \to Q \to M_1 \to N \to 0$ splits and suppose that the sequence $0 \to L \xrightarrow{\psi} M$ is exact. Prove that every R-module homomorphism f from L to Q can be lifted to an R-module homomorphism F from M to Q with $f = F \circ \psi$. [By the previous exercise, Q is contained in an injective R-module. Use the splitting property together with Exercise 4 (noting that Exercise 4 can be proved using (2) in Proposition 34 as the definition of an injective module).]

18. Prove that the injective hull of the $\mathbb{Z}$-module $\mathbb{Z}$ is $\mathbb{Q}$. [Let H be the injective hull of $\mathbb{Z}$ and argue that $\mathbb{Q}$ contains an isomorphic copy of H. Use the divisibility of H to show $1/n \in H$ for all nonzero integers n, and deduce that $H = \mathbb{Q}$.]

19. If F is a field, prove that the injective hull of F is F.

20. Prove that the polynomial ring $R[x]$ in the indeterminate x over the commutative ring R is a flat R-module.

21. Let R and S be rings with 1 and suppose M is a right R-module, and N is an (R, S)-bimodule. If M is flat over R and N is flat as an S-module prove that $M \otimes_R N$ is flat as a right S-module.

22. Suppose that R is a commutative ring and that M and N are flat R-modules. Prove that $M \otimes_R N$ is a flat R-module. [Use the previous exercise.]

23. Prove that the (right) module $M \otimes_R S$ obtained by changing the base from the ring R to the ring S (by some homomorphism $f : R \to S$ with $f(1_R) = 1_S$, cf. Example 6 following Corollary 12 in Section 4) of the flat (right) R-module M is a flat S-module.

24. Prove that A is a flat R-module if and only if for any left R-modules L and M where L is *finitely generated*, then $\psi : L \to M$ injective implies that also $1 \otimes \psi : A \otimes_R L \to A \otimes_R M$ is injective. [Use the techniques in the proof of Corollary 42.]

25. *(A Flatness Criterion)* Parts (a)-(c) of this exercise prove that A is a flat R-module if and only if for every finitely generated ideal I of R, the map from $A \otimes_R I \to A \otimes_R R \cong A$ induced by the inclusion $I \subseteq R$ is again injective (or, equivalently, $A \otimes_R I \cong AI \subseteq A$).

(a) Prove that if A is flat then $A \otimes_R I \to A \otimes_R R$ is injective.

(b) If $A \otimes_R I \to A \otimes_R R$ is injective for every finitely generated ideal I, prove that $A \otimes_R I \to A \otimes_R R$ is injective for every ideal I. Show that if K is any submodule of a finitely generated free module F then $A \otimes_R K \to A \otimes_R F$ is injective. Show that the same is true for any free module F. [Cf. the proof of Corollary 42.]

(c) Under the assumption in (b), suppose L and M are R-modules and $L \xrightarrow{\psi} M$ is injective. Prove that $A \otimes_R L \xrightarrow{1 \otimes \psi} A \otimes_R M$ is injective and conclude that A is flat. [Write M as a quotient of the free module F, giving a short exact sequence

$$0 \longrightarrow K \longrightarrow F \xrightarrow{f} M \longrightarrow 0.$$

Show that if $J = f^{-1}(\psi(L))$ and $\iota : J \to F$ is the natural injection, then the diagram

$$\begin{array}{ccccccccc} 0 & \longrightarrow & K & \longrightarrow & J & \longrightarrow & L & \longrightarrow & 0 \\ & & \downarrow{\scriptstyle id} & & \downarrow{\scriptstyle \iota} & & \downarrow{\scriptstyle \psi} & & \\ 0 & \longrightarrow & K & \longrightarrow & F & \longrightarrow & M & \longrightarrow & 0 \end{array}$$

is commutative with exact rows. Show that the induced diagram

$$\begin{array}{ccccccccc} A \otimes_R K & \longrightarrow & A \otimes_R J & \longrightarrow & A \otimes_R L & \longrightarrow & 0 \\ \downarrow{\scriptstyle id} & & \downarrow{\scriptstyle 1 \otimes \iota} & & \downarrow{\scriptstyle 1 \otimes \psi} & & \\ A \otimes_R K & \longrightarrow & A \otimes_R F & \longrightarrow & A \otimes_R M & \longrightarrow & 0 \end{array}$$

is commutative with exact rows. Use (b) to show that $1 \otimes \iota$ is injective, then use Exercise 1 to conclude that $1 \otimes \psi$ is injective.]

(d) *(A Flatness Criterion for quotients)* Suppose $A = F/K$ where F is flat (e.g., if F is free) and K is an R-submodule of F. Prove that A is flat if and only if $FI \cap K = KI$ for every finitely generated ideal I of R. [Use (a) to prove $F \otimes_R I \cong FI$ and observe the image of $K \otimes_R I$ is KI; tensor the exact sequence $0 \to K \to F \to A \to 0$ with I to prove that $A \otimes_R I \cong FI/KI$, and apply the flatness criterion.]

26. Suppose R is a P.I.D. This exercise proves that A is a flat R-module if and only if A is torsion free R-module (i.e., if $a \in A$ is nonzero and $r \in R$, then $ra = 0$ implies $r = 0$).

(a) Suppose that A is flat and for fixed $r \in R$ consider the map $\psi_r : R \to R$ defined by multiplication by r: $\psi_r(x) = rx$. If r is nonzero show that ψ_r is an injection. Conclude from the flatness of A that the map from A to A defined by mapping a to ra is injective and that A is torsion free.

(b) Suppose that A is torsion free. If I is a nonzero ideal of R, then $I = rR$ for some nonzero $r \in R$. Show that the map ψ_r in (a) induces an isomorphism $R \cong I$ of

R-modules and that the composite $R \xrightarrow{\psi} I \xrightarrow{\iota} R$ of ψ_r with the inclusion $\iota : I \subseteq R$ is multiplication by r. Prove that the composite $A \otimes_R R \xrightarrow{1\otimes\psi_r} A \otimes_R I \xrightarrow{1\otimes\iota} A \otimes_R R$ corresponds to the map $a \mapsto ra$ under the identification $A \otimes_R R = A$ and that this composite is injective since A is torsion free. Show that $1 \otimes \psi_r$ is an isomorphism and deduce that $1 \otimes \iota$ is injective. Use the previous exercise to conclude that A is flat.

27. Let M, A and B be R-modules.

(a) Suppose $f : A \to M$ and $g : B \to M$ are R-module homomorphisms. Prove that $X = \{(a, b) \mid a \in A,\ b \in B \text{ with } f(a) = g(b)\}$ is an R-submodule of the direct sum $A \oplus B$ (called the *pullback* or *fiber product* of f and g) and that there is a commutative diagram

$$\begin{array}{ccc} X & \xrightarrow{\pi_2} & B \\ {\scriptstyle \pi_1}\downarrow & & \downarrow{\scriptstyle g} \\ A & \xrightarrow{f} & M \end{array}$$

where π_1 and π_2 are the natural projections onto the first and second components.

(b) Suppose $f' : M \to A$ and $g' : M \to B$ are R-module homomorphisms. Prove that the quotient Y of $A \oplus B$ by $\{(f'(m), -g'(m)) \mid m \in M\}$ is an R-module (called the *pushout* or *fiber sum* of f' and g') and that there is a commutative diagram

$$\begin{array}{ccc} M & \xrightarrow{g'} & B \\ {\scriptstyle f'}\downarrow & & \downarrow{\scriptstyle \pi_2'} \\ A & \xrightarrow{\pi_1'} & Y \end{array}$$

where π_1' and π_2' are the natural maps to the quotient induced by the maps into the first and second components.

28. **(a)** *(Schanuel's Lemma)* If $0 \to K \to P \xrightarrow{\varphi} M \to 0$ and $0 \to K' \to P' \xrightarrow{\varphi'} M \to 0$ are exact sequences of R-modules where P and P' are projective, prove $P \oplus K' \cong P' \oplus K$ as R-modules. [Show that there is an exact sequence $0 \to \ker \pi \to X \xrightarrow{\pi} P \to 0$ with $\ker \pi \cong K'$, where X is the fiber product of φ and φ' as in the previous exercise. Deduce that $X \cong P \oplus K'$. Show similarly that $X \cong P' \oplus K$.]

(b) If $0 \to M \to Q \xrightarrow{\psi} L \to 0$ and $0 \to M \to Q' \xrightarrow{\psi'} L' \to 0$ are exact sequences of R-modules where Q and Q' are injective, prove $Q \oplus L' \cong Q' \oplus L$ as R-modules.

The R-modules M and N are said to be *projectively equivalent* if $M \oplus P \cong N \oplus P'$ for some projective modules P, P'. Similarly, M and N are *injectively equivalent* if $M \oplus Q \cong N \oplus Q'$ for some injective modules Q, Q'. The previous exercise shows K and K' are projectively equivalent and L and L' are injectively equivalent.

CHAPTER 11

Vector Spaces

In this chapter we review the basic theory of finite dimensional vector spaces over an arbitrary field F (some infinite dimensional vector space theory is covered in the exercises). Since the proofs are identical to the corresponding arguments for real vector spaces our treatment is very terse. For the most part we include only those results which are used in other parts of the text so basic topics such as Gauss–Jordan elimination, row echelon forms, methods for finding bases of subspaces, elementary properties of matrices, etc., are not covered or are discussed in the exercises. The reader should therefore consider this chapter as a refresher in linear algebra and as a prelude to field theory and Galois theory. Characteristic polynomials and eigenvalues will be reviewed and treated in a larger context in the next chapter.

11.1 DEFINITIONS AND BASIC THEORY

The terminology for vector spaces is slightly different from that of modules, that is, when the ring R is a field there are different names for many of the properties of R-modules which we defined in the last chapter. The following is a dictionary of these new terms (many of which may already be familiar). The definition of each corresponding vector space property is the same (verbatim) as the module-theoretic definition with the only added assumption being that the ring R is a field (so these definitions are not repeated here).

Terminology for R any Ring	**Terminology for R a Field**
M is an R-module	M is a vector space over R
m is an element of M	m is a vector in M
α is a ring element	α is a scalar
N is a submodule of M	N is a subspace of M
M/N is a quotient module	M/N is a quotient space
M is a free module of rank n	M is a vector space of dimension n
M is a finitely generated module	M is a finite dimensional vector space
M is a nonzero cyclic module	M is a 1-dimensional vector space
$\varphi : M \to N$ is an R-module homomorphism	$\varphi : M \to N$ is a linear transformation
M and N are isomorphic as R-modules	M and N are isomorphic vector spaces
the subset A of M generates M	the subset A of M spans M
$M = RA$	each element of M is a linear combination of elements of A i.e., $M = \text{Span}(A)$

For the remainder of this chapter F is a field and V is a vector space over F.

One of the first results we shall prove about vector spaces is that they are free F-modules, that is, they have bases. Although our arguments treat only the case of finite dimensional spaces, the corresponding result for arbitrary vector spaces is proved in the exercises as an application of Zorn's Lemma. The reader may first wish to review the section in the previous chapter on free modules, especially their properties pertaining to homomorphisms.

Definition.

(1) A subset S of V is called a set of *linearly independent* vectors if an equation $\alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_n v_n = 0$ with $\alpha_1, \alpha_2, \ldots, \alpha_n \in F$ and $v_1, v_2, \ldots, v_n \in S$ implies $\alpha_1 = \alpha_2 = \cdots = \alpha_n = 0$.

(2) A *basis* of a vector space V is an ordered set of linearly independent vectors which span V. In particular two bases will be considered different even if one is simply a rearrangement of the other. This is sometimes referred to as an *ordered basis.*

Examples

(1) The space $V = F[x]$ of polynomials in the variable x with coefficients from the field F is in particular a vector space over F. The elements $1, x, x^2, \ldots$ are linearly independent by definition (i.e., a polynomial is 0 if and only if all its coefficients are 0). Since these elements also span V by definition, they are a basis for V.

(2) The collection of solutions of a linear, homogeneous, constant coefficient differential equation (for example, $y'' - 3y' + 2y = 0$) over $\mathbb{C}$ form a vector space over $\mathbb{C}$ since differentiation is a linear operator. Elements of this vector space are linearly independent if they are linearly independent as functions. For example, e^t and e^{2t} are easily seen to be solutions of the equation $y'' - 3y' + 2y = 0$ (differentiation with respect to t). They are linearly independent functions since $ae^t + be^{2t} = 0$ implies $a + b = 0$ (let $t = 0$) and $ae + be^2 = 0$ (let $t = 1$) and the only solution to these two equations is $a = b = 0$. It is a theorem in differential equations that these elements span the set of solutions of this equation, hence are a basis for this space.

Proposition 1. Assume the set $\mathcal{A} = \{v_1, v_2, \ldots, v_n\}$ spans the vector space V but no proper subset of $\mathcal{A}$ spans V. Then $\mathcal{A}$ is a basis of V. In particular, any finitely generated (i.e., finitely spanned) vector space over F is a free F-module.

Proof: It is only necessary to prove that $v_1, v_2, \ldots, v_n$ are linearly independent. Suppose $\alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_n v_n = 0$ where not all of the α_i are 0. By reordering, we may assume that $\alpha_1 \neq 0$ and then

$$v_1 = -\frac{1}{\alpha_1}(\alpha_2 v_2 + \cdots + \alpha_n v_n).$$

It follows that $\{v_2, v_3, \ldots, v_n\}$ also spans V since any linear combination of $v_1, v_2, \ldots, v_n$ can be written as a linear combination of $v_2, v_3, \ldots, v_n$ using the equation above. This is a contradiction.

Example

Let F be a field and consider $F[x]/(f(x))$ where $f(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0$. The ideal $(f(x))$ is a subspace of the vector space $F[x]$ and the quotient $F[x]/(f(x))$ is also a vector space over F. By the Euclidean Algorithm, every polynomial $a(x) \in F[x]$ can be written uniquely in the form $a(x) = q(x)f(x) + r(x)$ where $r(x) \in F[x]$ and $0 \leq \deg r(x) \leq n - 1$. Since $q(x)f(x) \in (f(x))$, it follows that every element of the quotient is represented by a polynomial $r(x)$ of degree $\leq n - 1$. Two distinct such polynomials cannot be the same in the quotient since this would say their difference (which is a nonzero polynomial of degree at most $n - 1$) would be divisible by $f(x)$ (which is of degree n). It follows that the elements $\bar{1}, \bar{x}, \overline{x^2}, \ldots, \overline{x^{n-1}}$ (the bar denotes the image of these elements in the quotient, as usual) span $F[x]/(f(x))$ as a vector space over F and that no proper subset of these elements also spans, hence these elements give a basis for $F[x]/(f(x))$.

Corollary 2. Assume the finite set $\mathcal{A}$ spans the vector space V. Then $\mathcal{A}$ contains a basis of V.

Proof: Any subset $\mathcal{B}$ of $\mathcal{A}$ spanning V such that no proper subset of $\mathcal{B}$ also spans V (there clearly exist such subsets) is a basis for V by Proposition 1.

Theorem 3. (*A Replacement Theorem*) Assume $\mathcal{A} = \{a_1, a_2, \ldots, a_n\}$ is a basis for V containing n elements and $\{b_1, b_2, \ldots, b_m\}$ is a set of linearly independent vectors in V. Then there is an ordering $a_1, a_2, \ldots, a_n$ such that for each $k \in \{1, 2, \ldots, m\}$ the set $\{b_1, b_2, \ldots, b_k, a_{k+1}, a_{k+2}, \ldots, a_n\}$ is a basis of V. In other words, the elements $b_1, b_2, \ldots, b_m$ can be used to successively replace the elements of the basis $\mathcal{A}$, still retaining a basis. In particular, $n \geq m$.

Proof: Proceed by induction on k. If $k = 0$ there is nothing to prove, since $\mathcal{A}$ is given as a basis for V. Suppose now that $\{b_1, b_2, \ldots, b_k, a_{k+1}, a_{k+2}, \ldots, a_n\}$ is a basis for V. Then in particular this is a spanning set, so b_{k+1} is a linear combination:

$$b_{k+1} = \beta_1 b_1 + \cdots + \beta_k b_k + \alpha_{k+1}a_{k+1} + \cdots + \alpha_n a_n. \tag{11.1}$$

Not all of the α_i can be 0, since this would imply b_{k+1} is a linear combination of $b_1, b_2, \ldots, b_k$, contrary to the linear independence of these elements. By reordering if necessary, we may assume $\alpha_{k+1} \neq 0$. Then solving this last equation for a_{k+1} as a linear combination of b_{k+1} and $b_1, b_2, \ldots, b_k, a_{k+2}, \ldots, a_n$ shows

$$\text{Span}\{b_1, b_2, \ldots, b_k, b_{k+1}, a_{k+2}, \ldots, a_n\} = \text{Span}\{b_1, b_2, \ldots, b_k, a_{k+1}, a_{k+2}, \ldots, a_n\}$$

and so this is a spanning set for V. It remains to show $b_1, \ldots, b_k, b_{k+1}, a_{k+2}, \ldots, a_n$ are linearly independent. If

$$\beta_1 b_1 + \cdots + \beta_k b_k + \beta_{k+1}b_{k+1} + \alpha_{k+2}a_{k+2} + \cdots + \alpha_n a_n = 0 \tag{11.2}$$

then substituting for b_{k+1} from the expression for b_{k+1} in equation (1), we obtain a linear combination of $\{b_1, b_2, \ldots, b_k, a_{k+1}, a_{k+2}, \ldots, a_n\}$ equal to 0, where the coefficient of a_{k+1} is β_{k+1}. Since this last set is a basis by induction, all the coefficients in this linear combination, in particular β_{k+1}, must be 0. But then equation (2) is

$$\beta_1 b_1 + \cdots + \beta_k b_k + \alpha_{k+2}a_{k+2} + \cdots + \alpha_n a_n = 0.$$

Again by the induction hypothesis all the other coefficients must be 0 as well. **Thus** $\{b_1, b_2, \ldots, b_k, b_{k+1}, a_{k+2}, \ldots, a_n\}$ is a basis for V, and the induction is complete.

Corollary 4.

(1) Suppose V has a finite basis with n elements. Any set of linearly independent vectors has $\leq n$ elements. Any spanning set has $\geq n$ elements.

(2) If V has some finite basis then any two bases of V have the same cardinality.

Proof: (1) This is a restatement of the last result of Theorem 3 and Corollary 2. (2) This is immediate from (1) since a basis is both a spanning set and a linearly independent set.

Definition. If V is a finitely generated F-module (i.e., has a finite basis) the cardinality of any basis is called the *dimension* of V and is denoted by $\dim_F V$, or just $\dim V$ when F is clear from the context, and V is said to be *finite dimensional* over F. If V is not finitely generated, V is said to be infinite dimensional (written $\dim V = \infty$).

Examples

(1) The dimension of the space of solutions to the differential equation $y'' - 3y' + 2y = 0$ over $\mathbb{C}$ is 2 (with basis e^t, e^{2t}, for example). In general, it is a theorem in differential equations that the space of solutions of an n^{th} order linear, homogeneous, constant coefficient differential equation of degree n over $\mathbb{C}$ form a vector space over $\mathbb{C}$ of dimension n.

(2) The dimension over F of the quotient $F[x]/(f(x))$ by the nonzero polynomial $f(x)$ considered above is $n = \deg f(x)$. The space $F[x]$ and its subspace $(f(x))$ are infinite dimensional vector spaces over F.

Corollary 5. *(Building–Up Lemma)* If A is a set of linearly independent vectors in the finite dimensional space V then there exists a basis of V containing A.

Proof: This is also immediate from Theorem 3, since we can use the elements of A to successively replace the elements of any given basis for V (which exists by the assumption that V is finite dimensional).

Theorem 6. If V is an n dimensional vector space over F, then $V \cong F^n$. In particular, any two finite dimensional vector spaces over F of the same dimension are isomorphic.

Proof: Let $v_1, v_2, \ldots, v_n$ be a basis for V. Define the map

$$\varphi : F^n \to V \qquad \text{by} \qquad \varphi(\alpha_1, \alpha_2, \ldots, \alpha_n) = \alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_n v_n.$$

The map φ is clearly F-linear, is surjective since the v_i span V, and is injective since the v_i are linearly independent, hence is an isomorphism.

Examples

(1) Let $\mathbb{F}$ be a finite field with q elements and let W be a k-dimensional vector space over $\mathbb{F}$. We show that the number of distinct bases of W is

$$(q^k - 1)(q^k - q)(q^k - q^2) \dots (q^k - q^{k-1}).$$

Every basis of W can be built up as follows. Any nonzero vector w_1 can be the first element of a basis. Since W is isomorphic to $\mathbb{F}^k$, $|W| = q^k$, so there are $q^k - 1$ choices for w_1. Any vector not in the 1-dimensional space spanned by w_1 is linearly independent from w_1 and so may be chosen for the second basis element, w_2. A 1-dimensional space is isomorphic to $\mathbb{F}$ and so has q elements. Thus there are $q^k - q$ choices for w_2. Proceeding in this way one sees that at the i^{th} stage any vector not in the $(i-1)$-dimensional space spanned by $w_1, w_2, \dots, w_{i-1}$ will be linearly independent from $w_1, w_2, \dots, w_{i-1}$ and so may be chosen for the i^{th} basis vector w_i. An $(i-1)$-dimensional space is isomorphic to $\mathbb{F}^{i-1}$ and so has q^{i-1} elements. Thus there are $q^k - q^{i-1}$ choices for w_i. The process terminates when w_k is chosen, for then we have k linear independent vectors in a k-dimensional space, hence a basis.

(2) Let $\mathbb{F}$ be a finite field with q elements and let V be an n-dimensional vector space over $\mathbb{F}$. For each $k \in \{1, 2, \dots, n\}$ we show that the number of subspaces of V of dimension k is

$$\frac{(q^n - 1)(q^n - q) \dots (q^n - q^{k-1})}{(q^k - 1)(q^k - q) \dots (q^k - q^{k-1})}.$$

Any k-dimensional space is spanned by k independent vectors. By arguing as in the preceding example the numerator of the above expression is the number of ways of picking k independent vectors from an n-dimensional space. Two sets of k independent vectors span the same space W if and only if they are both bases of the k-dimensional space W. In order to obtain the formula for the number of distinct subspaces of dimension k we must divide by the number of repetitions, i.e., the number of bases of a fixed k-dimensional space. This factor which appears in the denominator is precisely the number computed in Example 1.

Next, we prove an important relation between the dimension of a subspace, the dimension of its associated quotient space and the dimension of the whole space:

Theorem 7. Let V be a vector space over F and let W be a subspace of V. Then V/W is a vector space with $\dim V = \dim W + \dim V/W$ (where if one side is infinite then both are).

Proof: Suppose W has dimension m and V has dimension n over F and let $w_1, w_2, \dots, w_m$ be a basis for W. By Corollary 5, these linearly independent elements of V can be extended to a basis $w_1, w_2, \dots, w_m, v_{m+1}, \dots, v_n$ of V. The natural surjective projection map of V into V/W maps each w_i to 0. No linear combination of the v_i is mapped to 0, since this would imply this linear combination is an element of W, contrary to the choice of the v_i. Hence, the image V/W of this projection map is isomorphic to the subspace of V spanned by the v_i, hence $\dim V/W = n - m$, which is the theorem when the dimensions are finite. If either side is infinite it is an easy exercise to produce an infinite number of linearly independent vectors showing the other side is also infinite.

Corollary 8. Let $\varphi : V \to U$ be a linear transformation of vector spaces over F. Then $\ker \varphi$ is a subspace of V, $\varphi(V)$ is a subspace of U and $\dim V = \dim \ker \varphi + \dim \varphi(V)$.

Proof: This follows immediately from Theorem 7. Note that the proof of Theorem 7 is in fact the special case of Corollary 8 where U is the quotient V/W and φ is the natural projection homomorphism.

Corollary 9. Let $\varphi : V \to W$ be a linear transformation of vector spaces of the same finite dimension. Then the following are equivalent:

(1) φ is an isomorphism
(2) φ is injective, i.e., $\ker \varphi = 0$
(3) φ is surjective, i.e., $\varphi(V) = W$
(4) φ sends a basis of V to a basis of W.

Proof: The equivalence of these conditions follows from Corollary 8 by counting dimensions.

Definition. If $\varphi : V \to U$ is a linear transformation of vector spaces over F, $\ker \varphi$ is sometimes called the *null space* of φ and the dimension of $\ker \varphi$ is called the *nullity* of φ. The dimension of $\varphi(V)$ is called the *rank* of φ. If $\ker \varphi = 0$, the transformation is said to be *nonsingular*.

Example

Let F be a finite field with q elements and let V be an n-dimensional vector space over F. Recall that the *general linear group* $GL(V)$ is the group of all nonsingular linear transformations from V to V (the group operation being composition). We show that the order of this group is

$$|GL(V)| = (q^n - 1)(q^n - q)(q^n - q^2) \ldots (q^n - q^{n-1}).$$

To see this, fix a basis $v_1, \ldots, v_n$ of V. A linear transformation is nonsingular if and only if it sends this basis to another basis of V. Moreover, if $w_1 \ldots, w_n$ is any basis of V, by Theorem 6 in Section 10.3 there is a unique linear transformation which sends v_i to w_i, $1 \leq i \leq n$. Thus the number of nonsingular linear transformations from V to itself equals the number of distinct bases of V. This number, which was computed in Example 1 above (with $k = n$), is the order of $GL(V)$.

EXERCISES

1. Let $V = \mathbb{R}^n$ and let $(a_1, a_2, \ldots, a_n)$ be a fixed vector in V. Prove that the collection of elements $(x_1, x_2, \ldots, x_n)$ of V with $a_1x_1 + a_2x_2 + \ldots + a_nx_n = 0$ is a subspace of V. Determine the dimension of this subspace and find a basis.
2. Let V be the collection of polynomials with coefficients in $\mathbb{Q}$ in the variable x of degree at most 5. Prove that V is a vector space over $\mathbb{Q}$ of dimension 6, with $1, x, x^2, \ldots, x^5$ as basis. Prove that $1, 1+x, 1+x+x^2, \ldots, 1+x+x^2+x^3+x^4+x^5$ is also a basis for V.

3. Let φ be the linear transformation $\varphi : \mathbb{R}^4 \to \mathbb{R}^1$ such that

$$\varphi((1,0,0,0)) = 1 \qquad \varphi((1,-1,0,0)) = 0$$
$$\varphi((1,-1,1,0)) = 1 \qquad \varphi((1,-1,1,-1)) = 0.$$

Determine $\varphi((a,b,c,d))$.

4. Prove that the space of real-valued functions on the closed interval $[a,b]$ is an infinite dimensional vector space over $\mathbb{R}$, where $a < b$.

5. Prove that the space of continuous real-valued functions on the closed interval $[a,b]$ is an infinite dimensional vector space over $\mathbb{R}$, where $a < b$.

6. Let V be a vector space of finite dimension. If φ is any linear transformation from V to V prove there is an integer m such that the intersection of the image of φ^m and the kernel of φ^m is $\{0\}$.

7. Let φ be a linear transformation from a vector space V of dimension n to itself that satisfies $\varphi^2 = 0$. Prove that the image of φ is contained in the kernel of φ and hence that the rank of φ is at most $n/2$.

8. Let V be a vector space over F and let φ be a linear transformation of the vector space V to itself. A nonzero element $v \in V$ satisfying $\varphi(v) = \lambda v$ for some $\lambda \in F$ is called an *eigenvector* of φ with *eigenvalue* λ. Prove that for any fixed $\lambda \in F$ the collection of eigenvectors of φ with eigenvalue λ together with 0 forms a subspace of V.

9. Let V be a vector space over F and let φ be a linear transformation of the vector space V to itself. Suppose for $i = 1, 2, \ldots, k$ that $v_i \in V$ is an eigenvector for φ with eigenvalue $\lambda_i \in F$ (cf. the preceding exercise) and that all the eigenvalues λ_i are distinct. Prove that $v_1, v_2, \ldots, v_k$ are linearly independent. [Use induction on k: write a linear dependence relation among the v_i and apply φ to get another linear dependence relation among the v_i involving the eigenvalues — now subtract a suitable multiple of the first linear relation to get a linear dependence relation on fewer elements.] Conclude that any linear transformation on an n-dimensional vector space has at most n distinct eigenvalues.

In the following exercises let V be a vector space of arbitrary dimension over a field F.

10. Prove that any vector space V has a basis (by convention the null set is the basis for the zero space). [Let $\mathcal{S}$ be the set of subsets of V consisting of linearly independent vectors, partially ordered under inclusion; apply Zorn's Lemma to $\mathcal{S}$ and show a maximal element of $\mathcal{S}$ is a basis.]

11. Refine your argument in the preceding exercise to prove that any set of linearly independent vectors of V is contained in a basis of V.

12. If F is a field with a finite or countable number of elements and V is an infinite dimensional vector space over F with basis $\mathcal{B}$, prove that the cardinality of V equals the cardinality of $\mathcal{B}$. Deduce in this case that any two bases of V have the same cardinality.

13. Prove that as vector spaces over $\mathbb{Q}$, $\mathbb{R}^n \cong \mathbb{R}$, for all $n \in \mathbb{Z}^+$ (note that, in particular, this means $\mathbb{R}^n$ and $\mathbb{R}$ are isomorphic as additive abelian groups).

14. Let $\mathcal{A}$ be a basis for the infinite dimensional space V. Prove that V is isomorphic to the *direct sum* of copies of the field F indexed by the set $\mathcal{A}$. Prove that the *direct product* of copies of F indexed by $\mathcal{A}$ is a vector space over F and it has strictly larger dimension than the dimension of V (see the exercises in Section 10.3 for the definitions of direct sum and direct product of infinitely many modules).

Throughout this section let V, W be vector spaces over the same field F, let $\mathcal{B} = \{v_1, v_2, \ldots, v_n\}$ be an (ordered) basis of V, let $\mathcal{E} = \{w_1, w_2, \ldots, w_m\}$ be an (ordered) basis of W and let $\varphi \in \mathrm{Hom}(V, W)$ be a linear transformation from V to W. For each $j \in \{1, 2, \ldots, n\}$ write the image of v_j under φ in terms of the basis $\mathcal{E}$:

$$\varphi(v_j) = \sum_{i=1}^{m} \alpha_{ij} w_i. \tag{11.3}$$

Let $M_{\mathcal{B}}^{\mathcal{E}}(\varphi) = (a_{ij})$ be the $m \times n$ matrix whose i, j entry is α_{ij} (that is, use the coefficients of the w_i's in the above computation of $\varphi(v_j)$ for the j^{th} column of this matrix). The matrix $M_{\mathcal{B}}^{\mathcal{E}}(\varphi)$ is called the *matrix of φ with respect to the bases $\mathcal{B}$, $\mathcal{E}$*. The domain basis is the lower and the codomain basis the upper letters appearing after the "M." Given this matrix, we can recover the linear transformation φ as follows: to compute $\varphi(v)$ for $v \in V$, write v in terms of the basis $\mathcal{B}$:

$$v = \sum_{i=1}^{n} \alpha_i v_i, \qquad \alpha_i \in F,$$

and then calculate the product of the $m \times n$ and $n \times 1$ matrices

$$M_{\mathcal{B}}^{\mathcal{E}}(\varphi) \times \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{pmatrix}.$$

The image of v under φ is given by

$$\varphi(v) = \sum_{i=1}^{m} \beta_i w_i \,,$$

i.e., the column vector of coordinates of $\varphi(v)$ with respect to the basis $\mathcal{E}$ are obtained by multiplying the matrix $M_{\mathcal{B}}^{\mathcal{E}}(\varphi)$ by the column vector of coordinates of v with respect to the basis $\mathcal{B}$ (sometimes denoted $[\varphi(v)]_{\mathcal{E}} = M_{\mathcal{B}}^{\mathcal{E}}(\varphi)[v]_{\mathcal{B}}$).

Definition. The $m \times n$ matrix $A = (a_{ij})$ associated to the linear transformation φ above is said to *represent* the linear transformation φ with respect to the bases $\mathcal{B}$, $\mathcal{E}$. Similarly, φ is the linear transformation represented by A with respect to the bases $\mathcal{B}$, $\mathcal{E}$.

Examples

(1) Let $V = \mathbb{R}^3$ with the standard basis $\mathcal{B} = \{(1, 0, 0), (0, 1, 0), (0, 0, 1)\}$ and let $W = \mathbb{R}^2$ with the standard basis $\mathcal{E} = \{(1, 0), (0, 1)\}$. Let φ be the linear transformation $\varphi(x, y, z) = (x + 2y, x + y + z)$. Since $\varphi(1, 0, 0) = (1, 1)$, $\varphi(0, 1, 0) = (2, 1)$, $\varphi(0, 0, 1) = (0, 1)$, the matrix $A = M_{\mathcal{B}}^{\mathcal{E}}(\varphi)$ is the matrix $\begin{pmatrix} 1 & 2 & 0 \\ 1 & 1 & 1 \end{pmatrix}$.

(2) Let $V = W$ be the 2-dimensional space of solutions of the differential equation $y'' - 3y' + 2y = 0$ over $\mathbb{C}$ and let $\mathcal{B} = \mathcal{E}$ be the basis $v_1 = e^t$, $v_2 = e^{2t}$. Since the coefficients of this equation are constants it is easy to check that if y is a solution then its derivative y' is also a solution. It follows that the map $\varphi = d/dt =$ differentiation (with respect to t) is a linear transformation from V to itself. Since $\varphi(v_1) = d(e^t)/dt = e^t = v_1$ and $\varphi(v_2) = d(e^{2t})/dt = 2e^{2t} = 2v_2$ we see that the corresponding matrix with respect to these bases is the diagonal matrix $\begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$.

(3) Let $V = W = \mathbb{Q}^3 = \{(x, y, z) \mid x, y, z \in \mathbb{Q}\}$ be the usual 3-dimensional vector space of ordered 3-tuples with entries from the field $F = \mathbb{Q}$ of rational numbers and suppose φ is the linear transformation

$$\varphi(x, y, z) = (9x + 4y + 5z, -4x - 3z, -6x - 4y - 2z), \qquad x, y, z \in \mathbb{Q}$$

from V to itself. Take the standard basis $e_1 = (1, 0, 0)$, $e_2 = (0, 1, 0)$, $e_3 = (0, 0, 1)$ for V and for $W = V$. Since $\varphi(1, 0, 0) = (9, -4, -6)$, $\varphi(0, 1, 0) = (4, 0, -4)$, $\varphi(0, 0, 1) = (5, -3, -2)$, the matrix A representing this linear transformation with respect to these bases is

$$A = \begin{pmatrix} 9 & 4 & 5 \\ -4 & 0 & -3 \\ -6 & -4 & -2 \end{pmatrix}.$$

Theorem 10. Let V be a vector space over F of dimension n and let W be a vector space over F of dimension m, with bases $\mathcal{B}$, $\mathcal{E}$ respectively. Then the map $\mathrm{Hom}_F(V, W) \to M_{m\times n}(F)$ from the space of linear transformations from V to W to the space of $m \times n$ matrices with coefficients in F defined by $\varphi \mapsto M_{\mathcal{B}}^{\mathcal{E}}(\varphi)$ is a vector space isomorphism. In particular, there is a bijective correspondence between linear transformations and their associated matrices with respect to a fixed choice of bases.

Proof: The columns of the matrix $M_{\mathcal{B}}^{\mathcal{E}}(\varphi)$ are determined by the action of φ on the basis $\mathcal{B}$ as in equation (3). This shows in particular that the map $\varphi \mapsto M_{\mathcal{B}}^{\mathcal{E}}(\varphi)$ is an F-linear map since φ is F-linear. This map is *surjective* since given a matrix M, the map φ *defined* by equation (3) on a basis and then extended by linearity is a linear transformation with matrix M. The map is *injective* since two linear transformations agreeing on a basis are the same.

Note that different choices of bases give rise to different isomorphisms, so in the same sense that there is no natural choice of basis for a vector space, there is no natural isomorphism between $\mathrm{Hom}_F(V, W)$ and $M_{m\times n}(F)$.

Corollary 11. The dimension of $\mathrm{Hom}_F(V, W)$ is $(\dim V)(\dim W)$.

Proof: The dimension of $M_{m\times n}(F)$ is mn.

Definition. An $m \times n$ matrix A is called *nonsingular* if $Ax = 0$ with $x \in F^n$ implies $x = 0$.

The connection of the term *nonsingular* applied to matrices and to linear transformations is the following: let $A = M_{\mathcal{B}}^{\mathcal{E}}(\varphi)$ be the matrix associated to the linear transformation φ (with some choice of bases $\mathcal{B}$, $\mathcal{E}$). Then independently of the choice of bases, the $m \times n$ matrix A is nonsingular if and only if the linear transformation φ is a nonsingular linear transformation from the n-dimensional space V to the m-dimensional space W (cf. the exercises).

Assume now that U, V and W are all finite dimensional vector spaces over F with ordered bases $\mathcal{D}$, $\mathcal{B}$ and $\mathcal{E}$ respectively, where $\mathcal{B}$ and $\mathcal{E}$ are as before and suppose $\mathcal{D} = \{u_1, u_2, \ldots, u_k\}$. Assume $\psi : U \to V$ and $\varphi : V \to W$ are linear transformations. Their composite, $\varphi \circ \psi$, is a linear transformation from U to W, so we can compute its matrix with respect to the appropriate bases; namely, $M_{\mathcal{D}}^{\mathcal{E}}(\varphi \circ \psi)$ is found by computing

$$\varphi \circ \psi(u_j) = \sum_{i=1}^{m} \gamma_{ij} w_i$$

and putting the coefficients γ_{ij} down the j^{th} column of $M_{\mathcal{D}}^{\mathcal{E}}(\varphi \circ \psi)$. Next, compute the matrices of ψ and φ separately:

$$\psi(u_j) = \sum_{p=1}^{n} \alpha_{pj} v_p \qquad \text{and} \qquad \varphi(v_p) = \sum_{i=1}^{m} \beta_{ip} w_i$$

so that $M_{\mathcal{D}}^{\mathcal{B}}(\psi) = (\alpha_{pj})$ and $M_{\mathcal{B}}^{\mathcal{E}}(\varphi) = (\beta_{ip})$.

Using these coefficients we can find an expression for the γ's in terms of the α's and β's as follows:

$$\begin{aligned}
\varphi \circ \psi(u_j) &= \varphi\left(\sum_{p=1}^{n} \alpha_{pj} v_p \right) \\
&= \sum_{p=1}^{n} \alpha_{pj} \varphi(v_p) \\
&= \sum_{p=1}^{n} \alpha_{pj} \sum_{i=1}^{m} \beta_{ip} w_i \\
&= \sum_{p=1}^{n} \sum_{i=1}^{m} \alpha_{pj} \beta_{ip} w_i .
\end{aligned}$$

By interchanging the order of summation in the above double sum we see that γ_{ij}, which is the coefficient of w_i in the above expression, is

$$\gamma_{ij} = \sum_{p=1}^{n} \alpha_{pj} \beta_{ip} .$$

Computing the product of the matrices for φ and ψ (in that order) we obtain

$$(\beta_{ij})(\alpha_{ij}) = (\delta_{ij}), \quad \text{where} \quad \delta_{ij} = \sum_{p=1}^{m} \beta_{ip} \alpha_{pj} .$$

By comparing the two sums above and using the commutativity of field multiplication, we see that for all i and j, $\gamma_{ij} = \delta_{ij}$. This computation proves the following result:

Theorem 12. With notations as above, $M_{\mathcal{D}}^{\mathcal{E}}(\varphi \circ \psi) = M_{\mathcal{B}}^{\mathcal{E}}(\varphi) M_{\mathcal{D}}^{\mathcal{B}}(\psi)$, i.e., with respect to a compatible choice of bases, the product of the matrices representing the linear transformations φ and ψ is the matrix representing the composite linear transformation $\varphi \circ \psi$.

Corollary 13. Matrix multiplication is associative and distributive (whenever the dimensions are such as to make products defined). An $n \times n$ matrix A is nonsingular if and only if it is invertible.

Proof: Let A, B and C be matrices such that the products $(AB)C$ and $A(BC)$ are defined, and let S, T and R denote the associated linear transformations. By Theorem 12, the linear transformation corresponding to AB is the composite $S \circ T$ so the linear transformation corresponding to $(AB)C$ is the composite $(S \circ T) \circ R$. Similarly, the linear transformation corresponding to $A(BC)$ is the composite $S \circ (T \circ R)$. Since function composition is associative, these two linear transformations are the same, and so $(AB)C = A(BC)$ by Theorem 10. The distributivity is proved similarly. Note also that it is possible to prove these results by straightforward (albeit tedious) calculations with matrices.

If A is invertible, then $Ax = 0$ implies $x = A^{-1}Ax = A^{-1}0 = 0$, so A is nonsingular. Conversely, if A is nonsingular, fix bases $\mathcal{B}$, $\mathcal{E}$ for V and let φ be the linear transformation of V to itself represented by A with respect to these bases. By Corollary 9, φ is an isomorphism of V to itself, hence has an inverse, φ^{-1}. Let B be the matrix representing φ^{-1} with respect to the bases $\mathcal{E}$, $\mathcal{B}$ (note the order). Then $AB = M_{\mathcal{B}}^{\mathcal{E}}(\varphi) M_{\mathcal{E}}^{\mathcal{B}}(\varphi^{-1}) = M_{\mathcal{E}}^{\mathcal{E}}(\varphi \circ \varphi^{-1}) = M_{\mathcal{E}}^{\mathcal{E}}(1) = I$. Similarly, $BA = I$ so B is the inverse of A.

Corollary 14.

(1) If $\mathcal{B}$ is a basis of the n-dimensional space V, the map $\varphi \mapsto M_{\mathcal{B}}^{\mathcal{B}}(\varphi)$ is a ring and a vector space isomorphism of $\mathrm{Hom}_F(V, V)$ onto the space $M_n(F)$ of $n \times n$ matrices with coefficients in F.

(2) $GL(V) \cong GL_n(F)$ where $\dim V = n$. In particular, if F is a finite field the order of the finite group $GL_n(F)$ (which equals $|GL(V)|$) is given by the formula at the end of Section 1.

Proof: (1) We have already seen in Theorem 10 that this map is an isomorphism of vector spaces over F. Corollary 13 shows that $M_n(F)$ is a ring under matrix multiplication, and then Theorem 12 shows that multiplication is preserved under this map, hence it is also a ring isomorphism.

(2) This is immediate from (1) since a ring isomorphism sends units to units.

Definition. If A is any $m \times n$ matrix with entries from F, the *row rank* (respectively, *column rank*) of A is the maximal number of linearly independent rows (respectively,

columns) of A (where the rows or columns of A are considered as vectors in affine n-space, m-space, respectively).

The relation between the rank of a matrix and the rank of the associated linear transformation is the following: the rank of φ as a linear transformation equals the column rank of the matrix $M_{\mathcal{B}}^{\mathcal{E}}(\varphi)$ (cf. the exercises). We shall also see that the row rank and the column rank of any matrix are the same.

We now consider the relation of two matrices associated to the same linear transformation of a vector space to itself but with respect to two different choices of bases (cf. the exercises for the general statement regarding a linear transformation from a vector space V to another vector space W).

Definition. Two $n \times n$ matrices A and B are said to be *similar* if there is an invertible (i.e., nonsingular) $n \times n$ matrix P such that $P^{-1}AP = B$. Two linear transformations φ and ψ from a vector space V to itself are said to be *similar* if there is a nonsingular linear transformation ξ from V to V such that $\xi^{-1}\varphi\xi = \psi$.

Suppose $\mathcal{B}$ and $\mathcal{E}$ are two bases of the same vector space V and let $\varphi \in \mathrm{Hom}_F(V, V)$. Let I be the identity map from V to V and let $P = M_{\mathcal{E}}^{\mathcal{B}}(I)$ be its associated matrix (in other words, write the elements of the basis $\mathcal{E}$ in terms of the basis $\mathcal{B}$ — note the order — and use the resulting coordinates for the columns of the matrix P). Note that if $\mathcal{B} \neq \mathcal{E}$ then P is *not* the identity matrix. Then $P^{-1}M_{\mathcal{B}}^{\mathcal{B}}(\varphi)P = M_{\mathcal{E}}^{\mathcal{E}}(\varphi)$. If $[v]_{\mathcal{B}}$ is the $n \times 1$ matrix of coordinates for $v \in V$ with respect to the basis $\mathcal{B}$, and similarly $[v]_{\mathcal{E}}$ is the $n \times 1$ matrix of coordinates for $v \in V$ with respect to the basis $\mathcal{E}$, then $[v]_{\mathcal{B}} = P[v]_{\mathcal{E}}$. The matrix P is called the *transition* or *change of basis* matrix from $\mathcal{B}$ to $\mathcal{E}$ and this similarity action on $M_{\mathcal{B}}^{\mathcal{B}}(\varphi)$ is called a *change of basis*. This shows that the matrices associated to the same linear transformation with respect to two different bases are similar.

Conversely, suppose A and B are $n \times n$ matrices similar by a nonsingular matrix P. Let $\mathcal{B}$ be a basis for the n-dimensional vector space V. Define the linear transformation φ of V (with basis $\mathcal{B}$) to V (again with basis $\mathcal{B}$) by equation (3) using the given matrix A, i.e.,

$$\varphi(v_j) = \sum_{i=1}^{n} \alpha_{ij} v_i.$$

Then $A = M_{\mathcal{B}}^{\mathcal{B}}(\varphi)$ by definition of φ. Define a new basis $\mathcal{E}$ of V by using the i^{th} column of P for the coordinates of w_i in terms of the basis $\mathcal{B}$ (so $P = M_{\mathcal{E}}^{\mathcal{B}}(I)$ by definition). Then $B = P^{-1}AP = P^{-1}M_{\mathcal{B}}^{\mathcal{B}}(\varphi)P = M_{\mathcal{E}}^{\mathcal{E}}(\varphi)$ is the matrix associated to φ with respect to the basis $\mathcal{E}$. This shows that any two similar $n \times n$ matrices arise in this fashion as the matrices representing the same linear transformation with respect to two different choices of bases.

Note that change of basis for a linear transformation from V to itself is the same as conjugation by some element of the group $GL(V)$ of nonsingular linear transformations of V to V. In particular, the relation "similarity" is an equivalence relation whose equivalence classes are the orbits of $GL(V)$ acting by conjugation on $\mathrm{Hom}_F(V, V)$. If

$\varphi \in GL(V)$ (i.e., φ is an invertible linear transformation), then the similarity class of φ is none other than the conjugacy class of φ in the group $GL(V)$.

Example

Let $V = \mathbb{Q}^3$ and let φ be the linear transformation

$$\varphi(x, y, z) = (9x + 4y + 5z, -4x - 3z, -6x - 4y - 2z), \qquad x, y, z \in \mathbb{Q}$$

from V to itself we considered in an earlier example. With respect to the standard basis, $\mathcal{B}, b_1 = (1, 0, 0), b_2 = (0, 1, 0), b_3 = (0, 0, 1)$ we saw that the matrix A representing this linear transformation is

$$A = M_{\mathcal{B}}^{\mathcal{B}}(\varphi) = \begin{pmatrix} 9 & 4 & 5 \\ -4 & 0 & -3 \\ -6 & -4 & -2 \end{pmatrix}.$$

Take now the basis, $\mathcal{E}$, $e_1 = (2, -1, -2)$, $e_2 = (1, 0, -1)$, $e_3 = (3, -2, -2)$ for V (we shall see that this is in fact a basis momentarily). Since

$$\varphi(e_1) = \varphi(2, -1, -2) = (4, -2, -4) = 2 \cdot e_1 + 0 \cdot e_2 + 0 \cdot e_3$$
$$\varphi(e_2) = \varphi(1, 0, -1) = (4, -1, -4) = 1 \cdot e_1 + 2 \cdot e_2 + 0 \cdot e_3$$
$$\varphi(e_3) = \varphi(3, -2, -2) = (9, -6, -6) = 0 \cdot e_1 + 0 \cdot e_2 + 3 \cdot e_3,$$

the matrix representing φ with respect to this basis is the matrix

$$B = M_{\mathcal{E}}^{\mathcal{E}}(\varphi) = \begin{pmatrix} 2 & 1 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix}.$$

Writing the elements of the basis $\mathcal{E}$ in terms of the basis $\mathcal{B}$ we have

$$e_1 = 2b_1 - b_2 - 2b_3$$
$$e_2 = b_1 - b_3$$
$$e_3 = 3b_1 - 2b_2 - 2b_3$$

so the matrix $P = M_{\mathcal{E}}^{\mathcal{B}}(I) = \begin{pmatrix} 2 & 1 & 3 \\ -1 & 0 & -2 \\ -2 & -1 & -2 \end{pmatrix}$ with inverse $P^{-1} = \begin{pmatrix} -2 & -1 & -2 \\ 2 & 2 & 1 \\ 1 & 0 & 1 \end{pmatrix}$ conjugates A into B, i.e., $P^{-1}AP = B$, as can easily be checked. (Note incidentally that since P is invertible this proves that $\mathcal{E}$ is indeed a basis for V.)

We observe in passing that the matrix B representing this linear transformation φ is much simpler than the matrix A representing φ. The study of the simplest possible matrix representing a given linear transformation (and which basis to choose to realize it) is the study of *canonical forms* considered in the next chapter.

Linear Transformations on Tensor Products of Vector Spaces

For convenience we reiterate Corollaries 18 and 19 of Section 10.4 for the special case of vector spaces.

Proposition 15. Let F be a subfield of the field K. If W is an m-dimensional vector space over F with basis $w_1, \ldots, w_m$, then $K \otimes_F W$ is an m-dimensional vector space over K with basis $1 \otimes w_1, \ldots, 1 \otimes w_m$.

Proposition 16. Let V and W be finite dimensional vector spaces over the field F with bases $v_1, \ldots, v_n$ and $w_1, \ldots, w_m$ respectively. Then $V \otimes_F W$ is a vector space over F of dimension nm with basis $v_i \otimes w_j$, $1 \leq i \leq n$ and $1 \leq j \leq m$.

Remark: If v and w are nonzero elements of V and W, respectively, then it follows from the proposition that $v \otimes w$ is a nonzero element of $V \otimes_F W$, because we may always build bases of V and W whose first basis vectors are v, w, respectively. In a tensor product $M \otimes_R N$ of two R-modules where R is not a field it is in general substantially more difficult to determine when the tensor product $m \otimes n$ of two nonzero elements is zero.

Now let V, W, X, Y be finite dimensional vector spaces over F and let

$$\varphi : V \to X \qquad \text{and} \qquad \psi : W \to Y$$

be linear transformations. We compute a matrix of the linear transformation

$$\varphi \otimes \psi : V \otimes W \to X \otimes Y.$$

Let $\mathcal{B}_1 = \{v_1, \ldots, v_n\}$ and $\mathcal{B}_2 = \{w_1, \ldots, w_m\}$ be (ordered) bases of V and W respectively, and let $\mathcal{E}_1 = \{x_1, \ldots, x_r\}$ and $\mathcal{E}_2 = \{y_1, \ldots, y_s\}$ be (ordered) bases of X and Y respectively. Let $\mathcal{B} = \{v_i \otimes w_j\}$ and $\mathcal{E} = \{x_i \otimes y_j\}$ be the bases of $V \otimes W$ and $X \otimes Y$ given by Proposition 16; we shall order these shortly. Suppose

$$\varphi(v_i) = \sum_{p=1}^{r} \alpha_{pi} x_p \qquad \text{and} \qquad \psi(w_j) = \sum_{q=1}^{s} \beta_{qj} y_q.$$

Then

$$\begin{aligned} (\varphi \otimes \psi)(v_i \otimes w_j) &= (\varphi(v_i)) \otimes (\psi(w_j)) \\ &= (\sum_{p=1}^{r} \alpha_{pi} x_p) \otimes (\sum_{q=1}^{s} \beta_{qj} y_q) \\ &= \sum_{p=1}^{r} \sum_{q=1}^{s} \alpha_{pi} \beta_{qj} (x_p \otimes y_q). \end{aligned} \tag{11.8}$$

In view of the order of summation in (11.8) we order the basis $\mathcal{E}$ into r ordered sets, with the p^{th} list being $x_p \otimes y_1, x_p \otimes y_2, \ldots, x_p \otimes y_s$, and similarly order the basis $\mathcal{B}$. Then equation (8) determines the column entries for the corresponding matrix of $\varphi \otimes \psi$. The resulting matrix $M_{\mathcal{B}}^{\mathcal{E}}(\varphi \otimes \psi)$ is an $r \times n$ block matrix whose p, q block is the $s \times m$ matrix $\alpha_{p,q} M_{\mathcal{B}_2}^{\mathcal{E}_2}(\psi)$. In other words, the matrix for $\varphi \otimes \psi$ is obtained by taking the matrix for φ and multiplying each entry by the matrix for ψ. Such matrices have a name:

Definition. Let $A = (\alpha_{ij})$ and B be $r \times n$ and $s \times m$ matrices, respectively, with coefficients from any commutative ring. The *Kronecker product* or *tensor product* of A and B, denoted by $A \otimes B$, is the $rs \times nm$ matrix consisting of an $r \times n$ block matrix whose i, j block is the $s \times m$ matrix $\alpha_{ij} B$.

With this terminology we have

Proposition 17. Let $\varphi : V \to X$ and $\psi : W \to Y$ be linear transformations of finite dimensional vector spaces. Then the Kronecker product of matrices representing φ and ψ is a matrix representation of $\varphi \otimes \psi$.

Example

Let $V = X = \mathbb{R}^3$, both with basis v_1, v_2, v_3, and $W = Y = \mathbb{R}^2$, both with basis w_1, w_2. Suppose $\varphi : \mathbb{R}^3 \to \mathbb{R}^3$ is the linear transformation given by $\varphi(av_1 + bv_2 + cv_3) = cv_1 + 2av_2 - 3bv_3$ and $\psi : \mathbb{R}^2 \to \mathbb{R}^2$ is the linear transformation given by $\psi(aw_1 + bw_2) = (a+3b)w_1 + (4b-2a)w_2$. With respect to the chosen bases, the matrices for φ and ψ are

$$\begin{pmatrix} 0 & 0 & 1 \\ 2 & 0 & 0 \\ 0 & -3 & 0 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 1 & 3 \\ -2 & 4 \end{pmatrix},$$

respectively. Then with respect to the ordered basis

$$\mathcal{B} = \{v_1 \otimes w_1,\ v_1 \otimes w_2,\ v_2 \otimes w_1,\ v_2 \otimes w_2,\ v_3 \otimes w_1,\ v_3 \otimes w_2\}$$

we have

$$M_{\mathcal{B}}^{\mathcal{B}}(\varphi \otimes \psi) = \left(\begin{array}{cc:cc:cc} 0 & 0 & 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & 0 & -2 & 4 \\ \hdashline 2 & 6 & 0 & 0 & 0 & 0 \\ -4 & 8 & 0 & 0 & 0 & 0 \\ \hdashline 0 & 0 & -3 & -9 & 0 & 0 \\ 0 & 0 & 6 & -12 & 0 & 0 \end{array}\right),$$

obtained (as indicated by the dashed lines) by multiplying the 2×2 matrix for ψ successively by the entries in the matrix for φ.

EXERCISES

1. Let V be the collection of polynomials with coefficients in $\mathbb{Q}$ in the variable x of degree at most 5. Determine the transition matrix from the basis $1, x, x^2, \ldots, x^5$ for V to the basis $1, 1+x, 1+x+x^2, \ldots, 1+x+x^2+x^3+x^4+x^5$ for V.
2. Let V be the vector space of the preceding exercise. Let $\varphi = d/dx$ be the linear transformation of V to itself given by usual differentiation of a polynomial with respect to x. Determine the matrix of φ with respect to the two bases for V in the previous exercise.
3. Let V be the collection of polynomials with coefficients in F in the variable x of degree at most n. Determine the transition matrix from the basis $1, x, x^2, \ldots, x^n$ for V to the elements
$$1,\ x - \lambda, \ldots,\ (x-\lambda)^{n-1},\ (x-\lambda)^n$$
where λ is a fixed element of F. Conclude that these elements are a basis for V.
4. Let φ be the linear transformation of $\mathbb{R}^2$ to itself given by rotation counterclockwise around the origin through an angle θ. Show that the matrix of φ with respect to the standard basis for $\mathbb{R}^2$ is $\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$.
5. Show that the $m \times n$ matrix A is nonsingular if and only if the linear transformation φ is a nonsingular linear transformation from the n-dimensional space V to the m-dimensional space W, where $A = M_{\mathcal{B}}^{\mathcal{E}}(\varphi)$, regardless of the choice of bases $\mathcal{B}$ and $\mathcal{E}$.

6. Prove if $\varphi \in \mathrm{Hom}_F(F^n, F^m)$, and $\mathcal{B}, \mathcal{E}$ are the natural bases of F^n, F^m respectively, then the range of φ equals the span of the set of columns of $M_{\mathcal{B}}^{\mathcal{E}}(\varphi)$. Deduce that the rank of φ (as a linear transformation) equals the column rank of $M_{\mathcal{B}}^{\mathcal{E}}(\varphi)$.

7. Prove that any two similar matrices have the same row rank and the same column rank.

8. Let V be an n-dimensional vector space over F and let φ be a linear transformation of the vector space V to itself.
   (a) Prove that if V has a basis consisting of eigenvectors for φ (cf. Exercise 8 of Section 1) then the matrix representing φ with respect to this basis (for both domain and range) is diagonal with the eigenvalues as diagonal entries.
   (b) If A is the $n \times n$ matrix representing φ with respect to a given basis for V (for both domain and range) prove that A is similar to a diagonal matrix if and only if V has a basis of eigenvectors for φ.

9. If W is a subspace of the vector space V stable under the linear transformation φ (i.e., $\varphi(W) \subseteq W$), show that φ induces linear transformations $\varphi|_W$ on W and $\bar{\varphi}$ on the quotient vector space V/W. If $\varphi|_W$ and $\bar{\varphi}$ are nonsingular prove φ is nonsingular. Prove the converse holds if V has finite dimension and give a counterexample with V infinite dimensional.

10. Let V be an n-dimensional vector space and let φ be a linear transformation of V to itself. Suppose W is a subspace of V of dimension m that is stable under φ.
    (a) Prove that there is a basis for V with respect to which the matrix for φ is of the form
    $$\begin{pmatrix} A & B \\ 0 & C \end{pmatrix}$$
    where A is an $m \times m$ matrix, B is an $m \times (n-m)$ matrix and C is an $(n-m) \times (n-m)$ matrix (such a matrix is called *block upper triangular*).
    (b) Prove that if there is a subspace W' invariant under φ so that $V = W \oplus W'$ decomposes as a direct sum then the bases for W and W' give a basis for V with respect to which the matrix for φ is *block diagonal*:
    $$\begin{pmatrix} A & 0 \\ 0 & C \end{pmatrix}$$
    where A is an $m \times m$ matrix and C is an $(n-m) \times (n-m)$ matrix.
    (c) Prove conversely that if there is a basis for V with respect to which φ is block diagonal as in (b) then there are φ-invariant subspaces W and W' of dimensions m and $n-m$, respectively, with $V = W \oplus W'$.

11. Let φ be a linear transformation from the finite dimensional vector space V to itself such that $\varphi^2 = \varphi$.
    (a) Prove that image $\varphi \cap \ker \varphi = 0$.
    (b) Prove that $V = \text{image } \varphi \oplus \ker \varphi$.
    (c) Prove that there is a basis of V such that the matrix of φ with respect to this basis is a diagonal matrix whose entries are all 0 or 1.

    A linear transformation φ satisfying $\varphi^2 = \varphi$ is called an *idempotent* linear transformation. This exercise proves that idempotent linear transformations are simply projections onto some subspace.

12. Let $V = \mathbb{R}^2$, $v_1 = (1, 0)$, $v_2 = (0, 1)$, so that v_1, v_2 are a basis for V. Let φ be the linear transformation of V to itself whose matrix with respect to this basis is $\begin{pmatrix} 2 & 1 \\ 0 & 2 \end{pmatrix}$. Prove that if W is the subspace generated by v_1 then W is stable under the action of φ. Prove that there is no subspace W' invariant under φ so that $V = W \oplus W'$.

13. Let V be a vector space of dimension n and let W be a vector space of dimension m over a field F. Suppose A is the $m \times n$ matrix representing a linear transformation φ from V to W with respect to the bases $\mathcal{B}_1$ for V and $\mathcal{E}_1$ for W. Suppose similarly that B is the $m \times n$ matrix representing φ with respect to the bases $\mathcal{B}_2$ for V and $\mathcal{E}_2$ for W. Let $P = M_{\mathcal{B}_2}^{\mathcal{B}_1}(I)$ where I denotes the identity map from V to V, and let $Q = M_{\mathcal{E}_2}^{\mathcal{E}_1}(I)$ where I denotes the identity map from W to W. Prove that $Q^{-1} = M_{\mathcal{E}_1}^{\mathcal{E}_2}(I)$ and that $Q^{-1}AP = B$, giving the general relation between matrices representing the same linear transformation but with respect to different choices of bases.

The following exercises recall the *Gauss–Jordan* elimination process. This is one of the fastest computational methods for the solution of a number of problems involving vector spaces — solving systems of linear equations, determining inverses of matrices, computing determinants, determining the span of a set of vectors, determining linear independence of a set of vectors etc.

Consider the system of m linear equations

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= c_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n &= c_2 \\ &\vdots \\ a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n &= c_m \end{aligned} \tag{11.4}$$

in the n unknowns $x_1, x_2, \ldots, x_n$ where a_{ij}, c_i, $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, n$ are elements of the field F. Associated to this system is the *coefficient matrix*:

$$A = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{pmatrix}$$

and the *augmented matrix*:

$$(A \mid C) = \left(\begin{array}{cccc|c} a_{11} & a_{12} & \ldots & a_{1n} & c_1 \\ a_{21} & a_{22} & \ldots & a_{2n} & c_2 \\ \vdots & \vdots & & \vdots & \vdots \\ a_{m1} & a_{m2} & \ldots & a_{mn} & c_m \end{array}\right)$$

(the term *augmented* refers to the presence of the column matrix $C = (c_i)$ in addition to the coefficient matrix $A = (a_{ij})$). The set of solutions in F of this system of equations is not altered if we perform any of the following three operations:

(1) interchange any two equations
(2) add a multiple of one equation to another
(3) multiply any equation by a nonzero element from F,

which correspond to the following three *elementary row operations* on the augmented matrix:

(1) interchange any two rows
(2) add a multiple of one row to another
(3) multiply any row by a unit in F, i.e., by any nonzero element in F.

If a matrix A can be transformed into a matrix C by a series of elementary row operations then A is said to be *row reduced* to C.

14. Prove that if A can be row reduced to C then C can be row reduced to A. Prove that the relation "$A \sim C$ if and only if A can be row reduced to C" is an equivalence relation. [Observe that the elementary row operations are reversible.]

Matrices lying in the same equivalence class under this equivalence relation are said to be *row equivalent*.

15. Prove that the row rank of two row equivalent matrices is the same. [It suffices to prove this for two matrices differing by an elementary row operation.]

An $m \times n$ matrix is said to be in *reduced row echelon form* if

(a) the first nonzero entry a_{ij_i} in row i is 1 and all other entries in the corresponding j_i^{th} column are zero, and

(b) $j_1 < j_2 < \ldots < j_r$ where r is the number of nonzero rows, i.e., the number of initial zeros in each row is strictly increasing (hence the term *echelon*).

An augmented matrix $(A \mid C)$ is said to be in reduced row echelon form if its coefficient matrix A is in reduced row echelon form. For example, the following two matrices are in reduced row echelon form:

$$\left(\begin{array}{cccccc|c} 1 & 0 & 5 & 7 & 0 & 3 & 0 \\ 0 & 1 & -1 & 1 & 0 & -4 & -1 \\ 0 & 0 & 0 & 0 & 1 & 6 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right) \qquad \left(\begin{array}{cccc|c} 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 0 & -3 \end{array}\right)$$

(with $j_1 = 1$, $j_2 = 2$, $j_3 = 5$ for the first matrix and $j_1 = 2$, $j_2 = 4$ for the second matrix). The first nonzero entry in any given row of the coefficient matrix of a reduced row echelon augmented matrix (in position (i, j_i) by definition) is sometimes referred to as a *pivotal* element (so the pivotal elements in the first matrix are in positions (1,1), (2,2) and (3,5) and the pivotal elements in the second matrix are in positions (1,2) and (2,4)). The columns containing pivotal elements will be called *pivotal* columns and the columns of the coefficient matrix not containing pivotal elements will be called *nonpivotal*.

16. Prove by induction that any augmented matrix can be put in reduced row echelon form by a series of elementary row operations.

17. Let A and C be two matrices in reduced row echelon form. Prove that if A and C are row equivalent then $A = C$.

18. Prove that the row rank of a matrix in reduced row echelon form is the number of nonzero rows.

19. Prove that the reduced row echelon forms of the matrices

$$\left(\begin{array}{cccccc|c} 1 & 1 & 4 & 8 & 0 & -1 & -1 \\ 1 & 2 & 3 & 9 & 0 & -5 & -2 \\ 0 & -2 & 2 & -2 & 1 & 14 & 3 \\ 1 & 4 & 1 & 11 & 0 & -13 & -4 \end{array}\right) \qquad \left(\begin{array}{cccc|c} 0 & -3 & 3 & 1 & 5 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 2 & -2 & 0 & -3 \end{array}\right)$$

are the two matrices preceding Exercise 16.

The point of the reduced row echelon form is that the corresponding system of linear equations is in a particularly simple form, from which the solutions to the system $AX = C$ in (4) can be determined immediately:

20. (*Solving Systems of Linear Equations*) Let $(A' \mid C')$ be the reduced row echelon form of the augmented matrix $(A \mid C)$. The number of zero rows of A' is clearly at least as great as the number of zero rows of $(A' \mid C')$.

(a) Prove that if the number of zero rows of A' is strictly larger than the number of zero rows of $(A' \mid C')$ then there are no solutions to $AX = C$.

By (a) we may assume that A' and $(A' \mid C')$ have the same number, r, of nonzero rows (so $n \geq r$).

(b) Prove that if $r = n$ then there is precisely one solution to the system of equations $AX = C$.

(c) Prove that if $r < n$ then there are infinitely many solutions to the system of equations $AX = C$. Prove in fact that the values of the $n - r$ variables corresponding to the nonpivotal columns of $(A' \mid C')$ can be chosen arbitrarily and that the remaining r variables corresponding to the pivotal columns of $(A' \mid C')$ are then determined uniquely.

21. Determine the solutions of the following systems of equations:

(a)

$$\begin{aligned} -3x + 3y + z &= 5 \\ x - y &= 0 \\ 2x - 2y &= -3 \end{aligned}$$

(b)

$$\begin{aligned} x - 2y + z &= 5 \\ x - 4y + 6z &= 10 \\ 4x - 11y + 11z &= 12 \end{aligned}$$

(c)

$$\begin{aligned} x - 2y + z &= 5 \\ y - 2z &= 17 \\ 2x - 3y &= 27 \end{aligned}$$

(d)

$$\begin{aligned} x + y - 3z + 2u &= 2 \\ 3x - 2y + 5z + u &= 1 \\ 6x + y - 4z + 3u &= 7 \\ 2x + 2y - 6z &= 4 \end{aligned}$$

(e)

$$\begin{aligned} x + y + 4z + 8u - w &= -1 \\ x + 2y + 3z + 9u - 5w &= -2 \\ -2y + 2z - 2u + v + 14w &= 3 \\ x + 4y + z + 11u - 13w &= -4 \end{aligned}$$

22. Suppose A and B are two row equivalent $m \times n$ matrices.

(a) Prove that the set

$$X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

of solutions to the homogeneous linear equations $AX = 0$ as in equation (4) above are the same as the set of solutions to the homogeneous linear equations $BX = 0$. [It suffices to prove this for two matrices differing by an elementary row operation.]

(b) Prove that any linear dependence relation satisfied by the columns of A viewed as vectors in F^m is also satisfied by the columns of B.

(c) Conclude from (b) that the number of linearly independent columns of A is the same as the number of linearly independent columns of B.

23. Let A' be a matrix in reduced row echelon form.

(a) Prove that the nonzero rows of A' are linearly independent. Prove that the pivotal columns of A' are linearly independent and that the nonpivotal columns of A' are linearly dependent on the pivotal columns. (Note the role the pivotal elements play.)

(b) Prove that the number of linearly independent columns of a matrix in reduced row echelon form is the same as the number of linearly independent rows, i.e., the row rank and the column rank of such a matrix are the same.

24. Use the previous two exercises and Exercise 15 above to prove in general that the row rank and the column rank of a matrix are the same.

25. (*Computing Inverses of Matrices*) Let A be an $n \times n$ matrix.

(a) Show that A has an inverse matrix B with columns $B_1, B_2, \ldots, B_n$ if and only if the systems of equations:

$$AB_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}, \quad AB_2 = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \\ 0 \end{pmatrix}, \quad \ldots, \quad AB_n = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}$$

have solutions.

(b) Prove that A has an inverse if and only if A is row equivalent to the $n \times n$ identity matrix.

(c) Prove that A has an inverse B if and only if the augmented matrix $(A \mid I)$ can be row reduced to the augmented matrix $(I \mid B)$ where I is the $n \times n$ identity matrix.

26. Determine the inverses of the following matrices using row reduction:

$$A = \begin{pmatrix} -7 & -1 & -4 \\ 7 & 1 & 3 \\ 1 & 0 & 0 \end{pmatrix} \qquad B = \begin{pmatrix} 1 & 1 & 0 & 2 \\ 0 & 2 & 1 & -1 \\ 0 & 2 & 0 & 0 \\ -1 & 1 & 1 & 0 \end{pmatrix}.$$

27. (*Computing Spans, Linear Independence and Linear Dependencies in Vector Spaces*) Let V be an m-dimensional vector space with basis $e_1, e_2, \ldots, e_m$ and let $v_1, v_2, \ldots, v_n$ be vectors in V. Let A be the $m \times n$ matrix whose columns are the coordinates of the vectors v_i (with respect to the basis $e_1, e_2, \ldots, e_m$) and let A' be the reduced row echelon form of A.

(a) Let B be any matrix row equivalent to A. Let $w_1, w_2, \ldots, w_n$ be the vectors whose coordinates (with respect to the basis $e_1, e_2, \ldots, e_m$) are the columns of B. Prove that any linear relation

$$x_1v_1 + x_2v_2 + \ldots + x_nv_n = 0 \tag{11.5}$$

satisfied by $v_1, v_2, \ldots, v_n$ is also satisfied when v_i is replaced by w_i, $i = 1, 2, \ldots, n$.

(b) Prove that the vectors whose coordinates are given by the pivotal columns of A' are linearly independent and that the vectors whose coordinates are given by the nonpivotal columns of A' are linearly dependent on these.

(c) (*Determining Linear Independence of Vectors*) Prove that the vectors $v_1, v_2, \ldots, v_n$ are linearly independent if and only if A' has n nonzero rows (i.e., has rank n).

(d) (*Determining Linear Dependencies of Vectors*) By (c), the vectors $v_1, v_2, \ldots, v_n$ are linearly dependent if and only if A' has nonpivotal columns. The solutions to (5)

defining linear dependence relations among $v_1, v_2, \ldots, v_n$ are given by the linear equations defined by A'. Show that each of the variables $x_1, x_2, \ldots, x_n$ in (5) corresponding to the nonpivotal columns of A' can be prescribed arbitrarily and the values of the remaining variables are then uniquely determined to give a linear dependence relation among $v_1, v_2, \ldots, v_n$ as in (5).

(e) (*Determining the Span of a Set of Vectors*) Prove that the subspace W spanned by $v_1, v_2, \ldots, v_n$ has dimension r where r is the number of nonzero rows of A' and that a basis for W is given by the original vectors v_{j_i} $(i = 1, 2, \ldots, r)$ corresponding to the pivotal columns of A'.

28. Let $V = \mathbb{R}^5$ with the standard basis and consider the vectors

$$v_1 = (1, 1, 3, -2, 3)\,,\ v_2 = (0, 1, 0, -1, 0)\,,\ v_3 = (2, 3, 6, -5, 6)$$

$$v_4 = (0, 3, 1, -3, 1)\,,\ v_5 = (2, -1, -1, -1, -1).$$

(a) Show that the reduced row echelon form of the matrix

$$A = \begin{pmatrix} 1 & 0 & 2 & 0 & 2 \\ 1 & 1 & 3 & 3 & -1 \\ 3 & 0 & 6 & 1 & -1 \\ -2 & -1 & -5 & -3 & -1 \\ 3 & 0 & 6 & 1 & -1 \end{pmatrix}$$

whose columns are the coordinates of v_1, v_2, v_3, v_4, v_5 is the matrix

$$A' = \begin{pmatrix} 1 & 0 & 2 & 0 & 2 \\ 0 & 1 & 1 & 0 & 18 \\ 0 & 0 & 0 & 1 & -7 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

where the 1^{st}, 2^{nd} and 4^{th} columns are pivotal and the remaining two are nonpivotal.

(b) Conclude that these vectors are linearly dependent, that the subspace W spanned by v_1, v_2, v_3, v_4, v_5 is 3-dimensional and that the vectors

$$v_1 = (1, 1, 3, -2, 3), \quad v_2 = (0, 1, 0, -1, 0) \quad \text{and} \quad v_4 = (0, 3, 1, -3, 1)$$

are a basis for W.

(c) Conclude from (a) that the coefficients x_1, x_2, x_3, x_4, x_5 of any linear relation

$$x_1v_1 + x_2v_2 + x_3v_3 + x_4v_4 + x_5v_5 = 0$$

satisfied by v_1, v_2, v_3, v_4, v_5 are given by the equations

$$\begin{aligned} x_1 \quad + 2x_3 \quad + \ 2x_5 &= 0 \\ x_2 + \ x_3 \quad + 18x_5 &= 0 \\ x_4 - \ 7x_5 &= 0. \end{aligned}$$

Deduce that the 3^{rd} and 5^{th} variables, namely x_3 and x_5, corresponding to the nonpivotal columns of A', can be prescribed arbitrarily and the remaining variables are then uniquely determined as:

$$\begin{aligned} x_1 &= -2x_3 - 2x_5 \\ x_2 &= -x_3 - 18x_5 \\ x_4 &= 7x_5 \end{aligned}$$

to give all the linear dependence relations satisfied by v_1, v_2, v_3, v_4, v_5. In particular show that

$$-2v_1 - v_2 + v_3 = 0$$

and

$$-2v_1 - 18v_2 + 7v_4 + v_5 = 0$$

corresponding to $(x_3 = 1, x_5 = 0)$ and $(x_3 = 0, x_5 = 1)$, respectively.

29. For each exercise below, determine whether the given vectors in $\mathbb{R}^4$ are linearly independent. If they are linearly dependent, determine an explicit linear dependence among them.

(a) $(1, -4, 3, 0), (0, -1, 4, -3), (1, -1, 1, -1), (2, 2, -1, -3)$.
(b) $(1, -2, 4, 1), (2, -3, 9, -1), (1, 0, 6, -5), (2, -5, 7, 5)$.
(c) $(1, -2, 0, 1), (2, -2, 0, 0), (-1, 3, 0, -2), (-2, 1, 0, 1)$.
(d) $(0, 1, 1, 0), (1, 0, 1, 1), (2, 2, 2, 0), (0, -1, 1, 1)$.

30. For each exercise below, determine the subspace spanned in $\mathbb{R}^4$ by the given vectors and give a basis for this subspace.

(a) $(1, -2, 5, 3), (2, 3, 1, -4), (3, 8, -3, -5)$.
(b) $(2, -5, 3, 0), (0, -2, 5, -3), (1, -1, 1, -1), (-3, 2, -1, 2)$.
(c) $(1, -2, 0, 1), (2, -2, 0, 0), (-1, 3, 0, -2), (-2, 1, 0, 1)$.
(d) $(1, 1, 0, -1), (1, 2, 3, 0), (2, 3, 3, -1), (1, 2, 2, -2), (2, 3, 2, -3), (1, 3, 4, -3)$.

31. (*Computing the Image and Kernel of a Linear Transformation*) Let V be an n-dimensional vector space with basis $e_1, e_2, \ldots, e_n$ and let W be an m-dimensional vector space with basis $f_1, f_2, \ldots, f_m$. Let φ be a linear transformation from V to W and let A be the corresponding $m \times n$ matrix with respect to these bases: $A = (a_{ij})$ where

$$\varphi(e_j) = \sum_{i=1}^{m} a_{ij} f_i , \qquad j = 1, 2, \ldots, n,$$

i.e., the columns of A are the coordinates of the vectors $\varphi(e_1), \varphi(e_2), \ldots, \varphi(e_n)$ with respect to the basis $f_1, f_2, \ldots, f_m$ of W. Let A' be the reduced row echelon form of A.

(a) (*Determining the Image of a Linear Transformation*) Prove that the image $\varphi(V)$ of V under φ has dimension r where r is the number of nonzero rows of A' and that a basis for $\varphi(V)$ is given by the vectors $\varphi(e_{j_i})$ $(i = 1, 2, \ldots, r)$, i.e., the columns of A corresponding to the pivotal columns of A' give the coordinates of a basis for the image of φ.

(b) (*Determining the Kernel of a Linear Transformation*) The elements in the kernel of φ are the vectors in V whose coordinates $(x_1, x_2, \ldots, x_n)$ with respect to the basis $e_1, e_2, \ldots, e_n$ satisfy the equation

$$A \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = 0,$$

and the solutions $x_1, x_2, \ldots, x_n$ to this system of linear equations are determined by the matrix A'.

(i) Prove that φ is injective if and only if A' has n nonzero rows (i.e., has rank n).

(ii) By (i), the kernel of φ is nontrivial if and only if A' has nonpivotal columns. Show that each of the variables $x_1, x_2, \ldots, x_n$ above corresponding to the nonpivotal columns of A' can be prescribed arbitrarily and the values of the remaining variables are then

uniquely determined to give an element $x_1e_1 + x_2e_2 + \ldots + x_ne_n$ in the kernel of φ. In particular, show that the coordinates of a basis for the kernel are obtained by successively setting one nonpivotal variable equal to 1 and all other nonpivotal variables to 0 and solving for the remaining pivotal variables. Conclude that the kernel of φ has dimension $n - r$ where r is the rank of A.

32. Let $V = \mathbb{R}^5$ and $W = \mathbb{R}^4$ with the standard bases. Let φ be the linear transformation $\varphi : V \to W$ defined by

$$\varphi(x, y, z, u, v) = (x + 2y + 3z + 4u + 4v, -2x - 4y + 2v, x + 2y + u - 2v, x + 2y - v).$$

(a) Prove that the matrix A corresponding to φ and these bases is

$$A = \begin{pmatrix} 1 & 2 & 3 & 4 & 4 \\ -2 & -4 & 0 & 0 & 2 \\ 1 & 2 & 0 & 1 & -2 \\ 1 & 2 & 0 & 0 & -1 \end{pmatrix}$$

and that the reduced row echelon matrix A' row equivalent to A is

$$A' = \begin{pmatrix} 1 & 2 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 & 3 \\ 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

where the 1st, 3rd and 4th columns are pivotal and the remaining two are nonpivotal.

(b) Conclude that the image of φ is 3-dimensional and that the image of the 1st, 3rd and 4th basis elements of V, namely, $(1, -2, 1, 1)$, $(3, 0, 0, 0)$ and $(4, 0, 1, 0)$ give a basis for the image $\varphi(V)$ of V.

(c) Conclude from (a) that the elements in the kernel of φ are the vectors (x, y, z, u, v) satisfying the equations

$$\begin{aligned} x + 2y \quad - \quad v &= 0 \\ z \quad + 3v &= 0 \\ u - \quad v &= 0. \end{aligned}$$

Deduce that the 2nd and 5th variables, namely y and v, corresponding to the nonpivotal columns of A' can be prescribed arbitrarily and the remaining variables are then uniquely determined as

$$\begin{aligned} x &= -2y + v \\ z &= -3v \\ u &= v. \end{aligned}$$

Show that $(-2, 1, 0, 0, 0)$ and $(1, 0, -3, 1, 1)$ give a basis for the 2-dimensional kernel of φ, corresponding to $(y = 1, v = 0)$ and $(y = 0, v = 1)$, respectively.

33. Let φ be the linear transformation from $\mathbb{R}^4$ to itself defined by the matrix

$$A = \begin{pmatrix} 1 & -1 & 0 & 3 \\ -1 & 2 & 1 & -1 \\ -1 & 1 & 0 & -3 \\ 1 & -2 & -1 & 1 \end{pmatrix}$$

with respect to the standard basis for $\mathbb{R}^4$. Determine a basis for the image and for the kernel of φ.

34. Let φ be the linear transformation $\varphi : \mathbb{R}^4 \to \mathbb{R}^2$ such that

$$\varphi((1,0,0,0)) = (1,-1) \qquad \varphi((1,-1,0,0)) = (0,0)$$
$$\varphi((1,-1,1,0)) = (1,-1) \qquad \varphi((1,-1,1,-1)) = (0,0).$$

Determine a basis for the image and for the kernel of φ.

35. Let V be the set of all 2×2 matrices with real entries and let $\varphi : V \to \mathbb{R}$ be the map defined by sending a matrix $A \in V$ to the sum of the diagonal entries of A (the *trace* of A).

(a) Show that

$$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$$

is a basis for V.

(b) Prove that φ is a linear transformation and determine the matrix of φ with respect to the basis in (a) for V. Determine the dimension of and a basis for the kernel of φ.

36. Let V be the 6-dimensional vector space over $\mathbb{Q}$ consisting of the polynomials in the variable x of degree at most 5. Let φ be the map of V to itself defined by $\varphi(f) = x^2 f'' - 6xf' + 12f$, where f'' denotes the usual second derivative (with respect to x) of the polynomial $f \in V$ and f' similarly denotes the usual first derivative.

(a) Prove that φ is a linear transformation of V to itself.

(b) Determine a basis for the image and for the kernel of φ.

37. Let V be the 7-dimensional vector space over the field F consisting of the polynomials in the variable x of degree at most 6. Let φ be the linear transformation of V to itself defined by $\varphi(f) = f'$, where f' denotes the usual derivative (with respect to x) of the polynomial $f \in V$. For each of the fields below, determine a basis for the image and for the kernel of φ:

(a) $F = \mathbb{R}$

(b) $F = \mathbb{F}_2$, the finite field of 2 elements (note that, for example, $(x^2)' = 2x = 0$ over this field)

(c) $F = \mathbb{F}_3$

(d) $F = \mathbb{F}_5$.

38. Let A and B be square matrices. Prove that the trace of their Kronecker product is the product of their traces: $\text{tr}(A \otimes B) = \text{tr}(A)\,\text{tr}(B)$. (Recall that the trace of a square matrix is the sum of its diagonal entries.)

39. Let F be a subfield of K and let $\psi : V \to W$ be a linear transformation of finite dimensional vector spaces over F.

(a) Prove that $1 \otimes \psi$ is a K–linear transformation from the vector spaces $K \otimes_F V$ to $K \otimes_F W$ over K. (Here 1 denotes the identity map from K to itself.)

(b) Let $\mathcal{B} = \{v_1, \ldots, v_n\}$ and $\mathcal{E} = \{w_1, \ldots, w_m\}$ be bases of V and W respectively. Prove that the matrix of $1 \otimes \psi$ with respect to the bases $\{1 \otimes v_1, \ldots, 1 \otimes v_n\}$ and $\{1 \otimes w_1, \ldots, 1 \otimes w_m\}$ is the same as the matrix of ψ with respect to $\mathcal{B}$ and $\mathcal{E}$.

11.3 DUAL VECTOR SPACES

Definition.

(1) For V any vector space over F let $V^* = \text{Hom}_F(V, F)$ be the space of linear transformations from V to F, called the *dual space* of V. Elements of V^* are called *linear functionals*.

(2) If $\mathcal{B} = \{v_1, v_2, \ldots, v_n\}$ is a basis of the finite dimensional space V, define $v_i^* \in V^*$ for each $i \in \{1, 2, \ldots, n\}$ by its action on the basis $\mathcal{B}$:

$$v_i^*(v_j) = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{if } i \neq j \end{cases} \qquad 1 \leq j \leq n. \tag{11.6}$$

Proposition 18. With notations as above, $\{v_1^*, v_2^*, \ldots, v_n^*\}$ is a basis of V^*. In particular, if V is finite dimensional then V^* has the same dimension as V.

Proof: Observe that since V is finite dimensional, $\dim V^* = \dim \mathrm{Hom}_F(V, F) = \dim V = n$ (Corollary 11), so since there are n of the v_i^*'s it suffices to prove that they are linearly independent. If

$$\alpha_1 v_1^* + \alpha_2 v_2^* + \cdots + \alpha_n v_n^* = 0 \quad \text{in } \mathrm{Hom}_F(V, F),$$

then applying this element to v_i and using equation (6) above we obtain $\alpha_i = 0$. Since i is arbitrary these elements are linearly independent.

Definition. The basis $\{v_1^*, v_2^*, \ldots, v_n^*\}$ of V^* is called the *dual basis* to $\{v_1, v_2, \ldots, v_n\}$.

The exercises later show that if V is infinite dimensional it is always true that $\dim V < \dim V^*$. For spaces of arbitrary dimension the space V^* is the "algebraic" dual space to V. If V has some additional structure, for example a continuous structure (i.e., a topology), then one may define other types of dual spaces (e.g., the continuous dual of V, defined by requiring the linear functionals to be *continuous* maps). One has to be careful when reading other works (particularly analysis books) to ascertain what qualifiers are implicit in the use of the terms "dual space" and "linear functional."

Example

Let $[a, b]$ be a closed interval in $\mathbb{R}$ and let V be the real vector space of all continuous functions $f : [a, b] \to \mathbb{R}$. If $a < b$, V is infinite dimensional. For each $g \in V$ the function $\varphi_g : V \to \mathbb{R}$ defined by $\varphi_g(f) = \int_a^b f(t)g(t)dt$ is a linear functional on V.

Definition. The dual of V^*, namely V^{**}, is called the *double dual* or *second dual* of V.

Note that for a finite dimensional space V, $\dim V = \dim V^*$ and also $\dim V^* = \dim V^{**}$, hence V and V^{**} are isomorphic vector spaces. For infinite dimensional spaces $\dim V < \dim V^{**}$ (cf. the exercises) so V and V^{**} cannot be isomorphic. In the case of finite dimensional spaces there is a *natural*, i.e., basis independent or coordinate free way of exhibiting the isomorphism between a vector space and its second dual. The basic idea, in a more general setting, is as follows: if X is any set and S is any set of functions of X into the field F, we normally think of choosing or fixing an $f \in S$ and computing $f(x)$ as x ranges over all of X. Alternatively, we could think of fixing a point x in X and computing $f(x)$ as f ranges over all of S. The latter process, called *evaluation at* x shows that for each $x \in X$ there is a function $E_x : S \to F$ defined by

$E_x(f) = f(x)$ (i.e., evaluate f at x). This gives a map $x \mapsto E_x$ of X into the set of F-valued functions on S. If S "separates points" in the sense that for distinct points x and y of X there is some $f \in S$ such that $f(x) \neq f(y)$, then the map $x \mapsto E_x$ is injective. The proof of the next lemma applies this "role reversal" process to the situation where $X = V$ and $S = V^*$, proves E_x is a linear F-valued function on S, that is, E_x belongs to the dual space of V^*, and proves the map $x \mapsto E_x$ is a linear transformation from V into V^{**}. Note that throughout this process there is no mention of the word "basis" (although it is convenient to know the dimension of V^{**} — a fact we established by picking bases). In particular, the proof does not start with the familiar phrase "pick a basis of $V \ldots$."

Theorem 19. There is a natural injective linear transformation from V to V^{**}. If V is finite dimensional then this linear transformation is an isomorphism.

Proof: Let $v \in V$. Define the map (*evaluation at v*)

$$E_v : V^* \to F \qquad \text{by} \qquad E_v(f) = f(v).$$

Then $E_v(f + \alpha g) = (f + \alpha g)(v) = f(v) + \alpha g(v) = E_v(f) + \alpha E_g(v)$, so that E_v is a linear transformation from V^* to F. Hence E_v is an element of $\mathrm{Hom}_F(V^*, F) = V^{**}$. This defines a natural map

$$\varphi : V \to V^{**} \qquad \text{by} \qquad \varphi(v) = E_v.$$

The map φ is a *linear* map, as follows: for $v, w \in V$ and $\alpha \in F$,

$$E_{v+\alpha w}(f) = f(v + \alpha w) = f(v) + \alpha f(w) = E_v(f) + \alpha E_w(f)$$

for every $f \in V^*$, and so

$$\varphi(v + \alpha w) = E_{v+\alpha w} = E_v + \alpha E_w = \varphi(v) + \alpha\varphi(w).$$

To see that φ is injective let v be any nonzero vector in V. By the Building Up Lemma there is a basis $\mathcal{B}$ containing v. Let f be the linear transformation from V to F defined by sending v to 1 and every element of $\mathcal{B} - \{v\}$ to zero. Then $f \in V^*$ and $E_v(f) = f(v) = 1$. Thus $\varphi(v) = E_v$ is not zero in V^{**}. This proves $\ker \varphi = 0$, i.e., φ is injective.

If V has finite dimension n then by Proposition 18, V^* and hence also V^{**} has dimension n. In this case φ is an injective linear transformation from V to a finite dimensional vector space of the same dimension, hence is an isomorphism.

Let V, W be finite dimensional vector spaces over F with bases $\mathcal{B}$, $\mathcal{E}$, respectively and let $\mathcal{B}^*$, $\mathcal{E}^*$ be the dual bases. Fix some $\varphi \in \mathrm{Hom}_F(V, W)$. Then for each $f \in W^*$, the composite $f \circ \varphi$ is a linear transformation from V to F, that is $f \circ \varphi \in V^*$. Thus the map $f \mapsto f \circ \varphi$ defines a function from W^* to V^*. We denote this induced function on dual spaces by φ^*.

Theorem 20. With notations as above, φ^* is a linear transformation from W^* to V^* and $M_{\mathcal{E}^*}^{\mathcal{B}^*}(\varphi^*)$ is the transpose of the matrix $M_{\mathcal{B}}^{\mathcal{E}}(\varphi)$ (recall that the transpose of the matrix (a_{ij}) is the matrix (a_{ji})).

Proof: The map φ^* is linear because $(f + \alpha g) \circ \varphi = (f \circ \varphi) + \alpha(g \circ \varphi)$. The equations which define φ are (from its matrix)

$$\varphi(v_j) = \sum_{i=1}^{m} \alpha_{ij} w_i \qquad 1 \leq j \leq n.$$

To compute the matrix for φ^*, observe that by the definitions of φ^* and w_k^*

$$\varphi^*(w_k^*)(v_j) = (w_k^* \circ \varphi)(v_j) = w_k^*\left(\sum_{i=1}^{m} \alpha_{ij} w_i\right) = \alpha_{kj}.$$

Also

$$\left(\sum_{i=1}^{n} \alpha_{ki} v_i^*\right)(v_j) = \alpha_{kj}$$

for all j. This shows that the two linear functionals below agree on a basis of V, hence they are the same element of V^*:

$$\varphi^*(w_k^*) = \sum_{i=1}^{n} \alpha_{ki} v_i^*.$$

This determines the matrix for φ^* with respect to the bases $\mathcal{E}^*$ and $\mathcal{B}^*$ as the transpose of the matrix for φ.

Corollary 21. For any matrix A, the row rank of A equals the column rank of A.

Proof: Let $\varphi : V \to W$ be a linear transformation whose matrix with respect to some fixed bases of V and W is A. By Theorem 20 the matrix of $\varphi^* : W^* \to V^*$ with respect to the dual bases is the transpose of A. The column rank of A is the rank of φ and the row rank of A (= the column rank of the transpose of A) is the rank of φ^* (cf. Exercise 6 of Section 2). It therefore suffices to show that φ and φ^* have the same rank. Now

$$\begin{aligned} f \in \ker \varphi^* &\Leftrightarrow \varphi^*(f) = 0 \Leftrightarrow f \circ \varphi(v) = 0, \quad \text{for all } v \in V \\ &\Leftrightarrow \varphi(V) \subseteq \ker f \Leftrightarrow f \in \operatorname{Ann}(\varphi(V)), \end{aligned}$$

where $\operatorname{Ann}(S)$ is the annihilator of S described in Exercise 3 below. Thus $\operatorname{Ann}(\varphi(V)) = \ker \varphi^*$. By Exercise 3, $\dim \operatorname{Ann}(\varphi(V)) = \dim W - \dim \varphi(V)$. By Corollary 8, $\dim \ker \varphi^* = \dim W^* - \dim \varphi^*(W^*)$. Since W and W^* have the same dimension, $\dim \varphi(V) = \dim \varphi^*(W^*)$ as needed.

EXERCISES

1. Let V be a finite dimensional vector space. Prove that the map $\varphi \mapsto \varphi^*$ in Theorem 20 gives a ring isomorphism of $\text{End}(V)$ with $\text{End}(V^*)$.
2. Let V be the collection of polynomials with coefficients in $\mathbb{Q}$ in the variable x of degree at most 5 with $1, x, x^2, \ldots, x^5$ as basis. Prove that the following are elements of the dual space of V and express them as linear combinations of the dual basis:
   (a) $E : V \to \mathbb{Q}$ defined by $E(p(x)) = p(3)$ (i.e., evaluation at $x = 3$).
   (b) $\varphi : V \to \mathbb{Q}$ defined by $\varphi(p(x)) = \int_0^1 p(t)dt$.
   (c) $\varphi : V \to \mathbb{Q}$ defined by $\varphi(p(x)) = \int_0^1 t^2 p(t)dt$.
   (d) $\varphi : V \to \mathbb{Q}$ defined by $\varphi(p(x)) = p'(5)$ where $p'(x)$ denotes the usual derivative of the polynomial $p(x)$ with respect to x.
3. Let S be any subset of V^* for some finite dimensional space V. Define $\text{Ann}(S) = \{v \in V \mid f(v) = 0 \text{ for all } f \in S\}$. ($\text{Ann}(S)$ is called the *annihilator of S in V*).
   (a) Prove that $\text{Ann}(S)$ is a subspace of V.
   (b) Let W_1 and W_2 be subspaces of V^*. Prove that $\text{Ann}(W_1 + W_2) = \text{Ann}(W_1) \cap \text{Ann}(W_2)$ and $\text{Ann}(W_1 \cap W_2) = \text{Ann}(W_1) + \text{Ann}(W_2)$.
   (c) Let W_1 and W_2 be subspaces of V^*. Prove that $W_1 = W_2$ if and only if $\text{Ann}(W_1) = \text{Ann}(W_2)$.
   (d) Prove that the annihilator of S is the same as the annihilator of the subspace of V^* spanned by S.
   (e) Assume V is finite dimensional with basis $v_1, \ldots, v_n$. Prove that if $S = \{v_1^*, \ldots, v_k^*\}$ for some $k \leq n$, then $\text{Ann}(S)$ is the subspace spanned by $\{v_{k+1}, \ldots, v_n\}$.
   (f) Assume V is finite dimensional. Prove that if W^* is any subspace of V^* then $\dim \text{Ann}(W^*) = \dim V - \dim W^*$.
4. If V is infinite dimensional with basis $\mathcal{A}$, prove that $\mathcal{A}^* = \{v^* \mid v \in \mathcal{A}\}$ does *not* span V^*.
5. If V is infinite dimensional with basis $\mathcal{A}$, prove that V^* is isomorphic to the direct product of copies of F indexed by $\mathcal{A}$. Deduce that $\dim V^* > \dim V$. [Use Exercise 14, Section 1.]

11.4 DETERMINANTS

Although we shall be using the theory primarily for vector spaces over a field, the theory of determinants can be developed with no extra effort over arbitrary commutative rings with 1. Thus in this section R is any commutative ring with 1 and $V_1, V_2, \ldots, V_n, V$ and W are R-modules. For convenience we repeat the definition of multilinear functions from Section 10.4.

Definition.

(1) A map $\varphi : V_1 \times V_2 \times \cdots \times V_n \to W$ is called *multilinear* if for each fixed i and fixed elements $v_j \in V_j$, $j \neq i$, the map

$$V_i \to W \qquad \text{defined by} \qquad x \mapsto \varphi(v_1, \ldots, v_{i-1}, x, v_{i+1}, \ldots, v_n)$$

is an R-module homomorphism. If $V_i = V$, $i = 1, 2, \ldots, n$, then φ is called an *n-multilinear function on V*, and if in addition $W = R$, φ is called an *n-multilinear form on V*.

(2) An n-multilinear function φ on V is called *alternating* if $\varphi(v_1, v_2, \ldots, v_n) = 0$ whenever $v_i = v_{i+1}$ for some $i \in \{1, 2, \ldots, n-1\}$ (i.e., φ is zero whenever two consecutive arguments are equal). The function φ is called *symmetric* if interchanging v_i and v_j for any i and j in $(v_1, v_2, \ldots, v_n)$ does not alter the value of φ on this n-tuple.

When $n = 2$ (respectively, 3) one says φ is *bilinear* (respectively, *trilinear*) rather than 2-multilinear (respectively, 3-multilinear). Also, when n is clear from the context we shall simply say φ is multilinear.

Example

For any fixed $m \geq 0$ the usual dot product on $V = \mathbb{R}^m$ is a bilinear form (here the ring R is the field of real numbers).

Proposition 22. Let φ be an n-multilinear alternating function on V. Then

(1) $\varphi(v_1, \ldots, v_{i-1}, v_{i+1}, v_i, v_{i+2}, \ldots, v_n) = -\varphi(v_1, v_2, \ldots, v_n)$ for any $i \in \{1, 2, \ldots, n-1\}$, i.e., the value of φ on an n-tuple is negated if two adjacent components are interchanged.

(2) For each $\sigma \in S_n$, $\varphi(v_{\sigma(1)}, v_{\sigma(2)}, \ldots, v_{\sigma(n)}) = \epsilon(\sigma)\varphi(v_1, v_2, \ldots, v_n)$, where $\epsilon(\sigma)$ is the sign of the permutation σ (cf. Section 3.5).

(3) If $v_i = v_j$ for any pair of distinct $i, j \in \{1, 2, \ldots, n\}$ then $\varphi(v_1, v_2, \ldots, v_n) = 0$.

(4) If v_i is replaced by $v_i + \alpha v_j$ in $(v_1, \ldots, v_n)$ for any $j \neq i$ and any $\alpha \in R$, the value of φ on this n-tuple is not changed.

Proof: (1) Let $\psi(x, y)$ be the function φ with variable entries x and y in positions i and $i+1$ respectively and fixed entries v_j in position j, for all other j. Thus (1) is the same as showing $\psi(y, x) = -\psi(x, y)$. Since φ is alternating $\psi(x + y, x + y) = 0$. Expanding $x + y$ in each variable in turn gives $\psi(x+y, x+y) = \psi(x, x)+\psi(x, y)+\psi(y, x) + \psi(y, y)$. Again, by the alternating property of φ, the first and last terms on the right hand side of the latter equation are zero. Thus $0 = \psi(x, y) + \psi(y, x)$, which gives (1).

(2) Every permutation can be written as a product of transpositions (cf. Section 3.5). Furthermore, every transposition may be written as a product of transpositions which interchange two successive integers (cf. Exercise 3 of Section 3.5). Thus every permutation σ can be written as $\tau_1 \cdots \tau_m$, where τ_k is a transposition interchanging two successive integers, for all k. It follows from m applications of (1) that

$$\varphi(v_{\sigma(1)}, v_{\sigma(2)}, \ldots, v_{\sigma(n)}) = \epsilon(\tau_m) \cdots \epsilon(\tau_1)\varphi(v_1, v_2, \ldots, v_n).$$

Finally, since ϵ is a homomorphism into the abelian group ± 1 (so the order of the factors ± 1 does not matter), $\epsilon(\tau_1) \cdots \epsilon(\tau_m) = \epsilon(\tau_1 \cdots \tau_m) = \epsilon(\sigma)$. This proves (2).

(3) Choose σ to be any permutation which fixes i and moves j to $i + 1$. Thus $(v_{\sigma(1)}, v_{\sigma(2)}, \ldots, v_{\sigma(n)})$ has two equal adjacent components so φ is zero on this n-tuple. By (2), $\varphi(v_{\sigma(1)}, v_{\sigma(2)}, \ldots, v_{\sigma(n)}) = \pm\varphi(v_1, v_2, \ldots, v_n)$. This implies (3).

(4) This follows immediately from (3) on expanding by linearity in the i^{th} position.

Proposition 23. Assume φ is an n-multilinear alternating function on V and that for some $v_1, v_2, \ldots, v_n$ and $w_1, w_2, \ldots, w_n \in V$ and some $\alpha_{ij} \in R$ we have

$$\begin{aligned} w_1 &= \alpha_{11}v_1 + \alpha_{21}v_2 + \cdots + \alpha_{n1}v_n \\ w_2 &= \alpha_{12}v_1 + \alpha_{22}v_2 + \cdots + \alpha_{n2}v_n \\ &\vdots \\ w_n &= \alpha_{1n}v_1 + \alpha_{2n}v_2 + \cdots + \alpha_{nn}v_n \end{aligned}$$

(we have purposely written the indices of the α_{ij} in "column format"). Then

$$\varphi(w_1, w_2, \ldots, w_n) = \sum_{\sigma \in S_n} \epsilon(\sigma)\alpha_{\sigma(1)1}\alpha_{\sigma(2)2} \cdots \alpha_{\sigma(n)n}\varphi(v_1, v_2, \ldots, v_n).$$

Proof: If we expand $\varphi(w_1, w_2, \ldots, w_n)$ by multilinearity we obtain a sum of n^n terms of the form $\alpha_{i_1 1}\alpha_{i_2 2} \ldots \alpha_{i_n n}\varphi(v_{i_1}, v_{i_2}, \ldots, v_{i_n})$, where the indices $i_1, i_2, \ldots, i_n$ each run over $1, 2, \ldots, n$. By Proposition 22(3), φ is zero on the terms where two or more of the i_j's are equal. Thus in this expansion we need only consider the terms where $i_1, \ldots, i_n$ are distinct. Such sequences are in bijective correspondence with permutations in S_n, so each nonzero term may be written as $\alpha_{\sigma(1)1}\alpha_{\sigma(2)2} \cdots \alpha_{\sigma(n)n}\varphi(v_{\sigma(1)}, v_{\sigma(2)}, \ldots, v_{\sigma(n)})$, for some $\sigma \in S_n$. Applying (2) of the previous proposition to each of these terms in the expansion of $\varphi(w_1, w_2, \ldots, w_n)$ gives the expression in the proposition.

Definition. An $n \times n$ *determinant function* on R is any function

$$\det : M_{n\times n}(R) \to R$$

that satisfies the following two axioms:

(1) det is an n-multilinear alternating form on $R^n (= V)$, where the n-tuples are the n columns of the matrices in $M_{n\times n}(R)$

(2) $\det(I) = 1$, where I is the $n \times n$ identity matrix.

On occasion we shall write $\det(A_1, A_2, \ldots, A_n)$ for $\det A$, where $A_1, A_2, \ldots, A_n$ are the columns of A.

Theorem 24. There is a unique $n \times n$ determinant function on R and it can be computed for any $n \times n$ matrix (α_{ij}) by the formula:

$$\det(\alpha_{ij}) = \sum_{\sigma \in S_n} \epsilon(\sigma)\alpha_{\sigma(1)1}\alpha_{\sigma(2)2} \cdots \alpha_{\sigma(n)n}.$$

Proof: Let $A_1, A_2, \ldots, A_n$ be the column vectors in a general $n \times n$ matrix (α_{ij}). We leave it as an exercise to check that the formula given in the statement of the theorem does satisfy the axioms of a determinant function — this gives existence of a determinant

function. To prove uniqueness let e_i be the column n-tuple with 1 in position i and zeros in all other positions. Then

$$\begin{aligned} A_1 &= \alpha_{11}e_1 + \alpha_{21}e_2 + \cdots + \alpha_{n1}e_n \\ A_2 &= \alpha_{12}e_1 + \alpha_{22}e_2 + \cdots + \alpha_{n2}e_n \\ &\vdots \\ A_n &= \alpha_{1n}e_1 + \alpha_{2n}e_2 + \cdots + \alpha_{nn}e_n. \end{aligned}$$

By Proposition 23, $\det A = \sum_{\sigma \in S_n} \epsilon(\sigma)\alpha_{\sigma(1)1}\alpha_{\sigma(2)2}\cdots\alpha_{\sigma(n)n} \det(e_1, e_2, \ldots, e_n)$. Since by axiom (2) of a determinant function $\det(e_1, e_2, \ldots, e_n) = 1$, the value of $\det A$ is as claimed.

Corollary 25. The determinant is an n-multilinear function of the rows of $M_{n\times n}(R)$ and for any $n \times n$ matrix A, $\det A = \det(A^t)$, where A^t is the transpose of A.

Proof: The first statement is an immediate consequence of the second, so it suffices to prove that a matrix and its transpose have the same determinant. For $A = (\alpha_{ij})$ one calculates that

$$\det A^t = \sum_{\sigma \in S_n} \epsilon(\sigma)\alpha_{1\sigma(1)}\alpha_{2\sigma(2)}\ldots\alpha_{n\sigma(n)}.$$

Each number from 1 to n appears exactly once among $\sigma(1), \ldots, \sigma(n)$ so we may rearrange the product $\alpha_{1\sigma(1)}\alpha_{2\sigma(2)}\ldots\alpha_{n\sigma(n)}$ as $\alpha_{\sigma^{-1}(1)1}\alpha_{\sigma^{-1}(2)2}\ldots\alpha_{\sigma^{-1}(n)n}$. Also, the homomorphism ϵ takes values in $\{\pm 1\}$ so $\epsilon(\sigma) = \epsilon(\sigma^{-1})$. Thus the sum for $\det A^t$ may be rewritten as

$$\sum_{\sigma \in S_n} \epsilon(\sigma^{-1})\alpha_{\sigma^{-1}(1)1}\alpha_{\sigma^{-1}(2)2}\ldots\alpha_{\sigma^{-1}(n)n}.$$

The latter sum is over all permutations, so the index σ^{-1} may be replaced by σ. The resulting expression is the sum for $\det A$. This completes the proof.

Theorem 26. *(Cramer's Rule)* If $A_1, A_2, \ldots, A_n$ are the columns of an $n \times n$ matrix A and $B = \beta_1 A_1 + \beta_2 A_2 + \cdots + \beta_n A_n$, for some $\beta_1, \ldots, \beta_n \in R$, then

$$\beta_i \det A = \det(A_1, \ldots, A_{i-1}, B, A_{i+1}, \ldots, A_n).$$

Proof: This follows immediately from Proposition 22(3) on replacing the given expression for B in the i^{th} position and expanding by multilinearity in that position.

Corollary 27. If R is an integral domain, then $\det A = 0$ for $A \in M_n(R)$ if and only if the columns of A are R-linearly dependent as elements of the free R-module of rank n. Also, $\det A = 0$ if and only if the rows of A are R-linearly dependent.

Proof: Since $\det A = \det A^t$ the first sentence implies the second.
Assume first that the columns of A are linearly dependent and

$$0 = \beta_1 A_1 + \beta_2 A_2 + \cdots + \beta_n A_n$$

is a dependence relation on the columns of A with, say, $\beta_i \neq 0$. By Cramer's Rule, $\beta_i \det A = 0$. Since R is an integral domain and $\beta_i \neq 0$, $\det A = 0$.

Conversely, assume the columns of A are independent. Consider the integral domain R as embedded in its quotient field F so that $M_{n\times n}(R)$ may be considered as a subring of $M_{n\times n}(F)$ (and note that the determinant function on the subring is the restriction of the determinant function from $M_{n\times n}(F)$). The columns of A in this way become elements of F^n. Any nonzero F-linear combination of the columns of A which is zero in F^n gives, by multiplying the coefficients by a common denominator, a nonzero R-linear dependence relation. The columns of A must therefore be independent vectors in F^n. Since A has n columns, these form a basis of F^n. Thus there are elements β_{ij} of F such that for each i, the i^{th} basis vector e_i in F^n may be expressed as

$$e_i = \beta_{1i}A_1 + \beta_{2i}A_2 + \cdots + \beta_{ni}A_n.$$

The $n \times n$ identity matrix is the one whose columns are $e_1, e_2, \ldots, e_n$. By Proposition 23 (with $\varphi = \det$), the determinant of the identity matrix is some F-multiple of $\det A$. Since the determinant of the identity matrix is 1, $\det A$ cannot be zero. This completes the proof.

Theorem 28. For matrices $A, B \in M_{n\times n}(R)$, $\det AB = (\det A)(\det B)$.

Proof: Let $B = (\beta_{ij})$ and let $A_1, A_2, \ldots, A_n$ be the columns of A. Then $C = AB$ is the $n \times n$ matrix whose j^{th} column is $C_j = \beta_{1j}A_1 + \beta_{2j}A_2 + \cdots + \beta_{nj}A_n$. By Proposition 23 applied to the multilinear function det we obtain

$$\det C = \det(C_1, \ldots, C_n) = \left[\sum_{\sigma\in S_n} \epsilon(\sigma)\beta_{\sigma(1)\,1}\beta_{\sigma(2)\,2}\cdots\beta_{\sigma(n)\,n}\right] \det(A_1, \ldots, A_n).$$

The sum inside the brackets is the formula for $\det B$, hence $\det C = (\det B)(\det A)$, as required ($R$ is commutative).

Definition. Let $A = (\alpha_{ij})$ be an $n \times n$ matrix. For each i, j, let A_{ij} be the $n-1 \times n-1$ matrix obtained from A by deleting its i^{th} row and j^{th} column (an $n-1 \times n-1$ *minor* of A). Then $(-1)^{i+j} \det(A_{ij})$ is called the ij *cofactor of* A.

Theorem 29. *(The Cofactor Expansion Formula along the i^{th} row)* If $A = (\alpha_{ij})$ is an $n \times n$ matrix, then for each fixed $i \in \{1, 2, \ldots, n\}$ the determinant of A can be computed from the formula

$$\det A = (-1)^{i+1}\alpha_{i1} \det A_{i1} + (-1)^{i+2}\alpha_{i2} \det A_{i2} + \cdots + (-1)^{i+n}\alpha_{in} \det A_{in}.$$

Proof: For each A let $D(A)$ be the element of R obtained from the cofactor expansion formula described above. We prove that D satisfies the axioms of a determinant function, hence is *the* determinant function. Proceed by induction on n. If $n = 1$, $D((\alpha)) = \alpha$, for all 1×1 matrices (α) and the result holds. Assume therefore that $n \geq 2$. To show that D is an alternating multilinear function of the columns, fix an index k and consider the k^{th} column as varying and all other columns as fixed. If $j \neq k$,

α_{ij} does not depend on k and $D(A_{ij})$ is linear in the k^{th} column by induction. Also, as the k^{th} column varies linearly so does α_{ik}, whereas $D(A_{ik})$ remains unchanged (the k^{th} column has been deleted from A_{ik}). Thus each term in the formula for D varies linearly in the k^{th} column. This proves D is multilinear in the columns.

To prove D is alternating assume columns k and $k+1$ of A are equal. If $j \neq k$ or $k+1$, the two equal columns of A become two equal columns in the matrix A_{ij}. By induction $D(A_{ij}) = 0$. The formula for D therefore has at most two nonzero terms: when $j = k$ and when $j = k+1$. The minor matrices A_{ik} and $A_{i\,k+1}$ are identical and $\alpha_{ik} = \alpha_{i\,k+1}$. Then the two remaining terms in the expansion for D, $(-1)^{i+k}\alpha_{ik}D(A_{ik})$ and $(-1)^{i+k+1}\alpha_{i\,k+1}D(A_{i\,k+1})$ are equal and appear with opposite signs, hence they cancel. Thus $D(A) = 0$ if A has two adjacent columns which are equal, i.e., D is alternating.

Finally, it follows easily from the formula and induction that $D(I) = 1$, where I is the identity matrix. This completes the induction.

Theorem 30. *(Cofactor Formula for the Inverse of a Matrix)* Let $A = (\alpha_{ij})$ be an $n \times n$ matrix and let B be the transpose of its matrix of cofactors, i.e., $B = (\beta_{ij})$, where $\beta_{ij} = (-1)^{i+j} \det A_{ji}$, $1 \leq i, j \leq n$. Then $AB = BA = (\det A)I$. Moreover, $\det A$ is a unit in R if and only if A is a unit in $M_{n\times n}(R)$; in this case the matrix $\dfrac{1}{\det A}B$ is the inverse of A.

Proof: The i, j entry of AB is $\alpha_{i1}\beta_{1j} + \alpha_{i2}\beta_{2j} + \cdots + \alpha_{in}\beta_{nj}$. By definition of the entries of B this equals

$$\alpha_{i1}(-1)^{j+1}D(A_{j1}) + \alpha_{i2}(-1)^{j+2}D(A_{j2}) + \cdots + \alpha_{in}(-1)^{j+n}D(A_{jn}). \qquad (11.7)$$

If $i = j$, this is the cofactor expansion for $\det A$ along the i^{th} row. The diagonal entries of AB are thus all equal to $\det A$. If $i \neq j$, let $\overline{A}$ be the matrix A with the j^{th} row replaced by the i^{th} row, so $\det \overline{A} = 0$. By inspection $\overline{A}_{jk} = A_{jk}$ and $\alpha_{ik} = \overline{\alpha}_{jk}$ for every $k \in \{1, 2, \ldots, n\}$. By making these substitutions in equation (7) for each $k = 1, 2, \ldots, n$ one sees that the i, j entry in AB equals $\overline{\alpha}_{j1}(-1)^{1+j}D(\overline{A}_{j1}) + \cdots + \overline{\alpha}_{jn}(-1)^{n+j}D(\overline{A}_{jn})$. This expression is the cofactor expansion for $\det \overline{A}$ along the j^{th} row. Since, as noted above, $\det \overline{A} = 0$, this proves that all off diagonal terms of AB are zero, which proves that $AB = (\det A)I$.

It follows directly from the definition of B that the pair (A^t, B^t) satisfies the same hypotheses as the pair (A, B). By what has already been shown it follows that $(BA)^t = A^tB^t = (\det A^t)I$. Since $\det A^t = \det A$ and the transpose of a diagonal matrix is itself, we obtain $BA = (\det A)I$ as well.

If $d = \det A$ is a unit in R, then $d^{-1}B$ is a matrix with entries in R whose product with A (on either side) is the identity, i.e., A is a unit in $M_{n\times n}(R)$. Conversely, assume that A is a unit in R with (2-sided) inverse matrix C. Since $\det C \in R$ and

$$1 = \det I = \det AC = (\det A)(\det C) = (\det C)(\det A),$$

it follows that $\det A$ has a 2-sided inverse in R, as needed. This completes all parts of the proof.

EXERCISES

1. Formulate and prove the cofactor expansion formula along the j^{th} column of a square matrix A.
2. Let F be a field and let $A_1, A_2, \ldots, A_n$ be (column) vectors in F^n. Form the matrix A whose i^{th} column is A_i. Prove that these vectors form a basis of F^n if and only if $\det A \neq 0$.
3. Let R be any commutative ring with 1, let V be an R-module and let $x_1, x_2, \ldots, x_n \in V$. Assume that for some $A \in M_{n\times n}(R)$,
$$A\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = 0.$$
Prove that $(\det A)x_i = 0$, for all $i \in \{1, 2, \ldots, n\}$.
4. (*Computing Determinants of Matrices*) This exercise outlines the use of Gauss–Jordan elimination (cf. the exercises in Section 2) to compute determinants. This is the most efficient general procedure for computing large determinants. Let A be an $n \times n$ matrix.
   (a) Prove that the elementary row operations have the following effect on determinants:
      (i) interchanging two rows changes the sign of the determinant
      (ii) adding a multiple of one row to another does not alter the determinant
      (iii) multiplying any row by a nonzero element u from F multiplies the determinant by u.
   (b) Prove that $\det A$ is nonzero if and only if A is row equivalent to the $n \times n$ identity matrix. Suppose A can be row reduced to the identity matrix using a total of s row interchanges as in (i) and by multiplying rows by the nonzero elements $u_1, u_2, \ldots, u_t$ as in (iii). Prove that $\det A = (-1)^s(u_1u_2 \ldots u_t)^{-1}$.
5. Compute the determinants of the following matrices using row reduction:
$$A = \begin{pmatrix} 5 & 4 & -6 \\ -2 & 0 & 2 \\ 3 & 4 & -2 \end{pmatrix} \qquad B = \begin{pmatrix} 1 & 2 & -4 & 4 \\ 2 & -1 & 4 & -8 \\ 1 & 0 & 1 & -2 \\ 0 & 1 & -2 & 3 \end{pmatrix}.$$
6. (*Minkowski's Criterion*) Suppose A is an $n \times n$ matrix with real entries such that the diagonal elements are all positive, the off-diagonal elements are all negative and the row sums are all positive. Prove that $\det A \neq 0$. [Consider the corresponding system of equations $AX = 0$ and suppose there is a nontrivial solution $(x_1, \ldots, x_n)$. If x_i has the largest absolute value show that the i^{th} equation leads to a contradiction.]

11.5 TENSOR ALGEBRAS, SYMMETRIC AND EXTERIOR ALGEBRAS

In this section R is any commutative ring with 1, and we assume the left and right actions of R on each R-module are the same. We shall primarily be interested in the special case when $R = F$ is a field, but the basic constructions hold in general.

Suppose M is an R-module. When tensor products were first introduced in Section 10.4 we spoke heuristically of forming "products" m_1m_2 of elements of M, and we constructed a new module $M \otimes M$ generated by such "products" $m_1 \otimes m_2$. The "value" of this product is not in M, so this does not give a ring structure on M itself. If, however,

we iterate this by taking the "products" $m_1m_2m_3$ and $m_1m_2m_3m_4$, and all finite sums of such products, we can construct a ring containing M that is "universal" with respect to rings containing M (and, more generally, with respect to homomorphic images of M), as we now show.

For each integer $k \geq 1$, define

$$\mathcal{T}^k(M) = M \otimes_R M \otimes_R \cdots \otimes_R M \quad (k \text{ factors}),$$

and set $\mathcal{T}^0(M) = R$. The elements of $\mathcal{T}^k(M)$ are called *k-tensors*. Define

$$\mathcal{T}(M) = R \oplus \mathcal{T}^1(M) \oplus \mathcal{T}^2(M) \oplus \mathcal{T}^3(M) \cdots = \bigoplus_{k=0}^{\infty} \mathcal{T}^k(M).$$

Every element of $\mathcal{T}(M)$ is a finite linear combination of k-tensors for various $k \geq 0$. We identify M with $\mathcal{T}^1(M)$, so that M is an R-submodule of $\mathcal{T}(M)$.

Theorem 31. If M is any R-module over the commutative ring R then

(1) $\mathcal{T}(M)$ is an R-algebra containing M with multiplication defined by mapping

$$(m_1 \otimes \cdots \otimes m_i)(m'_1 \otimes \cdots \otimes m'_j) = m_1 \otimes \cdots \otimes m_i \otimes m'_1 \otimes \cdots \otimes m'_j$$

and extended to sums via the distributive laws. With respect to this multiplication $\mathcal{T}^i(M)\mathcal{T}^j(M) \subseteq \mathcal{T}^{i+j}(M)$.

(2) *(Universal Property)* If A is any R-algebra and $\varphi : M \to A$ is an R-module homomorphism, then there is a unique R-algebra homomorphism $\Phi : \mathcal{T}(M) \to A$ such that $\Phi|_M = \varphi$.

Proof: The map

$$\underbrace{M \times M \times \cdots \times M}_{i \text{ factors}} \times \underbrace{M \times M \times \cdots \times M}_{j \text{ factors}} \to \mathcal{T}^{i+j}(M)$$

defined by

$$(m_1, \ldots, m_i, m'_1, \ldots, m'_j) \mapsto m_1 \otimes \ldots \otimes m_i \otimes m'_1 \otimes \ldots \otimes m'_j$$

is R-multilinear, so induces a bilinear map $\mathcal{T}^i(M) \times \mathcal{T}^j(M)$ to $\mathcal{T}^{i+j}(M)$ which is easily checked to give a well defined multiplication satisfying (1) (cf. the proof of Proposition 21 in Section 10.4). To prove (2), assume that $\varphi : M \to A$ is an R-algebra homomorphism. Then

$$(m_1, m_2, \ldots, m_k) \mapsto \varphi(m_1)\varphi(m_2)\ldots\varphi(m_k)$$

defines an R-multilinear map from $M \times \cdots \times M$ (k times) to A. This in turn induces a unique R-module homomorphism Φ from $\mathcal{T}^k(M)$ to A (Corollary 16 of Section 10.4) mapping $m_1 \otimes \ldots \otimes m_k$ to the element on the right hand side above. It is easy to check from the definition of the multiplication in (1) that the resulting uniquely defined map $\Phi : \mathcal{T}(M) \to A$ is an R-algebra homomorphism.

Definition. The ring $\mathcal{T}(M)$ is called the *tensor algebra* of M.

Proposition 32. Let V be a finite dimensional vector space over the field F with basis $\mathcal{B} = \{v_1, \ldots, v_n\}$. Then the k-tensors

$$v_{i_1} \otimes v_{i_2} \otimes \cdots \otimes v_{i_k} \qquad \text{with } v_{i_j} \in \mathcal{B}$$

are a vector space basis of $\mathcal{T}^k(V)$ over F (with the understanding that the basis vector is the element $1 \in F$ when $k = 0$). In particular, $\dim_F(\mathcal{T}^k(V)) = n^k$.

Proof: This follows immediately from Proposition 16 of Section 2.

Theorem 31 and Proposition 32 show that the space $\mathcal{T}(V)$ may be regarded as the *noncommutative polynomial algebra* over F in the (noncommuting) variables $v_1, \ldots, v_n$. The analogous result also holds for finitely generated free modules over any commutative ring (using Corollary 19 in Section 10.4).

Examples

(1) Let $R = \mathbb{Z}$ and let $M = \mathbb{Q}/\mathbb{Z}$. Then $(\mathbb{Q}/\mathbb{Z}) \otimes_{\mathbb{Z}} (\mathbb{Q}/\mathbb{Z}) = 0$ (Example 4 following Corollary 12 in Section 10.4). Thus $\mathcal{T}(\mathbb{Q}/\mathbb{Z}) = \mathbb{Z} \oplus (\mathbb{Q}/\mathbb{Z})$, where addition is componentwise and the multiplication is given by $(r, \overline{p})(s, \overline{q}) = (rs, \overline{rq + sp})$. The ring $R/(x)$ of Exercise 4(d) in Section 9.3 is isomorphic to $\mathcal{T}(\mathbb{Q}/\mathbb{Z})$.

(2) Let $R = \mathbb{Z}$ and let $M = \mathbb{Z}/n\mathbb{Z}$. Then $(\mathbb{Z}/n\mathbb{Z}) \otimes_{\mathbb{Z}} (\mathbb{Z}/n\mathbb{Z}) \cong \mathbb{Z}/n\mathbb{Z}$ (Example 3 following Corollary 12 in Section 10.4). Thus $\mathcal{T}^i(M) \cong M$ for all $i > 0$ and so $\mathcal{T}(\mathbb{Z}/n\mathbb{Z}) \cong \mathbb{Z} \oplus (\mathbb{Z}/n\mathbb{Z}) \oplus (\mathbb{Z}/n\mathbb{Z}) \cdots$. It follows easily that $\mathcal{T}(\mathbb{Z}/n\mathbb{Z}) \cong \mathbb{Z}[x]/(nx)$.

Since $\mathcal{T}^i(M)\mathcal{T}^j(M) \subseteq \mathcal{T}^{i+j}(M)$, the tensor algebra $\mathcal{T}(M)$ has a natural "grading" or "degree" structure reminiscent of a polynomial ring.

Definition.

(1) A ring S is called a *graded ring* if it is the direct sum of additive subgroups: $S = S_0 \oplus S_1 \oplus S_2 \oplus \cdots$ such that $S_i S_j \subseteq S_{i+j}$ for all $i, j \geq 0$. The elements of S_k are said to be *homogeneous of degree k*, and S_k is called the *homogeneous component of S of degree k*.

(2) An ideal I of the graded ring S is called a *graded ideal* if $I = \oplus_{k=0}^{\infty}(I \cap S_k)$.

(3) A ring homomorphism $\varphi : S \to T$ between two graded rings is called a *homomorphism of graded rings* if it respects the grading structures on S and T, i.e., if $\varphi(S_k) \subseteq T_k$ for $k = 0, 1, 2, \ldots$.

Note that $S_0 S_0 \subseteq S_0$, which implies that S_0 is a subring of the graded ring S and then S is an S_0-module. If S_0 is in the center of S and it contains an identity of S, then S is an S_0-algebra. Note also that the ideal I is graded if whenever a sum $i_{k_1} + \cdots + i_{k_n}$ of homogeneous elements with distinct degrees $k_1, \ldots, k_n$ is in I then each of the individual summands $i_{k_1}, \ldots, i_{k_n}$ is itself in I.

Example

The polynomial ring $S = R[x_1, x_2, \ldots, x_n]$ in n variables over the commutative ring R is an example of a graded ring. Here $S_0 = R$ and the homogeneous component of degree k is the subgroup of all R-linear combinations of monomials of degree k.

The ideal I generated by $x_1, \ldots, x_n$ is a graded ideal: every polynomial with zero constant term may be written uniquely as a sum of homogeneous polynomials of degree $k > 1$, and each of these has zero constant term hence lies in I. More generally, an ideal is a graded ideal if and only if it can be generated by homogeneous polynomials (cf. Exercise 17 in Section 9.1).

Not every ideal of a graded ring need be a graded ideal. For example in the graded ring $\mathbb{Z}[x]$ the principal ideal J generated by $1 + x$ is not graded: $1 + x \in J$ and $1 \notin J$ so $1 + x$ cannot be written as a sum of homogeneous polynomials each of which belongs to J.

The next result shows that quotients of graded rings by graded ideals are again graded rings.

Proposition 33. Let S be a graded ring, let I be a graded ideal in S and let $I_k = I \cap S_k$ for all $k \geq 0$. Then S/I is naturally a graded ring whose homogeneous component of degree k is isomorphic to S_k/I_k.

Proof: The map

$$S = \oplus_{k=0}^{\infty} S_k \longrightarrow \oplus_{k=0}^{\infty} (S_k/I_k)$$
$$(\ldots, s_k, \ldots) \longmapsto (\ldots, s_k \bmod I_k, \ldots)$$

is surjective with kernel $I = \oplus_{k=0}^{\infty} I_k$ and defines an isomorphism of graded rings. The details are left for the exercises.

Symmetric Algebras

The first application of Proposition 33 is in the construction of a commutative quotient ring of $\mathcal{T}(M)$ through which R-module homomorphisms from M to any *commutative* R-algebra must factor. This gives an "abelianized" version of Theorem 31. The construction is analogous to forming the commutator quotient G/G' of a group (cf. Section 5.4).

Definition. The *symmetric algebra* of an R-module M is the R-algebra obtained by taking the quotient of the tensor algebra $\mathcal{T}(M)$ by the ideal $\mathcal{C}(M)$ generated by all elements of the form $m_1 \otimes m_2 - m_2 \otimes m_1$, for all $m_1, m_2 \in M$. The symmetric algebra $\mathcal{T}(M)/\mathcal{C}(M)$ is denoted by $\mathcal{S}(M)$.

The tensor algebra $\mathcal{T}(M)$ is generated as a ring by $R = \mathcal{T}^0(M)$ and $M = \mathcal{T}^1(M)$, and these elements commute in the quotient ring $\mathcal{S}(M)$ by definition. It follows that the symmetric algebra $\mathcal{S}(M)$ is a commutative ring. The ideal $\mathcal{C}(M)$ is generated by homogeneous tensors of degree 2 and it follows easily that $\mathcal{C}(M)$ is a graded ideal. Then by Proposition 33 the symmetric algebra is a graded ring whose homogeneous component of degree k is $\mathcal{S}^k(M) = \mathcal{T}^k(M)/\mathcal{C}^k(M)$. Since $\mathcal{C}(M)$ consists of k-tensors

with $k \geq 2$, we have $\mathcal{C}(M) \cap M = 0$ and so the image of $M = \mathcal{T}^1(M)$ in $\mathcal{S}(M)$ is isomorphic to M. Identifying M with its image we see that $\mathcal{S}^1(M) = M$ and the symmetric algebra contains M. In a similar way $\mathcal{S}^0(M) = R$, so the symmetric algebra is also an R-algebra. The R-module $\mathcal{S}^k(M)$ is called the k^{th} *symmetric power* of M.

The first part of the next theorem shows that the elements of the k^{th} symmetric power of M can be considered as finite sums of simple tensors $m_1 \otimes \cdots \otimes m_k$ where tensors with the order of the factors permuted are identified. Recall also from Section 4 that a k-multilinear map $\varphi : M \times \cdots \times M \to N$ is said to be *symmetric* if $\varphi(m_1, \ldots, m_k) = \varphi(m_{\sigma(1)}, \ldots, m_{\sigma(k)})$ for all permutations σ of $1, 2, \ldots, k$. (The definition is the same for modules over any commutative ring R as for vector spaces.)

Theorem 34. Let M be an R-module over the commutative ring R and let $\mathcal{S}(M)$ be its symmetric algebra.

(1) The k^{th} symmetric power, $\mathcal{S}^k(M)$, of M is equal to $M \otimes \cdots \otimes M$ (k factors) modulo the submodule generated by all elements of the form

$$(m_1 \otimes m_2 \otimes \cdots \otimes m_k) - (m_{\sigma(1)} \otimes m_{\sigma(2)} \otimes \cdots \otimes m_{\sigma(k)})$$

for all $m_i \in M$ and all permutations σ in the symmetric group S_k.

(2) *(Universal Property for Symmetric Multilinear Maps)* If $\varphi : M \times \cdots \times M \to N$ is a symmetric k-multilinear map over R then there is a unique R-module homomorphism $\Phi : \mathcal{S}^k(M) \to N$ such that $\varphi = \Phi \circ \iota$, where

$$\iota : M \times \cdots \times M \to \mathcal{S}^k(M)$$

is the map defined by

$$\iota(m_1, \ldots, m_k) = m_1 \otimes \cdots \otimes m_n \bmod \mathcal{C}(M).$$

(3) *(Universal Property for maps to commutative R-algebras)* If A is any commutative R-algebra and $\varphi : M \to A$ is an R-module homomorphism, then there is a unique R-algebra homomorphism $\Phi : \mathcal{S}(M) \to A$ such that $\Phi|_M = \varphi$.

Proof: The k-tensors $\mathcal{C}^k(M)$ in the ideal $\mathcal{C}(M)$ are finite sums of elements of the form

$$m_1 \otimes \ldots \otimes m_{i-1} \otimes (m_i \otimes m_{i+1} - m_{i+1} \otimes m_i) \otimes m_{i+2} \otimes \ldots \otimes m_k$$

with $m_1, \ldots, m_k \in M$ (where $k \geq 2$ and $1 \leq i < k$). This product gives a difference of two k-tensors which are equal except that two entries (in positions i and $i+1$) have been transposed, i.e., gives the element in (1) of the theorem corresponding to the transposition $(i\ i{+}1)$ in the symmetric group S_k. Conversely, since any permutation σ in S_k can be written as a product of such transpositions it is easy to see that every element in (1) can be written as a sum of elements of the form above. This gives (1).

The proofs of (2) and (3) are very similar to the proofs of the corresponding "asymmetric" results (Corollary 16 of Section 10.4 and Theorem 31) noting that $\mathcal{C}^k(M)$ is contained in the kernel of any symmetric map from $\mathcal{T}^k(M)$ to N by part (1).

Corollary 35. Let V be an n-dimensional vector space over the field F. Then $\mathcal{S}(V)$ is isomorphic as a graded F-algebra to the ring of polynomials in n variables over F (i.e., the isomorphism is also a vector space isomorphism from $\mathcal{S}^k(V)$ onto the space of all homogeneous polynomials of degree k). In particular, $\dim_F(\mathcal{S}^k(V)) = \binom{k+n-1}{n-1}$.

Proof: Let $\mathcal{B} = \{v_1, \ldots, v_n\}$ be a basis of V. By Proposition 32 there is a bijection between a basis of $\mathcal{T}^k(V)$ and the set $\mathcal{B}^k$ of ordered k-tuples of elements from $\mathcal{B}$. Define two k-tuples in $\mathcal{B}^k$ to be equivalent if there is some permutation of the entries of one that gives the other — this is easily seen to be an equivalence relation on $\mathcal{B}^k$. Let $S(\mathcal{B}^k)$ denote the corresponding set of equivalence classes. Any symmetric k-multilinear function from V^k to a vector space over F will be constant on all of the basis tensors whose corresponding k-tuples lie in the same equivalence class; conversely, any function from $S(\mathcal{B}^k)$ can be uniquely extended to a symmetric k-multilinear function on V^k. It follows that the vector space over F with basis $S(\mathcal{B}^k)$ satisfies the universal property of $\mathcal{S}^k(V)$ in Theorem 34(2), hence is isomorphic to $\mathcal{S}^k(V)$. Each equivalence class has a unique representative of the form $(v_1^{a_1}, v_2^{a_2}, \ldots, v_n^{a_n})$, where v_i^a denotes the sequence $v_i, v_i, \ldots, v_i$ taken a times, each $a_i \geq 0$, and $a_1 + \cdots + a_n = k$. Thus there is a bijection between the basis $S^k(\mathcal{B})$ and the set $x_1^{a_1} \cdots x_n^{a_n}$ of monic monomials of degree k in the polynomial ring $F[x_1, \ldots, x_n]$. This bijection extends to an isomorphism of graded F-algebras, proving the first part of the corollary. The computation of the dimension of $\mathcal{S}^k(V)$ (i.e., the number of monic monomials of degree k) is left as an exercise.

Exterior Algebras

Recall from Section 4 that a multilinear map $\varphi : M \times \cdots \times M \to N$ is called *alternating* if $\varphi(m_1, \ldots, m_k) = 0$ whenever $m_i = m_{i+1}$ for some i. (The definition is the same for any R-module as for vector spaces.) We saw that the determinant map was alternating, and was uniquely determined by some additional constraints. We can apply Proposition 33 to construct an algebra through which alternating multilinear maps must factor in a manner similar to the construction of the symmetric algebra (through which symmetric multilinear maps factor).

Definition. The *exterior algebra* of an R-module M is the R-algebra obtained by taking the quotient of the tensor algebra $\mathcal{T}(M)$ by the ideal $\mathcal{A}(M)$ generated by all elements of the form $m \otimes m$, for $m \in M$. The exterior algebra $\mathcal{T}(M)/\mathcal{A}(M)$ is denoted by $\bigwedge(M)$ and the image of $m_1 \otimes m_2 \otimes \cdots \otimes m_k$ in $\bigwedge(M)$ is denoted by $m_1 \wedge m_2 \wedge \cdots \wedge m_k$.

As with the symmetric algebra, the ideal $\mathcal{A}(M)$ is generated by homogeneous elements hence is a graded ideal. By Proposition 33 the exterior algebra is graded, with k^{th} homogeneous component $\bigwedge^k(M) = \mathcal{T}^k(M)/\mathcal{A}^k(M)$. We can again identify R with $\bigwedge^0(M)$ and M with $\bigwedge^1(M)$ and so consider M as an R-submodule of the R-algebra $\bigwedge(M)$. The R-module $\bigwedge^k(M)$ is called the k^{th} *exterior power* of M.

The multiplication

$$(m_1 \wedge \cdots \wedge m_i) \wedge (m_1' \wedge \cdots \wedge m_j') = m_1 \wedge \cdots \wedge m_i \wedge m_1' \wedge \cdots \wedge m_j'$$

in the exterior algebra is called the *wedge* (or *exterior*) *product*. By definition of the quotient, this multiplication is alternating in the sense that the product $m_1 \wedge \cdots \wedge m_k$ is 0 in $\bigwedge(M)$ if $m_i = m_{i+1}$ for any $1 \leq i < k$. Then

$$\begin{aligned} 0 &= (m + m') \wedge (m + m') \\ &= (m \wedge m) + (m \wedge m') + (m' \wedge m) + (m' \wedge m') \\ &= (m \wedge m') + (m' \wedge m) \end{aligned}$$

shows that the multiplication is also anticommutative on simple tensors:

$$m \wedge m' = -m' \wedge m \qquad \text{for all } m, m' \in M.$$

This anticommutativity does not extend to arbitrary products, however, i.e., we need not have $ab = -ba$ for all $a, b \in \bigwedge(M)$ (cf. Exercise 4).

Theorem 36. Let M be an R-module over the commutative ring R and let $\bigwedge(M)$ be its exterior algebra.

(1) The k^{th} exterior power, $\bigwedge^k(M)$, of M is equal to $M \otimes \cdots \otimes M$ (k factors) modulo the submodule generated by all elements of the form

$$m_1 \otimes m_2 \otimes \cdots \otimes m_k \quad \text{where } m_i = m_j \text{ for some } i \neq j.$$

In particular,

$$m_1 \wedge m_2 \wedge \cdots \wedge m_k = 0 \quad \text{if } m_i = m_j \text{ for some } i \neq j.$$

(2) *(Universal Property for Alternating Multilinear Maps)* If $\varphi : M \times \cdots \times M \to N$ is an alternating k-multilinear map then there is a unique R-module homomorphism $\Phi : \bigwedge^k(M) \to N$ such that $\varphi = \Phi \circ \iota$, where

$$\iota : M \times \cdots \times M \to \bigwedge^k(M)$$

is the map defined by

$$\iota(m_1, \ldots, m_k) = m_1 \wedge \cdots \wedge m_k.$$

Remark: The exterior algebra also satisfies a universal property similar to (3) of Theorem 34, namely with respect to R-module homomorphisms from M to R-algebras A satisfying $a^2 = 0$ for all $a \in A$ (cf. Exercise 6).

Proof: The k-tensors $\mathcal{A}^k(M)$ in the ideal $\mathcal{A}(M)$ are finite sums of elements of the form

$$m_1 \otimes \ldots \otimes m_{i-1} \otimes (m \otimes m) \otimes m_{i+2} \otimes \ldots \otimes m_k$$

with $m_1, \ldots, m_k, m \in M$ (where $k \geq 2$ and $1 \leq i < k$), which is a k-tensor with two equal entries (in positions i and $i+1$), so is of the form in (1). For the reverse inclusion, note that since

$$\begin{aligned} m' \otimes m &= -m \otimes m' + \left[(m + m') \otimes (m + m') - m \otimes m - m' \otimes m'\right] \\ &\equiv -m \otimes m' \bmod \mathcal{A}(M), \end{aligned}$$

interchanging any two consecutive entries and multiplying by -1 in a simple k-tensor gives an equivalent tensor modulo $\mathcal{A}^k(M)$. Using such a sequence of interchanges and sign changes we can arrange for the equal entries m_i and m_j of a simple tensor as in (1) to be adjacent, which gives an element of $\mathcal{A}^k(M)$. It follows that the generators in (1) are contained in $\mathcal{A}^k(M)$, which proves the first part of the theorem.

As in Theorem 34, the proof of (2) follows easily from the corresponding result for the tensor algebra in Theorem 31 since $\mathcal{A}^k(M)$ is contained in the kernel of any alternating map from $\mathcal{T}^k(M)$ to N.

Examples

(1) Suppose V is a one-dimensional vector space over F with basis element v. Then $\bigwedge^k(V)$ consists of finite sums of elements of the form $\alpha_1 v \wedge \alpha_2 v \wedge \cdots \wedge \alpha_k v$, i.e., $\alpha_1\alpha_2 \cdots \alpha_k(v \wedge v \wedge \cdots \wedge v)$ for $\alpha_1, \ldots, \alpha_k \in F$. Since $v \wedge v = 0$, it follows that $\bigwedge^0(V) = F$, $\bigwedge^1(V) = V$, and $\bigwedge^i(V) = 0$ for $i \geq 2$, so as a graded F-algebra we have

$$\bigwedge(V) = F \oplus V \oplus 0 \oplus 0 \oplus \ldots.$$

(2) Suppose now that V is a two-dimensional vector space over F with basis v, v'. Here $\bigwedge^k(V)$ consists of finite sums of elements of the form $(\alpha_1 v + \alpha_1' v') \wedge \cdots \wedge (\alpha_k v + \alpha_k' v')$. Such an element is a sum of elements that are simple wedge products involving only v and v'. For example, an element in $\bigwedge^2(V)$ is a sum of elements of the form

$$\begin{aligned}(av + bv') \wedge (cv + dv') &= ac(v \wedge v) + ad(v \wedge v') + bc(v' \wedge v) \\ &\quad + bd(v' \wedge v') \\ &= (ad - bc)v \wedge v'.\end{aligned}$$

It follows that $\bigwedge^i(V) = 0$ for $i \geq 3$ since then at least one of v, v' appears twice in such simple products.

We can see directly from $\bigwedge^2(V) = \mathcal{T}^2(V)/\mathcal{A}^2(V)$ that $v \wedge v' \neq 0$, as follows. The vector space $\mathcal{T}^2(V)$ is 4-dimensional with $v \otimes v$, $v \otimes v'$, $v' \otimes v$, $v' \otimes v'$ as basis (Proposition 16). The elements $v \otimes v$, $v \otimes v' + v' \otimes v$, $v' \otimes v'$ and $v \otimes v'$ are therefore also a basis for $\mathcal{T}^2(V)$. The subspace $\mathcal{A}^2(V)$ consists of all the 2-tensors in the ideal generated by the tensors

$$(av + bv') \otimes (av + bv') = a^2(v \otimes v) + ab(v \otimes v' + v' \otimes v) + b^2(v' \otimes v'),$$

from which it is clear that $\mathcal{A}^2(V)$ is contained in the 3-dimensional subspace having $v \otimes v$, $v \otimes v' + v' \otimes v$, and $v' \otimes v'$ as basis. In particular, the basis element $v \otimes v'$ of $\mathcal{T}^2(V)$ is not contained in $\mathcal{A}^2(V)$, i.e., $v \wedge v' \neq 0$ in $\bigwedge^2(V)$.

It follows that $\bigwedge^0(V) = F$, $\bigwedge^1(V) = V$, $\bigwedge^2(V) = F(v \wedge v')$, and $\bigwedge^i(V) = 0$ for $i \geq 3$, so as a graded F-algebra we have

$$\bigwedge(V) = F \oplus V \oplus F(v \wedge v') \oplus 0 \oplus \ldots.$$

As the previous examples illustrate, unlike the tensor and symmetric algebras, for finite dimensional vector spaces the exterior algebra is finite dimensional:

Corollary 37. Let V be a finite dimensional vector space over the field F with basis $\mathcal{B} = \{v_1, \ldots, v_n\}$. Then the vectors

$$v_{i_1} \wedge v_{i_2} \wedge \cdots \wedge v_{i_k} \qquad \text{for } 1 \leq i_1 < i_2 < \cdots < i_k \leq n$$

are a basis of $\bigwedge^k(V)$, and $\bigwedge^k(V) = 0$ when $k > n$ (when $k = 0$ the basis vector is the element $1 \in F$). In particular, $\dim_F(\bigwedge^k(V)) = \binom{n}{k}$.

Proof: As the proof of Theorem 36 shows, modulo $\mathcal{A}^k(M)$, the order of the terms in any simple k-tensor can be rearranged up to introducing a sign change. It follows that the k-tensors in the corollary (which have been arranged with increasing subscripts on the v_i and with no repeated entries) are generators for $\bigwedge^k(V)$. To show these vectors are linearly independent it suffices to exhibit an alternating k-multilinear function from V^k to F which is 1 on a given $v_{i_1} \wedge v_{i_2} \wedge \cdots \wedge v_{i_k}$ and zero on all other generators. Such a function f is defined on the basis of $\mathcal{T}^k(V)$ in Proposition 32 by $f(v_{j_1} \otimes v_{j_2} \otimes \cdots \otimes v_{j_k}) = \epsilon(\sigma)$ if σ is the unique permutation of $(j_1, j_2, \ldots, j_k)$ into $(i_1, i_2, \ldots, i_k)$, and f is zero on every basis tensor whose k-tuple of indices cannot be permuted to $(i_1, i_2, \ldots, i_k)$ (where $\epsilon(\sigma)$ is the sign of σ). Note that f is zero on any basis tensor with repeated entries. The value $\epsilon(\sigma)$ ensures that when f is extended to all elements of $\mathcal{T}^k(V)$ it gives an alternating map, i.e., f factors through $\mathcal{A}^k(V)$. Hence f is the desired function. The computation of the dimension of $\bigwedge^k(V)$ (i.e., of the number of increasing sequences of k-tuples of indices) is left to the exercises.

The results in Corollary 37 are true for any *free* R-module of rank n. In particular if $M \cong R^n$ with R-module basis $m_1, \ldots, m_n$ then

$$\bigwedge^n(M) = R(m_1 \wedge \cdots \wedge m_n)$$

is a free (rank 1) R-module with generator $m_1 \wedge \cdots \wedge m_n$ and

$$\bigwedge^{n+1}(M) = \bigwedge^{n+2}(M) = \cdots = 0.$$

Example

Let R be the polynomial ring $\mathbb{Z}[x, y]$ in the variables x and y. If $M = R$, then $\bigwedge^2(M) = 0$ so, for example, there are no nontrivial alternating bilinear maps on $R \times R$ by the universal property of $\bigwedge^2(R)$ with respect to such maps (Theorem 36).

Suppose now that $M = I$ is the ideal (x, y) generated by x and y in R. Then $I \bigwedge I \neq 0$. Perhaps the easiest way to see this is to construct a nontrivial alternating bilinear map on $I \times I$. The map

$$\varphi(ax + by, cx + dy) = (ad - bc) \bmod (x, y)$$

is a well defined alternating R-bilinear map from $I \times I$ to $\mathbb{Z} = R/I$ (cf. Exercise 7). Since $\varphi(x, y) = 1$, it follows that $x \wedge y \in \bigwedge^2(I)$ is nonzero. Unlike the situation of free modules as in the examples following Theorem 36 (where arguments involving *bases* could be used), in this case it is not at all a trivial matter to give a direct verification that $x \wedge y \neq 0$ in $\bigwedge^2(I)$.

Remark: The ideal I is an example of a rank 1 (but *not* free) R-module (the rank of a module over an integral domain is defined in Section 12.1), and this example shows that the results of Corollary 37 are not true in general if the R-module is not free over R.

Homomorphisms of Tensor Algebras

If $\varphi : M \to N$ is any R-module homomorphism, then there is an induced map on the k^{th} tensor power:

$$\mathcal{T}^k(\varphi) : m_1 \otimes m_2 \otimes \cdots \otimes m_k \longmapsto \varphi(m_1) \otimes \varphi(m_2) \otimes \cdots \otimes \varphi(m_k).$$

It follows directly that this map sends generators of each of the homogeneous components of the ideals $\mathcal{C}(M)$ and $\mathcal{A}(M)$ to themselves. Thus φ induces R-module homomorphisms on the quotients:

$$\mathcal{S}^k(\varphi) : \mathcal{S}^k(M) \longrightarrow \mathcal{S}^k(N) \qquad \text{and} \qquad \bigwedge^k(\varphi) : \bigwedge^k(M) \longrightarrow \bigwedge^k(N).$$

Moreover, each of these three maps is a ring homomorphism (hence they are graded R-algebra homomorphisms).

Of particular interest is the case when $M = V$ is an n-dimensional vector space over the field F and $\varphi : V \to V$ is an endomorphism. In this case by Corollary 37, $\bigwedge^n(\varphi)$ maps the 1-dimensional space $\bigwedge^n(V)$ to itself. Let $v_1, \ldots, v_n$ be a basis of V, so that $v_1 \wedge \cdots \wedge v_n$ is a basis of $\bigwedge^n(V)$. Then

$$\bigwedge^n(\varphi)(v_1 \wedge \cdots \wedge v_n) = \varphi(v_1) \wedge \cdots \wedge \varphi(v_n) = D(\varphi)v_1 \wedge \cdots \wedge v_n$$

for some scalar $D(\varphi) \in F$.

For any $n \times n$ matrix A over F we can define the associated endomorphism φ (with respect to the given basis $v_1, \ldots, v_n$), which gives a map $D : M_{n\times n}(F) \to F$ where $D(A) = D(\varphi)$. It is easy to check that this map D satisfies the three axioms for a determinant function in Section 4. Then the uniqueness statement of Theorem 24 gives:

Proposition 38. If φ is an endomorphism on a n-dimensional vector space V, then $\bigwedge^n(\varphi)(w) = \det(\varphi)w$ for all $w \in \bigwedge^n(V)$.

Note that Proposition 38 characterizes the determinant of the endomorphism φ as a certain naturally induced *linear* map on $\bigwedge^n(V)$. The fact that the determinant arises naturally when considering alternating multilinear maps also explains the source of the map φ in the example above.

As with the tensor product, the maps $\mathcal{S}^k(\varphi)$ and $\bigwedge^k(\varphi)$ induced from an injective map from M to N need not remain injective (so $\bigwedge^2(M)$ need not be a submodule of $\bigwedge^2(N)$ when M is a submodule of N, for example).

Example

The inclusion $\varphi : I \hookrightarrow R$ of the ideal (x, y) into the ring $R = \mathbb{Z}[x, y]$, both considered as R-modules, induces a map

$$\bigwedge^2(\varphi) : \bigwedge^2(I) \to \bigwedge^2(R).$$

Since $\bigwedge^2(R) = 0$ and $\bigwedge^2(I) \neq 0$, the map cannot be injective.

One can show that if M is an R-module *direct summand* of N, then $\mathcal{T}(M)$ (respectively, $\mathcal{S}(M)$ and $\bigwedge(M)$) is an R-subalgebra of $\mathcal{T}(N)$ (respectively, $\mathcal{S}(N)$ and $\bigwedge(N)$) (cf. the exercises). When $R = F$ is a field then *every* subspace M of N is a direct summand of N and so the corresponding algebra for M is a subalgebra of the algebra for N.

Symmetric and Alternating Tensors

The symmetric and exterior algebras can in some instances also be defined in terms of *symmetric* and *alternating* tensors (defined below), which identify these algebras as *sub*algebras of the tensor algebra rather than as quotient algebras.

For any R-module M there is a natural left group action of the symmetric group S_k on $M \times M \times \cdots \times M$ (k factors) given by permuting the factors:

$$\sigma(m_1, m_2, \ldots, m_k) = (m_{\sigma^{-1}(1)}, m_{\sigma^{-1}(2)}, \ldots, m_{\sigma^{-1}(k)}) \qquad \text{for each } \sigma \in S_k$$

(the reason for σ^{-1} is to make this a *left* group action, cf. Exercise 8 of Section 5.1). This map is clearly R-multilinear, so there is a well defined R-linear left group action of S_k on $\mathcal{T}^k(M)$ which is defined on simple tensors by

$$\sigma(m_1 \otimes m_2 \otimes \cdots \otimes m_k) = m_{\sigma^{-1}(1)} \otimes m_{\sigma^{-1}(2)} \otimes \cdots \otimes m_{\sigma^{-1}(k)} \qquad \text{for each } \sigma \in S_k.$$

Definition.

(1) An element $z \in \mathcal{T}^k(M)$ is called a *symmetric* k-tensor if $\sigma z = z$ for all σ in the symmetric group S_k.

(2) An element $z \in \mathcal{T}^k(M)$ is called an *alternating* k-tensor if $\sigma z = \epsilon(\sigma) z$ for all σ in the symmetric group S_k, where $\epsilon(\sigma)$ is the sign, ± 1, of the permutation σ.

It is immediate from the definition that the collection of symmetric (respectively, alternating) k-tensors is an R-submodule of the module of all k-tensors.

Example

The elements $m \otimes m$ and $m_1 \otimes m_2 + m_2 \otimes m_1$ are symmetric 2-tensors. The element $m_1 \otimes m_2 - m_2 \otimes m_1$ is an alternating 2-tensor.

It is also clear from the definition that both $\mathcal{C}^k(M)$ and $\mathcal{A}^k(M)$ are stable under the action of S_k, hence there is an induced action on the quotients $\mathcal{S}^k(M)$ and $\bigwedge^k(M)$.

Proposition 39. Let σ be an element in the symmetric group S_k and let $\epsilon(\sigma)$ be the sign of the permutation σ. Then

(1) for every $w \in \mathcal{S}^k(M)$ we have $\sigma w = w$, and

(2) for every $w \in \bigwedge^k(M)$ we have $\sigma w = \epsilon(\sigma) w$.

Proof: The first statement is immediate from (1) in Theorem 34. We showed in the course of the proof of Theorem 36 that

$$m_1 \wedge \cdots \wedge m_i \wedge m_{i+1} \wedge \cdots \wedge m_k = -m_1 \wedge \cdots \wedge m_{i+1} \wedge m_i \wedge \cdots \wedge m_k,$$

which shows that the formula in (2) is valid on simple products for the transposition $\sigma = (i\, i{+}1)$. Since these transpositions generate S_k and ϵ is a group homomorphism it follows that (2) is valid for any $\sigma \in S_k$ on simple products w. Since both sides are R-linear in w, it follows that (2) holds for all $w \in \bigwedge^k(M)$.

By Proposition 39, the symmetric group S_k acts trivially on both the *sub*module of symmetric k-tensors and the *quotient* module $\mathcal{S}^k(M)$, the k^{th} symmetric power of M. Similarly, S_k acts the same way on the submodule of alternating k-tensors as on $\bigwedge^k(M)$, the k^{th} exterior power of M. We now show that when $k!$ is a unit in R that these respective submodules and quotient modules are isomorphic (where $k!$ is the sum of the 1 of R with itself $k!$ times).

For any k-tensor $z \in \mathcal{T}^k(M)$ define

$$Sym(z) = \sum_{\sigma \in S_k} \sigma z$$

$$Alt(z) = \sum_{\sigma \in S_k} \epsilon(\sigma)\, \sigma z.$$

For any k-tensor z, the k-tensor $Sym(z)$ is symmetric and the k-tensor $Alt(z)$ is alternating. For example, for any $\tau \in S_k$

$$\begin{aligned} \tau\, Alt(z) &= \sum_{\sigma \in S_k} \epsilon(\sigma)\, \tau\sigma\, z \\ &= \sum_{\sigma' \in S_k} \epsilon(\tau^{-1}\sigma')\, \sigma' z \quad \text{(letting } \sigma' = \tau\sigma\text{)} \\ &= \epsilon(\tau^{-1}) \sum_{\sigma' \in S_k} \epsilon(\sigma')\, \sigma' z = \epsilon(\tau) Alt(z). \end{aligned}$$

The tensor $Sym(z)$ is sometimes called the *symmetrization* of z and $Alt(z)$ the *skew-symmetrization* of z.

If z is already a symmetric (respectively, alternating) tensor then $Sym(z)$ (respectively, $Alt(z)$) is just $k!z$. It follows that Sym (respectively, Alt) is an R-module endomorphism of $\mathcal{T}^k(M)$ whose image lies in the submodule of symmetric (respectively, alternating) tensors. In general these maps are not surjective, but if $k!$ is a unit in R then

$$\frac{1}{k!} Sym(z) = z \quad \text{for any symmetric tensor } z, \text{ and}$$

$$\frac{1}{k!} Alt(z) = z \quad \text{for any alternating tensor } z$$

so that in this case the maps $(1/k!)Sym$ and $(1/k!)Alt$ give surjective R-module homomorphisms from $\mathcal{T}^k(M)$ to the submodule of symmetric (respectively, alternating) tensors.

Proposition 40. Suppose $k!$ is a unit in the ring R and M is an R-module. Then

(1) The map $(1/k!)Sym$ induces an R-module isomorphism between the k^{th} symmetric power of M and the R-submodule of symmetric k-tensors:

$$\frac{1}{k!}Sym : \mathcal{S}^k(M) \cong \{\text{symmetric } k\text{-tensors}\}.$$

(2) The map $(1/k!)Alt$ induces an R-module isomorphism between the k^{th} exterior power of M and the R-submodule of alternating k-tensors:

$$\frac{1}{k!}Alt : \bigwedge^k(M) \cong \{\text{alternating } k\text{-tensors}\}.$$

Proof: We have seen that the respective maps are surjective R-homomorphisms from $\mathcal{T}^k(M)$ so to prove the proposition it suffices to check that their kernels are $\mathcal{C}^k(M)$ and $\mathcal{A}^k(M)$, respectively. We show the first and leave the second to the exercises. It is clear that Sym is 0 on any difference of two k-tensors which differ only in the order of their factors, so $\mathcal{C}^k(M)$ is contained in the kernel of $(1/k!)Sym$ by (1) of Theorem 34. For the reverse inclusion, observe that

$$z - \frac{1}{k!}Sym(z) = \frac{1}{k!}\sum_{\sigma \in S_k}(z - \sigma z)$$

for any k-tensor z. If z is in the kernel of Sym then the left hand side of this equality is just z; and since $z - \sigma z \in \mathcal{C}^k(M)$ for every $\sigma \in S_k$ (again by (1) of Theorem 34), it follows that $z \in \mathcal{C}^k(M)$, completing the proof.

The maps $(1/k!)Sym$ and $(1/k!)Alt$ are *projections* (cf. Exercise 11 in Section 2) onto the submodules of symmetric and antisymmetric tensors, respectively. Equivalently, if $k!$ is a unit in R, we have R-module direct sums

$$\mathcal{T}^k(M) = \ker(\pi) \oplus \text{image}(\pi)$$

for $\pi = (1/k!)Sym$ or $\pi = (1/k!)Alt$. In the former case the kernel consists of $\mathcal{C}^k(M)$ and the image is the collection of symmetric tensors (in which case $\mathcal{C}^k(M)$ is said to form an R-module *complement* to the symmetric tensors). In the latter case the kernel is $\mathcal{A}^k(M)$ and the image consists of the alternating tensors.

The R-linear left group action of S_k on $\mathcal{T}^k(M)$ makes $\mathcal{T}^k(M)$ into a module over the group ring RS_k (analogous to the formation of $F[x]$-modules described in Section 10.1). In terms of this module structure these projections give RS_k-submodule complements to the RS_k-submodules $\mathcal{C}^k(M)$ and $\mathcal{A}^k(M)$. The "averaging" technique used to construct these maps can be used to prove a very general result (Maschke's Theorem in Section 18.1) related to actions of finite groups on vector spaces (which is the subject of the "representation theory" of finite groups in Part VI).

If $k!$ is not invertible in R then in general we do not have such S_k-invariant direct sum decompositions so it is not in general possible to identify, for example, the k^{th} exterior power of M with the alternating k-tensors of M.

Note also that when $k!$ is invertible it is possible to *define* the k^{th} exterior power of M as the collection of alternating k-tensors (this equivalent approach is sometimes found

in the literature when the theory is developed over fields such as $\mathbb{R}$ and $\mathbb{C}$). In this case the multiplication of two alternating tensors z and w is defined by first taking the product $zw = z \otimes w$ in $\mathcal{T}(M)$ and then projecting the resulting tensor into the submodule of alternating tensors. Note that the simple product of two alternating tensors need not be alternating (for example, the square of an alternating tensor is a symmetric tensor).

Example

Let V be a vector space over a field F in which $k! \neq 0$. There are many *vector space complements* to $\mathcal{A}^k(V)$ in $\mathcal{T}^k(V)$ (just extend a basis for the subspace $\mathcal{A}^k(V)$ to a basis for $\mathcal{T}^k(V)$, for example). These complements depend on choices of bases for $\mathcal{T}^k(V)$ and so are indistinguishable from each other from vector space considerations alone. The additional structure on $\mathcal{T}^k(V)$ given by the action of S_k singles out a unique complement to $\mathcal{A}^k(V)$, namely the subspace of alternating tensors in Proposition 40.

Suppose that $k! \neq 0$ in F for all $k \geq 2$ (i.e., the field F has "characteristic 0," cf. Exercise 26 in Section 7.3), for example, $F = \mathbb{Q}$. Then the full exterior algebra $\bigwedge(V) = \oplus_{k \geq 0} \bigwedge^k(V)$ can be identified with the collection of tensors whose homogeneous components are alternating (with respect to the appropriate symmetric groups S_k).

Multiplication in $\bigwedge(V)$ in terms of alternating tensors is rather cumbersome, however. For example let v_1, v_2, v_3 be distinct basis vectors in V. The product of the two alternating tensors $z = v_1$ and $w = v_2 \otimes v_3 - v_3 \otimes v_2$ is obtained by first computing

$$z \otimes w = v_1 \otimes v_2 \otimes v_3 - v_1 \otimes v_3 \otimes v_2$$

in the full tensor algebra. This 3-tensor is not alternating — for example,

$$(1\,2)(z \otimes w) = v_2 \otimes v_1 \otimes v_3 - v_3 \otimes v_1 \otimes v_2 \neq -z \otimes w$$

and also $(1\,2\,3)(z \otimes w) = v_3 \otimes v_1 \otimes v_2 - v_2 \otimes v_1 \otimes v_3 \neq z \otimes w$. The multiplication requires that we project this tensor into the subspace of alternating tensors. This projection is given by $(1/3!)Alt(z \otimes w)$ and an easy computation shows that

$$\frac{1}{6}Alt(z \otimes w) = \frac{1}{3}[v_1 \otimes v_2 \otimes v_3 + v_2 \otimes v_3 \otimes v_1 + v_3 \otimes v_1 \otimes v_2 - v_1 \otimes v_3 \otimes v_2 - v_2 \otimes v_1 \otimes v_3 - v_3 \otimes v_2 \otimes v_1],$$

so the right hand side is the product of z and w in terms of alternating tensors. The same product in terms of the quotient algebra $\bigwedge(V)$ is simply

$$v_1 \wedge (2v_2 \wedge v_3) = 2v_1 \wedge v_2 \wedge v_3.$$

EXERCISES

In these exercises R is a commutative ring with 1 and M is an R-module; F is a field and V is a finite dimensional vector space over F.

1. Prove that if M is a cyclic R-module then $\mathcal{T}(M) = \mathcal{S}(M)$, i.e., the tensor algebra $\mathcal{T}(M)$ is commutative.
2. Fill in the details for the proof of Proposition 33 that $S/I = \oplus_{k=0}^{\infty} S_k/I_k$. [Show first that $S_i I_j \subseteq I_{i+j}$. Use this to show that the multiplication $(S_i/I_i)(S_j/I_j) \subseteq S_{i+j}/I_{i+j}$ is well defined, and then check the ring axioms and verify the statements made in the proof of Proposition 33.]

3. Show that the image of the map Sym_2 for the $\mathbb{Z}$-module $\mathbb{Z}$ consists of the 2-tensors $a(1\otimes 1)$ where a is an even integer. Conclude in particular that the symmetric tensor $1\otimes 1$ in $\mathbb{Z}\otimes_{\mathbb{Z}}\mathbb{Z}$ is not contained in the image of the map Sym.

4. Prove that $m \wedge n_1 \wedge n_2 \wedge \cdots \wedge n_k = (-1)^k(n_1 \wedge n_2 \wedge \cdots \wedge n_k \wedge m)$. In particular, $x \wedge (y \wedge z) = (y \wedge z) \wedge x$ for all $x, y, z \in M$.

5. Prove that if M is a free R-module of rank n then $\bigwedge^i(M)$ is a free R-module of rank $\binom{n}{i}$ for $i = 0, 1, 2, \ldots$.

6. If A is any R-algebra in which $a^2 = 0$ for all $a \in A$ and $\varphi : M \to A$ is an R-module homomorphism, prove there is a unique R-algebra homomorphism $\Phi : \bigwedge(M) \to A$ such that $\Phi|_M = \varphi$.

7. Let $R = \mathbb{Z}[x, y]$ and $I = (x, y)$.
   (a) Prove that if $ax + by = a'x + b'y$ in R then $a' = a + yf$ and $b' = b - xf$ for some polynomial $f(x, y) \in R$.
   (b) Prove that the map $\varphi(ax+by, cx+dy) = ad-bc \bmod (x, y)$ in the example following Corollary 37 is a well defined alternating R-bilinear map from $I \times I$ to $\mathbb{Z} = R/I$.

8. Let R be an integral domain and let F be its field of fractions.
   (a) Considering F as an R-module, prove that $\bigwedge^2 F = 0$.
   (b) Let I be any R-submodule of F (for example, any ideal in R). Prove that $\bigwedge^i I$ is a torsion R-module for $i \geq 2$ (i.e., for every $x \in \bigwedge^i I$ there is some nonzero $r \in R$ with $rx = 0$).
   (c) Give an example of an integral domain R and an R-module I in F with $\bigwedge^i I \neq 0$ for every $i \geq 0$ (cf. the example following Corollary 37).

9. Let $R = \mathbb{Z}[G]$ be the group ring of the group $G = \{1, \sigma\}$ of order 2. Let $M = \mathbb{Z}e_1 + \mathbb{Z}e_2$ be a free $\mathbb{Z}$-module of rank 2 with basis e_1 and e_2. Define $\sigma(e_1) = e_1 + 2e_2$ and $\sigma(e_2) = -e_2$. Prove that this makes M into an R-module and that the R-module $\bigwedge^2 M$ is a group of order 2 with $e_1 \wedge e_2$ as generator.

10. Prove that $z - (1/k!)Alt(z) = (1/k!)\sum_{\sigma\in S_k}(z - \epsilon(\sigma)\sigma z)$ for any k-tensor z and use this to prove that the kernel of the R-module homomorphism $(1/k!)Alt$ in Proposition 40 is $\mathcal{A}^k(M)$.

11. Prove that the image of Alt_k is the unique largest subspace of $\mathcal{T}^k(V)$ on which each permutation σ in the symmetric group S_k acts as multiplication by the scalar $\epsilon(\sigma)$.

12. (a) Prove that if $f(x, y)$ is an alternating bilinear map on V (i.e., $f(x, x) = 0$ for all $x \in V$) then $f(x, y) = -f(y, x)$ for all $x, y \in V$.
    (b) Suppose that $-1 \neq 1$ in F. Prove that $f(x, y)$ is an alternating bilinear map on V (i.e., $f(x, x) = 0$ for all $x \in V$) if and only if $f(x, y) = -f(y, x)$ for all $x, y \in V$.
    (c) Suppose that $-1 = 1$ in F. Prove that every alternating bilinear form $f(x, y)$ on V is symmetric (i.e., $f(x, y) = f(y, x)$ for all $x, y \in V$). Prove that there is a symmetric bilinear map on V that is not alternating. [One approach: show that $\mathcal{C}^2(V) \subset \mathcal{A}^2(V)$ and $\mathcal{C}^2(V) \neq \mathcal{A}^2(V)$ by counting dimensions. Alternatively, construct an explicit symmetric map that is not alternating.]

13. Let F be any field in which $-1 \neq 1$ and let V be a vector space over F. Prove that $V \otimes_F V = \mathcal{S}^2(V) \oplus \bigwedge^2(V)$ i.e., that every 2-tensor may be written uniquely as a sum of a symmetric and an alternating tensor.

14. Prove that if M is an R-module *direct factor* of the R-module N then $\mathcal{T}(M)$ (respectively, $\mathcal{S}(M)$ and $\bigwedge(M)$) is an R-subalgebra of $\mathcal{T}(N)$ (respectively, $\mathcal{S}(N)$ and $\bigwedge(N)$).

CHAPTER 12

Modules over Principal Ideal Domains

The main purpose of this chapter is to prove a structure theorem for finitely generated modules over particularly nice rings, namely Principal Ideal Domains. This theorem is an example of the ideal structure of the ring (which is particularly simple for P.I.D.s) being reflected in the structure of its modules. If we apply this result in the case where the P.I.D. is the ring of integers $\mathbb{Z}$ then we obtain a proof of the Fundamental Theorem of Finitely Generated Abelian Groups (which we examined in Chapter 5 without proof). If instead we apply this structure theorem in the case where the P.I.D. is the ring $F[x]$ of polynomials in x with coefficients in a field F we shall obtain the basic results on the so-called rational and Jordan canonical forms for a matrix. Before proceeding to the proof we briefly discuss these two important applications.

We have already discussed in Chapter 5 the result that any finitely generated abelian group is isomorphic to the direct sum of cyclic abelian groups, either $\mathbb{Z}$ or $\mathbb{Z}/n\mathbb{Z}$ for some positive integer $n \neq 0$. Recall also that an abelian group is the same thing as a $\mathbb{Z}$-module. Since the ideals of $\mathbb{Z}$ are precisely the trivial ideal (0) and the principal ideals $(n) = n\mathbb{Z}$ generated by positive integers n, we see that the Fundamental Theorem of Finitely Generated Abelian Groups in the language of modules says that any finitely generated $\mathbb{Z}$-module is the direct sum of modules of the form $\mathbb{Z}/I$ where I is an ideal of $\mathbb{Z}$ (these are the cyclic $\mathbb{Z}$-modules), together with a uniqueness statement when the direct sum is written in a particular form. Note the correspondence between the ideal structure of $\mathbb{Z}$ and the structure of its (finitely generated) modules, the finitely generated abelian groups.

The Fundamental Theorem of Finitely Generated Modules over a P.I.D. states that the same result holds when the Principal Ideal Domain $\mathbb{Z}$ is replaced by *any* P.I.D. In particular, we have seen in Chapter 10 that a module over the ring $F[x]$ of polynomials in x with coefficients in the field F is the same thing as a vector space V together with a fixed linear transformation T of V (where the element x acts on V by the linear transformation T). The Fundamental Theorem in this case will say that such a vector space is the direct sum of modules of the form $F[x]/I$ where I is an ideal of $F[x]$, hence is either the trivial ideal (0) or a principal ideal $(f(x))$ generated by some nonzero polynomial $f(x)$ (these are the cyclic $F[x]$-modules), again with a uniqueness statement when the direct sum is written in a particular form. If this is translated back into the language of vector spaces and linear transformations we can obtain information on the

linear transformation T.

For example, suppose V is a vector space of dimension n over F and we choose a basis for V. Then giving a linear transformation T of V to itself is the same thing as giving an $n \times n$ matrix A with coefficients in F (and choosing a different basis for V gives a different matrix B for T which is similar to A i.e., is of the form $P^{-1}AP$ for some invertible matrix P which defines the change of basis). We shall see that the Fundamental Theorem in this situation implies (under the assumption that the field F contains all the "eigenvalues" for the given linear transformation T) that there is a basis for V so that the associated matrix for T is *as close to being a diagonal matrix as possible* and so has a particularly simple form. This is the *Jordan canonical form.* The *rational canonical form* is another simple form for the matrix for T (that does not require the eigenvalues for T to be elements of F). In this way we shall be able to give canonical forms for arbitrary $n \times n$ matrices over fields F, that is, find matrices which are similar to a given $n \times n$ matrix and which are particularly simple (almost diagonal, for example).

Example

Let $V = \mathbb{Q}^3 = \{(x, y, z) \mid x, y, z \in \mathbb{Q}\}$ be the usual 3-dimensional vector space of ordered 3-tuples with entries from the field $F = \mathbb{Q}$ of rational numbers and suppose T is the linear transformation

$$T(x, y, z) = (9x + 4y + 5z, -4x - 3z, -6x - 4y - 2z), \qquad x, y, z \in \mathbb{Q}.$$

If we take the standard basis $e_1 = (1, 0, 0)$, $e_2 = (0, 1, 0)$, $e_3 = (0, 0, 1)$ for V then the matrix A representing this linear transformation is

$$A = \begin{pmatrix} 9 & 4 & 5 \\ -4 & 0 & -3 \\ -6 & -4 & -2 \end{pmatrix}.$$

We shall see that the Jordan canonical form for this matrix A is the much simpler matrix

$$B = \begin{pmatrix} 2 & 1 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix}$$

obtained by taking instead the basis $f_1 = (2, -1, -2)$, $f_2 = (1, 0, -1)$, $f_3 = (3, -2, -2)$ for V, since in this case

$$\begin{aligned} T(f_1) &= T(2, -1, -2) = (4, -2, -4) = 2 \cdot f_1 + 0 \cdot f_2 + 0 \cdot f_3 \\ T(f_2) &= T(1, 0, -1) = (4, -1, -4) = 1 \cdot f_1 + 2 \cdot f_2 + 0 \cdot f_3 \\ T(f_3) &= T(3, -2, -2) = (9, -6, -6) = 0 \cdot f_1 + 0 \cdot f_2 + 3 \cdot f_3, \end{aligned}$$

so the columns of the matrix representing T with respect to this basis are $(2, 0, 0)$, $(1, 2, 0)$ and $(0, 0, 3)$, i.e., T has matrix B with respect to this basis. In particular A is similar to the simpler matrix B.

In fact this linear transformation T *cannot* be diagonalized (i.e., there is no choice of basis for V for which the corresponding matrix is a diagonal matrix) so that the matrix B is as close to a diagonal matrix for T as is possible.

The first section below gives some general definitions and states and proves the Fundamental Theorem over an arbitrary P.I.D., after which we return to the application to canonical forms (the application to abelian groups appears in Chapter 5). These applications can be read independently of the general proof. An alternate and computationally useful proof valid for Euclidean Domains (so in particular for the rings $\mathbb{Z}$ and $F[x]$) along the lines of row and column operations is outlined in the exercises.

12.1 THE BASIC THEORY

We first describe some general finiteness conditions. Let R be a ring and let M be a left R-module.

Definition.

(1) The left R-module M is said to be a *Noetherian R-module* or to satisfy the *ascending chain condition on submodules* (or *A.C.C. on submodules*) if there are no infinite increasing chains of submodules, i.e., whenever

$$M_1 \subseteq M_2 \subseteq M_3 \subseteq \cdots$$

is an increasing chain of submodules of M, then there is a positive integer m such that for all $k \geq m$, $M_k = M_m$ (so the chain becomes stationary at stage m: $M_m = M_{m+1} = M_{m+2} = \ldots$).

(2) The ring R is said to be *Noetherian* if it is Noetherian as a left module over itself, i.e., if there are no infinite increasing chains of left ideals in R.

One can formulate analogous notions of A.C.C. on right and on two-sided ideals in a (possibly noncommutative) ring R. For noncommutative rings these properties need not be related.

Theorem 1. Let R be a ring and let M be a left R-module. Then the following are equivalent:

(1) M is a Noetherian R-module.

(2) Every nonempty set of submodules of M contains a maximal element under inclusion.

(3) Every submodule of M is finitely generated.

Proof: [(1) implies (2)] Assume M is Noetherian and let Σ be any nonempty collection of submodules of M. Choose any $M_1 \in \Sigma$. If M_1 is a maximal element of Σ, (2) holds, so assume M_1 is not maximal. Then there is some $M_2 \in \Sigma$ such that $M_1 \subset M_2$. If M_2 is maximal in Σ, (2) holds, so we may assume there is an $M_3 \in \Sigma$ properly containing M_2. Proceeding in this way one sees that if (2) fails we can produce by the Axiom of Choice an infinite strictly increasing chain of elements of Σ, contrary to (1).

[(2) implies (3)] Assume (2) holds and let N be any submodule of M. Let Σ be the collection of all finitely generated submodules of N. Since $\{0\} \in \Sigma$, this collection is nonempty. By (2) Σ contains a maximal element N'. If $N' \neq N$, let $x \in N - N'$. Since $N' \in \Sigma$, the submodule N' is finitely generated by assumption, hence also the

submodule generated by N' and x is finitely generated. This contradicts the maximality of N', so $N = N'$ is finitely generated.

[(3) implies (1)] Assume (3) holds and let $M_1 \subseteq M_2 \subseteq M_3 \ldots$ be a chain of submodules of M. Let

$$N = \bigcup_{i=1}^{\infty} M_i$$

and note that N is a submodule. By (3) N is finitely generated by, say, $a_1, a_2, \ldots, a_n$. Since $a_i \in N$ for all i, each a_i lies in one of the submodules in the chain, say M_{j_i}. Let $m = \max\{j_1, j_2, \ldots, j_n\}$. Then $a_i \in M_m$ for all i so the module they generate is contained in M_m, i.e., $N \subseteq M_m$. This implies $M_m = N = M_k$ for all $k \geq m$, which proves (1).

Corollary 2. If R is a P.I.D. then every nonempty set of ideals of R has a maximal element and R is a Noetherian ring.

Proof: The P.I.D. R satisfies condition (3) in the theorem with $M = R$.

Recall that even if M itself is a finitely generated R-module, submodules of M need not be finitely generated, so the condition that M be a Noetherian R-module is in general stronger than the condition that M be a finitely generated R-module.

We require a result on "linear dependence" before turning to the main results of this chapter.

Proposition 3. Let R be an integral domain and let M be a free R-module of rank $n < \infty$. Then any $n + 1$ elements of M are R-linearly dependent, i.e., for any $y_1, y_2, \ldots, y_{n+1} \in M$ there are elements $r_1, r_2, \ldots, r_{n+1} \in R$, not all zero, such that

$$r_1 y_1 + r_2 y_2 + \ldots + r_{n+1} y_{n+1} = 0.$$

Proof: The quickest way of proving this is to embed R in its quotient field F (since R is an integral domain) and observe that since $M \cong R \oplus R \oplus \cdots \oplus R$ (n times) we obtain $M \subseteq F \oplus F \oplus \cdots \oplus F$. The latter is an n-dimensional vector space over F so any $n+1$ elements of M are F-linearly dependent. By clearing the denominators of the scalars (by multiplying through by the product of all the denominators, for example), we obtain an R-linear dependence relation among the $n + 1$ elements of M.

Alternatively, let $e_1, \ldots, e_n$ be a basis of the free R-module M and let $y_1, \ldots, y_{n+1}$ be any $n + 1$ elements of M. For $1 \leq i \leq n+1$ write $y_i = a_{1i}e_i + a_{2i}e_2 + \ldots + a_{ni}e_i$ in terms of the basis $e_1, e_2, \ldots, e_n$. Let A be the $(n + 1) \times (n+1)$ matrix whose i, j entry is a_{ij}, $1 \leq i \leq n$, $1 \leq j \leq n + 1$ and whose last row is zero, so certainly $\det A = 0$. Since R is an integral domain, Corollary 27 of Section 11.4 shows that the columns of A are R-linearly dependent. Any dependence relation on the columns of A gives a dependence relation on the y_i's, completing the proof.

If R is any integral domain and M is any R-module recall that

$$\mathrm{Tor}(M) = \{x \in M \mid rx = 0 \text{ for some nonzero } r \in R\}$$

is a submodule of M (called *the* torsion submodule of M) and if N is any submodule of $\mathrm{Tor}(M)$, N is called *a* torsion submodule of M (so the torsion submodule of M is the union of all torsion submodules of M, i.e., is the maximal torsion submodule of M). If $\mathrm{Tor}(M) = 0$, the module M is said to be *torsion free*.

For any submodule N of M, the *annihilator* of N is the ideal of R defined by

$$\mathrm{Ann}(N) = \{r \in R \mid rn = 0 \text{ for all } n \in N\}.$$

Note that if N is not a torsion submodule of M then $\mathrm{Ann}(N) = (0)$. It is easy to see that if N, L are submodules of M with $N \subseteq L$, then $\mathrm{Ann}(L) \subseteq \mathrm{Ann}(N)$. If R is a P.I.D. and $N \subseteq L \subseteq M$ with $\mathrm{Ann}(N) = (a)$ and $\mathrm{Ann}(L) = (b)$, then $a \mid b$. In particular, the annihilator of any element x of M divides the annihilator of M (this is implied by Lagrange's Theorem when $R = \mathbb{Z}$).

Definition. For any integral domain R the *rank* of an R-module M is the maximum number of R-linearly independent elements of M.

The preceding proposition states that for a free R-module M over an integral domain the rank of a submodule is bounded by the rank of M. This notion of rank agrees with previous uses of the same term. If the ring $R = F$ is a field, then the rank of an R-module M is the dimension of M as a vector space over F and any maximal set of F-linearly independent elements is a basis for M. For a general integral domain, however, an R-module M of rank n need not have a "basis," i.e., need not be a *free* R-module even if M is torsion free, so some care is necessary with the notion of rank, particularly with respect to the torsion elements of M. Exercises 1 to 6 and 20 give an alternate characterization of the rank and provide some examples of (torsion free) R-modules (of rank 1) that are not free.

The next important result shows that if N is a submodule of a free module of finite rank over a P.I.D. then N is again a free module of finite rank and furthermore it is possible to choose generators for the two modules which are related in a simple way.

Theorem 4. Let R be a Principal Ideal Domain, let M be a free R-module of finite rank n and let N be a submodule of M. Then

(1) N is free of rank m, $m \leq n$ and

(2) there exists a basis $y_1, y_2, \ldots, y_n$ of M so that $a_1y_1, a_2y_2, \ldots, a_my_m$ is a basis of N where $a_1, a_2, \ldots, a_m$ are nonzero elements of R with the divisibility relations

$$a_1 \mid a_2 \mid \cdots \mid a_m.$$

Proof: The theorem is trivial for $N = \{0\}$, so assume $N \neq \{0\}$. For each R-module homomorphism φ of M into R, the image $\varphi(N)$ of N is a submodule of R, i.e., an ideal in R. Since R is a P.I.D. this ideal must be principal, say $\varphi(N) = (a_\varphi)$, for some $a_\varphi \in R$. Let

$$\Sigma = \{(a_\varphi) \mid \varphi \in \mathrm{Hom}_R(M, R)\}$$

be the collection of the principal ideals in R obtained in this way from the R-module homomorphisms of M into R. The collection Σ is certainly nonempty since taking φ

to be the trivial homomorphism shows that $(0) \in \Sigma$. By Corollary 2, Σ has at least one maximal element i.e., there is at least one homomorphism ν of M to R so that the principal ideal $\nu(N) = (a_\nu)$ is not properly contained in any other element of Σ. Let $a_1 = a_\nu$ for this maximal element and let $y \in N$ be an element mapping to the generator a_1 under the homomorphism ν: $\nu(y) = a_1$.

We now show the element a_1 is nonzero. Let $x_1, x_2, \ldots, x_n$ be any basis of the free module M and let $\pi_i \in \mathrm{Hom}_R(M, R)$ be the natural projection homomorphism onto the i^{th} coordinate with respect to this basis. Since $N \neq \{0\}$, there exists an i such that $\pi_i(N) \neq 0$, which in particular shows that Σ contains more than just the trivial ideal (0). Since (a_1) is a maximal element of Σ it follows that $a_1 \neq 0$.

We next show that this element a_1 divides $\varphi(y)$ for every $\varphi \in \mathrm{Hom}_R(M, R)$. To see this let d be a generator for the principal ideal generated by a_1 and $\varphi(y)$. Then d is a divisor of both a_1 and $\varphi(y)$ in R and $d = r_1 a_1 + r_2 \varphi(y)$ for some $r_1, r_2 \in R$. Consider the homomorphism $\psi = r_1 \nu + r_2 \varphi$ from M to R. Then $\psi(y) = (r_1 \nu + r_2 \varphi)(y) = r_1 a_1 + r_2 \varphi(y) = d$ so that $d \in \psi(N)$, hence also $(d) \subseteq \psi(N)$. But d is a divisor of a_1 so we also have $(a_1) \subseteq (d)$. Then $(a_1) \subseteq (d) \subseteq \psi(N)$ and by the maximality of (a_1) we must have equality: $(a_1) = (d) = \psi(N)$. In particular $(a_1) = (d)$ shows that $a_1 \mid \varphi(y)$ since d divides $\varphi(y)$.

If we apply this to the projection homomorphisms π_i we see that a_1 divides $\pi_i(y)$ for all i. Write $\pi_i(y) = a_1 b_i$ for some $b_i \in R$, $1 \leq i \leq n$ and define

$$y_1 = \sum_{i=1}^{n} b_i x_i.$$

Note that $a_1 y_1 = y$. Since $a_1 = \nu(y) = \nu(a_1 y_1) = a_1 \nu(y_1)$ and a_1 is a nonzero element of the integral domain R this shows

$$\nu(y_1) = 1.$$

We now verify that this element y_1 can be taken as one element in a basis for M and that $a_1 y_1$ can be taken as one element in a basis for N, namely that we have

(a) $M = Ry_1 \oplus \ker \nu$, and

(b) $N = Ra_1 y_1 \oplus (N \cap \ker \nu)$.

To see (a) let x be an arbitrary element in M and write $x = \nu(x)y_1 + (x - \nu(x)y_1)$. Since

$$\begin{aligned} \nu(x - \nu(x)y_1) &= \nu(x) - \nu(x)\nu(y_1) \\ &= \nu(x) - \nu(x) \cdot 1 \\ &= 0 \end{aligned}$$

we see that $x - \nu(x)y_1$ is an element in the kernel of ν. This shows that x can be written as the sum of an element in Ry_1 and an element in the kernel of ν, so $M = Ry_1 + \ker \nu$. To see that the sum is direct, suppose ry_1 is also an element in the kernel of ν. Then $0 = \nu(ry_1) = r\nu(y_1) = r$ shows that this element is indeed 0.

For (b) observe that $\nu(x')$ is divisible by a_1 for every $x' \in N$ by the definition of a_1 as a generator for $\nu(N)$. If we write $\nu(x') = ba_1$ where $b \in R$ then the decomposition we used in (a) above is $x' = \nu(x')y_1 + (x' - \nu(x')y_1) = ba_1 y_1 + (x' - ba_1 y_1)$ where the second summand is in the kernel of ν and is an element of N. This shows that

$N = Ra_1y_1 + (N \cap \ker \nu)$. The fact that the sum in (b) is direct is a special case of the directness of the sum in (a).

We now prove part (1) of the theorem by induction on the rank, m, of N. If $m = 0$, then N is a torsion module, hence $N = 0$ since a free module is torsion free, so (1) holds trivially. Assume then that $m > 0$. Since the sum in (b) above is direct we see easily that $N \cap \ker \nu$ has rank $m - 1$ (cf. Exercise 3). By induction $N \cap \ker \nu$ is then a free R-module of rank $m - 1$. Again by the directness of the sum in (b) we see that adjoining a_1y_1 to any basis of $N \cap \ker \nu$ gives a basis of N, so N is also free (of rank m), which proves (1).

Finally, we prove (2) by induction on n, the rank of M. Applying (1) to the submodule $\ker \nu$ shows that this submodule is free and because the sum in (a) is direct it is free of rank $n - 1$. By the induction assumption applied to the module $\ker \nu$ (which plays the role of M) and its submodule $\ker \nu \cap N$ (which plays the role of N), we see that there is a basis $y_2, y_3, \ldots, y_n$ of $\ker \nu$ such that $a_2y_2, a_3y_3, \ldots, a_my_m$ is a basis of $N \cap \ker \nu$ for some elements $a_2, a_3, \ldots, a_m$ of R with $a_2 \mid a_3 \mid \cdots \mid a_m$. Since the sums (a) and (b) are direct, $y_1, y_2, \ldots, y_n$ is a basis of M and $a_1y_1, a_2y_2, \ldots, a_my_m$ is a basis of N. To complete the induction it remains to show that a_1 divides a_2. Define a homomorphism φ from M to R by defining $\varphi(y_1) = \varphi(y_2) = 1$ and $\varphi(y_i) = 0$, for all $i > 2$, on the basis for M. Then for this homomorphism φ we have $a_1 = \varphi(a_1y_1)$ so $a_1 \in \varphi(N)$ hence also $(a_1) \subseteq \varphi(N)$. By the maximality of (a_1) in Σ it follows that $(a_1) = \varphi(N)$. Since $a_2 = \varphi(a_2y_2) \in \varphi(N)$ we then have $a_2 \in (a_1)$ i.e., $a_1 \mid a_2$. This completes the proof of the theorem.

Recall that the left R-module C is a *cyclic* R-module (for any ring R, not necessarily commutative nor with 1) if there is an element $x \in C$ such that $C = Rx$. We can then define an R-module homomorphism

$$\pi : R \to C$$

by $\pi(r) = rx$, which will be surjective by the assumption $C = Rx$. The First Isomorphism Theorem gives an isomorphism of (left) R-modules

$$R/\ker \pi \cong C.$$

If R is a P.I.D., $\ker \pi$ is a principal ideal, (a), so we see that the cyclic R-modules C are of the form $R/(a)$ where $(a) = \text{Ann}(C)$.

The cyclic modules are the simplest modules (since they require only one generator). The existence portion of the Fundamental Theorem states that any finitely generated module over a P.I.D. is isomorphic to the direct sum of finitely many cyclic modules.

Theorem 5. *(Fundamental Theorem, Existence: Invariant Factor Form)* Let R be a P.I.D. and let M be a finitely generated R-module.

(1) Then M is isomorphic to the direct sum of finitely many cyclic modules. More precisely,

$$M \cong R^r \oplus R/(a_1) \oplus R/(a_2) \oplus \cdots \oplus R/(a_m)$$

for some integer $r \geq 0$ and nonzero elements $a_1, a_2, \ldots, a_m$ of R which are not units in R and which satisfy the divisibility relations

$$a_1 \mid a_2 \mid \cdots \mid a_m.$$

(2) M is torsion free if and only if M is free.

(3) In the decomposition in (1),

$$\mathrm{Tor}(M) \cong R/(a_1) \oplus R/(a_2) \oplus \cdots \oplus R/(a_m).$$

In particular M is a torsion module if and only if $r = 0$ and in this case the annihilator of M is the ideal (a_m).

Proof: The module M can be generated by a finite set of elements by assumption so let $x_1, x_2, \ldots, x_n$ be a set of generators of M of minimal cardinality. Let R^n be the free R-module of rank n with basis $b_1, b_2, \ldots, b_n$ and define the homomorphism $\pi : R^n \to M$ by defining $\pi(b_i) = x_i$ for all i, which is automatically surjective since $x_1, \ldots, x_n$ generate M. By the First Isomorphism Theorem for modules we have $R^n/\ker \pi \cong M$. Now, by Theorem 4 applied to R^n and the submodule $\ker \pi$ we can choose another basis $y_1, y_2, \ldots, y_n$ of R^n so that $a_1y_1, a_2y_2, \ldots, a_my_m$ is a basis of $\ker \pi$ for some elements $a_1, a_2, \ldots, a_m$ of R with $a_1 \mid a_2 \mid \cdots \mid a_m$. This implies

$$M \cong R^n/\ker \pi = (Ry_1 \oplus Ry_2 \oplus \cdots \oplus Ry_n)/(Ra_1y_1 \oplus Ra_2y_2 \oplus \cdots \oplus Ra_my_m).$$

To identify the quotient on the right hand side we use the natural surjective R-module homomorphism

$$Ry_1 \oplus Ry_2 \oplus \cdots \oplus Ry_n \to R/(a_1) \oplus R/(a_2) \oplus \cdots \oplus R/(a_m) \oplus R^{n-m}$$

that maps $(\alpha_1y_1, \ldots, \alpha_ny_n)$ to $(\alpha_1 \bmod (a_1), \ldots, \alpha_m \bmod (a_m), \alpha_{m+1}, \ldots, \alpha_n)$. The kernel of this map is clearly the set of elements where a_i divides α_i, $i = 1, 2, \ldots, m$, i.e., $Ra_1y_1 \oplus Ra_2y_2 \oplus \cdots \oplus Ra_my_m$ (cf. Exercise 7). Hence we obtain

$$M \cong R/(a_1) \oplus R/(a_2) \oplus \cdots \oplus R/(a_m) \oplus R^{n-m}.$$

If a is a unit in R then $R/(a) = 0$, so in this direct sum we may remove any of the initial a_i which are units. This gives the decomposition in (1) (with $r = n - m$).

Since $R/(a)$ is a torsion R-module for any nonzero element a of R, (1) immediately implies M is a torsion free module if and only if $M \cong R^r$, which is (2). Part (3) is immediate from the definitions since the annihilator of $R/(a)$ is evidently the ideal (a).

We shall shortly prove the uniqueness of the decomposition in Theorem 5, namely that if we have

$$M \cong R^{r'} \oplus R/(b_1) \oplus R/(b_2) \oplus \cdots \oplus R/(b_{m'})$$

for some integer $r' \geq 0$ and nonzero elements $b_1, b_2, \ldots, b_{m'}$ of R which are not units with

$$b_1 \mid b_2 \mid \cdots \mid b_{m'},$$

then $r = r'$, $m = m'$ and $(a_i) = (b_i)$ (so $a_i = b_i$ up to units) for all i. It is precisely the divisibility condition $a_1 \mid a_2 \mid \cdots \mid a_m$ which gives this uniqueness.

Definition. The integer r in Theorem 5 is called the *free rank* or the *Betti number* of M and the elements $a_1, a_2, \ldots, a_m \in R$ (defined up to multiplication by units in R) are called the *invariant factors* of M.

Note that until we have proved that the invariant factors of M are unique we should properly refer to *a* set of invariant factors for M (and similarly for the free rank), by which we mean any elements giving a decomposition for M as in (1) of the theorem above.

Using the Chinese Remainder Theorem it is possible to decompose the cyclic modules in Theorem 5 further so that M is the direct sum of cyclic modules whose annihilators are as simple as possible (namely (0) or generated by powers of primes in R). This gives an alternate decomposition which we shall also see is unique and which we now describe.

Suppose a is a nonzero element of the Principal Ideal Domain R. Then since R is also a Unique Factorization Domain we can write

$$a = up_1^{\alpha_1} p_2^{\alpha_2} \cdots p_s^{\alpha_s}$$

where the p_i are distinct primes in R and u is a unit. This factorization is unique up to units, so the ideals $(p_i^{\alpha_i})$, $i = 1, \ldots, s$ are uniquely defined. For $i \neq j$ we have $(p_i^{\alpha_i}) + (p_j^{\alpha_j}) = R$ since the sum of these two ideals is generated by a greatest common divisor, which is 1 for distinct primes p_i, p_j. Put another way, the ideals $(p_i^{\alpha_i})$, $i = 1, \ldots, s$, are comaximal in pairs. The intersection of all these ideals is the ideal (a) since a is the least common multiple of $p_1^{\alpha_1}, p_2^{\alpha_2}, \ldots, p_s^{\alpha_s}$. Then the Chinese Remainder Theorem (Theorem 7.17) shows that

$$R/(a) \cong R/(p_1^{\alpha_1}) \oplus R/(p_2^{\alpha_2}) \oplus \cdots \oplus R/(p_s^{\alpha_s})$$

as rings and also as R-modules.

Applying this to the modules in Theorem 5 allows us to write each of the direct summands $R/(a_i)$ for the invariant factor a_i of M as a direct sum of cyclic modules whose annihilators are the prime power divisors of a_i. This proves:

Theorem 6. *(Fundamental Theorem, Existence: Elementary Divisor Form)* Let R be a P.I.D. and let M be a finitely generated R-module. Then M is the direct sum of a finite number of cyclic modules whose annihilators are either (0) or generated by powers of primes in R, i.e.,

$$M \cong R^r \oplus R/(p_1^{\alpha_1}) \oplus R/(p_2^{\alpha_2}) \oplus \cdots \oplus R/(p_t^{\alpha_t})$$

where $r \geq 0$ is an integer and $p_1^{\alpha_1}, \ldots, p_t^{\alpha_t}$ are positive powers of (not necessarily distinct) primes in R.

We proved Theorem 6 by using the prime power factors of the invariant factors for M. In fact we shall see that the decomposition of M into a direct sum of cyclic modules whose annihilators are (0) or prime powers as in Theorem 6 is unique, i.e., the integer r and the ideals $(p_1^{\alpha_1}), \ldots, (p_t^{\alpha_t})$ are uniquely defined for M. These prime powers are given a name:

Definition. Let R be a P.I.D. and let M be a finitely generated R-module as in Theorem 6. The prime powers $p_1^{\alpha_1}, \ldots, p_t^{\alpha_t}$ (defined up to multiplication by units in R) are called the *elementary divisors* of M.

Suppose M is a finitely generated torsion module over the Principal Ideal Domain R. If for the *distinct* primes $p_1, p_2, \ldots, p_n$ occurring in the decomposition in Theorem 6 we group together all the cyclic factors corresponding to the same prime p_i we see in particular that M can be written as a direct sum

$$M = N_1 \oplus N_2 \oplus \cdots \oplus N_n$$

where N_i consists of all the elements of M which are annihilated by some power of the prime p_i. This result holds also for modules over R which may not be finitely generated:

Theorem 7. *(The Primary Decomposition Theorem)* Let R be a P.I.D. and let M be a nonzero torsion R-module (not necessarily finitely generated) with nonzero annihilator a. Suppose the factorization of a into distinct prime powers in R is

$$a = up_1^{\alpha_1} p_2^{\alpha_2} \cdots p_n^{\alpha_n}$$

and let $N_i = \{x \in M \mid p_i^{\alpha_i} x = 0\}$, $1 \leq i \leq n$. Then N_i is a submodule of M with annihilator $p_i^{\alpha_i}$ and is the submodule of M of all elements annihilated by some power of p_i. We have

$$M = N_1 \oplus N_2 \oplus \cdots \oplus N_n.$$

If M is finitely generated then each N_i is the direct sum of finitely many cyclic modules whose annihilators are divisors of $p_i^{\alpha_i}$.

Proof: We have already proved these results in the case where M is finitely generated over R. In the general case it is clear that N_i is a submodule of M with annihilator dividing $p_i^{\alpha_i}$. Since R is a P.I.D. the ideals $(p_i^{\alpha_i})$ and $(p_j^{\alpha_j})$ are comaximal for $i \neq j$, so the direct sum decomposition of M can be proved easily by modifying the argument in the proof of the Chinese Remainder Theorem to apply it to modules. Using this direct sum decomposition it is easy to see that the annihilator of N_i is precisely $p_i^{\alpha_i}$.

Definition. The submodule N_i in the previous theorem is called the *p_i-primary component* of M.

Notice that with this terminology the elementary divisors of a finitely generated module M are just the invariant factors of the primary components of $\mathrm{Tor}(M)$.

We now prove the uniqueness statements regarding the decompositions in the Fundamental Theorem.

Note that if M is any module over a commutative ring R and a is an element of R then $aM = \{am \mid m \in M\}$ is a submodule of M. Recall also that in a Principal Ideal Domain R the nonzero prime ideals are maximal, hence the quotient of R by a nonzero prime ideal is a field.

Lemma 8. Let R be a P.I.D. and let p be a prime in R. Let F denote the field $R/(p)$.

(1) Let $M = R^r$. Then $M/pM \cong F^r$.

(2) Let $M = R/(a)$ where a is a nonzero element of R. Then

$$M/pM \cong \begin{cases} F & \text{if } p \text{ divides } a \text{ in } R \\ 0 & \text{if } p \text{ does not divide } a \text{ in } R. \end{cases}$$

(3) Let $M = R/(a_1) \oplus R/(a_2) \oplus \cdots \oplus R/(a_k)$ where each a_i is divisible by p. Then $M/pM \cong F^k$.

Proof: (1) There is a natural map from R^r to $(R/(p))^r$ defined by mapping $(\alpha_1, \ldots, \alpha_r)$ to $(\alpha_1 \bmod (p), \ldots, \alpha_r \bmod (p))$. This is clearly a surjective R-module homomorphism with kernel consisting of the r-tuples all of whose coordinates are divisible by p, i.e., pR^r, so $R^r/pR^r \cong (R/(p))^r$, which is (1).

(2) This follows from the Isomorphism Theorems: note first that $p(R/(a))$ is the image of the ideal (p) in the quotient $R/(a)$, hence is $(p)+(a)/(a)$. The ideal $(p)+(a)$ is generated by a greatest common divisor of p and a, hence is (p) if p divides a and is $R = (1)$ otherwise. Hence $pM = (p)/(a)$ if p divides a and is $R/(a) = M$ otherwise. If p divides a then $M/pM = (R/(a))/((p)/(a)) \cong R/(p)$, and if p does not divide a then $M/pM = M/M = 0$, which proves (2).

(3) This follows from (2) as in the proof of part (1) of Theorem 5.

Theorem 9. *(Fundamental Theorem, Uniqueness)* Let R be a P.I.D.

(1) Two finitely generated R-modules M_1 and M_2 are isomorphic if and only if they have the same free rank and the same list of invariant factors.

(2) Two finitely generated R-modules M_1 and M_2 are isomorphic if and only if they have the same free rank and the same list of elementary divisors.

Proof: If M_1 and M_2 have the same free rank and list of invariant factors or the same free rank and list of elementary divisors then they are clearly isomorphic.

Suppose that M_1 and M_2 are isomorphic. Any isomorphism between M_1 and M_2 maps the torsion in M_1 to the torsion in M_2 so we must have $\mathrm{Tor}(M_1) \cong \mathrm{Tor}(M_2)$. Then $R^{r_1} \cong M_1/\mathrm{Tor}(M_1) \cong M_2/\mathrm{Tor}(M_2) \cong R^{r_2}$ where r_1 is the free rank of M_1 and r_2 is the free rank of M_2. Let p be any nonzero prime in R. Then from $R^{r_1} \cong R^{r_2}$ we obtain $R^{r_1}/pR^{r_1} \cong R^{r_2}/pR^{r_2}$. By (1) of the previous lemma, this implies $F^{r_1} \cong F^{r_2}$ where F is the field R/pR. Hence we have an isomorphism of an r_1-dimensional vector space over F with an r_2-dimensional vector space over F, so that $r_1 = r_2$ and M_1 and M_2 have the same free rank.

We are reduced to showing that M_1 and M_2 have the same lists of invariant factors and elementary divisors. To do this we need only work with the isomorphic torsion modules $\mathrm{Tor}(M_1)$ and $\mathrm{Tor}(M_2)$, i.e., we may as well assume that both M_1 and M_2 are torsion R-modules.

We first show they have the same elementary divisors. It suffices to show that for any fixed prime p the elementary divisors which are a power of p are the same for both M_1 and M_2. If $M_1 \cong M_2$ then the p-primary submodule of M_1 (= the direct

sum of the cyclic factors whose elementary divisors are powers of p) is isomorphic to the p-primary submodule of M_2, since these are the submodules of elements which are annihilated by some power of p. We are therefore reduced to the case of proving that if two modules M_1 and M_2 which have annihilator a power of p are isomorphic then they have the same elementary divisors.

We proceed by induction on the power of p in the annihilator of M_1 (which is the same as the annihilator of M_2 since M_1 and M_2 are isomorphic). If this power is 0, then both M_1 and M_2 are 0 and we are done. Otherwise M_1 (and M_2) have nontrivial elementary divisors. Suppose the elementary divisors of M_1 are given by

$$\text{elementary divisors of } M_1\text{: } \underbrace{p, p, \ldots, p}_{m \text{ times}}, \ p^{\alpha_1}, p^{\alpha_2}, \ldots, p^{\alpha_s},$$

where $2 \leq \alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_s$, i.e., M_1 is the direct sum of cyclic modules with generators $x_1, x_2, \ldots, x_m, x_{m+1}, \ldots, x_{m+s}$, say, whose annihilators are $(p), (p), \ldots, (p)$, $(p^{\alpha_1}), \ldots, (p^{\alpha_s})$, respectively. Then the submodule pM_1 has elementary divisors

$$\text{elementary divisors of } pM_1\text{: } p^{\alpha_1 - 1}, p^{\alpha_2 - 1}, \ldots, p^{\alpha_s - 1}$$

since pM_1 is the direct sum of the cyclic modules with generators $px_1, px_2, \ldots, px_m$, $px_{m+1}, \ldots, px_{m+s}$ whose annihilators are $(1), (1), \ldots, (1), (p^{\alpha_1 - 1}), \ldots, (p^{\alpha_s - 1})$, respectively. Similarly, if the elementary divisors of M_2 are given by

$$\text{elementary divisors of } M_2\text{: } \underbrace{p, p, \ldots, p}_{n \text{ times}}, \ p^{\beta_1}, p^{\beta_2}, \ldots, p^{\beta_t},$$

where $2 \leq \beta_1 \leq \beta_2 \leq \cdots \leq \beta_t$, then pM_2 has elementary divisors

$$\text{elementary divisors of } pM_2\text{: } p^{\beta_1 - 1}, p^{\beta_2 - 1}, \ldots, p^{\beta_t - 1}.$$

Since $M_1 \cong M_2$, also $pM_1 \cong pM_2$ and the power of p in the annihilator of pM_1 is one less than the power of p in the annihilator of M_1. By induction, the elementary divisors for pM_1 are the same as the elementary divisors for pM_2, i.e., $s = t$ and $\alpha_i - 1 = \beta_i - 1$ for $i = 1, 2, \ldots, s$, hence $\alpha_i = \beta_i$ for $i = 1, 2, \ldots, s$. Finally, since also $M_1/pM_1 \cong M_2/pM_2$ we see from (3) of the lemma above that $F^{m+s} \cong F^{n+t}$, which shows that $m + s = n + t$ hence $m = n$ since we have already seen $s = t$. This proves that the set of elementary divisors for M_1 is the same as the set of elementary divisors for M_2.

We now show that M_1 and M_2 must have the same invariant factors. Suppose $a_1 \mid a_2 \mid \cdots \mid a_m$ are invariant factors for M_1. We obtain a set of elementary divisors for M_1 by taking the prime power factors of these elements. Note that then the divisibility relations on the invariant factors imply that a_m is the product of the largest of the prime powers among these elementary divisors, a_{m-1} is the product of the largest prime powers among these elementary divisors once the factors for a_m have been removed, and so on. If $b_1 \mid b_2 \mid \cdots \mid b_n$ are invariant factors for M_2 then we similarly obtain a set of elementary divisors for M_2 by taking the prime power factors of these elements. But we showed above that the elementary divisors for M_1 and M_2 are the same, and it follows that the same is true of the invariant factors.

Corollary 10. Let R be a P.I.D. and let M be a finitely generated R-module.

(1) The elementary divisors of M are the prime power factors of the invariant factors of M.

(2) The largest invariant factor of M is the product of the largest of the distinct prime powers among the elementary divisors of M, the next largest invariant factor is the product of the largest of the distinct prime powers among the remaining elementary divisors of M, and so on.

Proof: The procedure in (1) gives *a* set of elementary divisors and since the elementary divisors for M are unique by the theorem, it follows that the procedure in (1) gives *the* set of elementary divisors. Similarly for (2).

Corollary 11. *(The Fundamental Theorem of Finitely Generated Abelian Groups)* See Theorem 5.3 and Theorem 5.5.

Proof: Take $R = \mathbb{Z}$ in Theorems 5, 6 and 9 (note however that the invariant factors are listed in reverse order in Chapter 5 for computational convenience).

The procedure for passing between elementary divisors and invariant factors in Corollary 10 is described in some detail in Chapter 5 in the case of finitely generated abelian groups.

Note also that if a finitely generated module M is written as a direct sum of cyclic modules of the form $R/(a)$ then the ideals (a) which occur are not in general unique unless some additional conditions are imposed (such as the divisibility condition for the invariant factors or the condition that a be the power of a prime in the case of the elementary divisors). To decide whether two modules are isomorphic it is necessary to first write them in such a standard (or *canonical*) form.

EXERCISES

1. Let M be a module over the integral domain R.

(a) Suppose x is a nonzero torsion element in M. Show that x and 0 are "linearly dependent." Conclude that the rank of $\mathrm{Tor}(M)$ is 0, so that in particular any torsion R-module has rank 0.

(b) Show that the rank of M is the same as the rank of the (torsion free) quotient $M/\mathrm{Tor}M$.

2. Let M be a module over the integral domain R.

(a) Suppose that M has rank n and that $x_1, x_2, \ldots, x_n$ is any maximal set of linearly independent elements of M. Let $N = Rx_1 + \ldots + Rx_n$ be the submodule generated by $x_1, x_2, \ldots, x_n$. Prove that N is isomorphic to R^n and that the quotient M/N is a torsion R-module (equivalently, the elements $x_1, \ldots, x_n$ are linearly independent and for any $y \in M$ there is a nonzero element $r \in R$ such that ry can be written as a linear combination $r_1x_1 + \ldots + r_nx_n$ of the x_i).

(b) Prove conversely that if M contains a submodule N that is free of rank n (i.e., $N \cong R^n$) such that the quotient M/N is a torsion R-module then M has rank n. [Let $y_1, y_2, \ldots, y_{n+1}$ be any $n+1$ elements of M. Use the fact that M/N is torsion to write r_iy_i as a linear combination of a basis for N for some nonzero elements $r_1, \ldots, r_{n+1}$ of R. Use an argument as in the proof of Proposition 3 to see that the r_iy_i, and hence also the y_i, are linearly dependent.]

3. Let R be an integral domain and let A and B be R-modules of ranks m and n, respectively. Prove that the rank of $A \oplus B$ is $m + n$. [Use the previous exercise.]

4. Let R be an integral domain, let M be an R-module and let N be a submodule of M. Suppose M has rank n, N has rank r and the quotient M/N has rank s. Prove that $n = r + s$. [Let $x_1, x_2, \ldots, x_s$ be elements of M whose images in M/N are a maximal set of independent elements and let $x_{s+1}, x_{s+2}, \ldots, x_{s+r}$ be a maximal set of independent elements in N. Prove that $x_1, x_2, \ldots, x_{s+r}$ are linearly independent in M and that for any element $y \in M$ there is a nonzero element $r \in R$ such that ry is a linear combination of these elements. Then use Exercise 2.]

5. Let $R = \mathbb{Z}[x]$ and let $M = (2, x)$ be the ideal generated by 2 and x, considered as a submodule of R. Show that $\{2, x\}$ is not a basis of M. [Find a nontrivial R-linear dependence between these two elements.] Show that the rank of M is 1 but that M is not free of rank 1 (cf. Exercise 2).

6. Show that if R is an integral domain and M is any nonprincipal ideal of R then M is torsion free of rank 1 but is not a free R-module.

7. Let R be any ring, let $A_1, A_2, \ldots, A_m$ be R-modules and let B_i be a submodule of A_i, $1 \leq i \leq m$. Prove that

$$(A_1 \oplus A_2 \oplus \cdots \oplus A_m)/(B_1 \oplus B_2 \oplus \cdots \oplus B_m) \cong (A_1/B_1) \oplus (A_2/B_2) \oplus \cdots \oplus (A_m/B_m).$$

8. Let R be a P.I.D., let B be a torsion R-module and let p be a prime in R. Prove that if $pb = 0$ for some nonzero $b \in B$, then $\text{Ann}(B) \subseteq (p)$.

9. Give an example of an integral domain R and a nonzero torsion R-module M such that $\text{Ann}(M) = 0$. Prove that if N is a finitely generated torsion R-module then $\text{Ann}(N) \neq 0$.

10. For p a prime in the P.I.D. R and N an R-module prove that the p-primary component of N is a submodule of N and prove that N is the direct sum of its p-primary components (there need not be finitely many of them).

11. Let R be a P.I.D., let a be a nonzero element of R and let $M = R/(a)$. For any prime p of R prove that

$$p^{k-1}M/p^kM \cong \begin{cases} R/(p) & \text{if } k \leq n \\ 0 & \text{if } k > n, \end{cases}$$

where n is the power of p dividing a in R.

12. Let R be a P.I.D. and let p be a prime in R.
    (a) Let M be a finitely generated torsion R-module. Use the previous exercise to prove that $p^{k-1}M/p^kM \cong F^{n_k}$ where F is the field $R/(p)$ and n_k is the number of elementary divisors of M which are powers p^α with $\alpha \geq k$.
    (b) Suppose M_1 and M_2 are isomorphic finitely generated torsion R-modules. Use (a) to prove that, for every $k \geq 0$, M_1 and M_2 have the same number of elementary divisors p^α with $\alpha \geq k$. Prove that this implies M_1 and M_2 have the same set of elementary divisors.

13. If M is a finitely generated module over the P.I.D. R, describe the structure of $M/\text{Tor}(M)$.

14. Let R be a P.I.D. and let M be a torsion R-module. Prove that M is irreducible (cf. Exercises 9 to 11 of Section 10.3) if and only if $M = Rm$ for any nonzero element $m \in M$ where the annihilator of m is a nonzero prime ideal (p).

15. Prove that if R is a Noetherian ring then R^n is a Noetherian R-module. [Fix a basis of R^n. If M is a submodule of R^n show that the collection of first coordinates of elements of M is a submodule of R hence is finitely generated. Let $m_1, m_2, \ldots, m_k$ be elements of M

whose first coordinates generate this submodule of R. Show that any element of M can be written as an R-linear combination of $m_1, m_2, \ldots, m_k$ plus an element of M whose first coordinate is 0. Prove that $M \cap R^{n-1}$ is a submodule of R^{n-1} where R^{n-1} is the set of elements of R^n with first coordinate 0 and then use induction on n.

The following set of exercises outlines a proof of Theorem 5 in the special case where R is a Euclidean Domain using a matrix argument involving row and column operations. This applies in particular to the cases $R = \mathbb{Z}$ and $R = F[x]$ of interest in the applications and is computationally useful.

Let R be a Euclidean Domain and let M be an R-module.

16. Prove that M is finitely generated if and only if there is a surjective R-homomorphism $\varphi : R^n \to M$ for some integer n (this is true for any ring R).

Suppose $\varphi : R^n \to M$ is a surjective R-module homomorphism. By Exercise 15, $\ker \varphi$ is finitely generated. If $x_1, x_2, \ldots, x_n$ is a basis for R^n and $y_1, \ldots, y_m$ are generators for $\ker \varphi$ we have

$$y_i = a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n \quad i = 1, 2, \ldots, m$$

with coefficients $a_{ij} \in R$. It follows that the homomorphism φ (hence the module structure of M) is determined by the choice of generators for R^n and the matrix $A = (a_{ij})$. Such a matrix A will be called a *relations matrix*.

17. **(a)** Show that interchanging x_i and x_j in the basis for R^n interchanges the i^{th} column with the j^{th} column in the corresponding relations matrix.

(b) Show that, for any $a \in R$, replacing the element x_j by $x_j - ax_i$ in the basis for R^n gives another basis for R^n and that the corresponding relations matrix for this basis is the same as the original relations matrix except that a times the j^{th} column has been added to the i^{th} column. [Note that $\cdots + a_i x_i + \cdots + a_j x_j + \cdots = \cdots + (a_i + aa_j)x_i + \cdots + a_j(x_j - ax_i) + \ldots$.]

18. **(a)** Show that interchanging the generators y_i and y_j interchanges the i^{th} row with the j^{th} row in the relations matrix.

(b) Show that, for any $a \in R$, replacing the element y_j by $y_j - ay_i$ gives another set of generators for $\ker \varphi$ and that the corresponding relations matrix for this choice of generators is the same as the original relations matrix except that $-a$ times the i^{th} row has been added to the j^{th} row.

19. By the previous two exercises we may perform elementary row and column operations on a given relations matrix by choosing different generators for R^n and $\ker \varphi$. If all relation matrices are the zero matrix then $\ker \varphi = 0$ and $M \cong R^n$. Otherwise let a_1 be the (nonzero) g.c.d. (recall R is a Euclidean Domain) of all the entries in a fixed initial relations matrix for M.

(a) Prove that by elementary row and column operations we may assume a_1 occurs in a relations matrix of the form

$$\begin{pmatrix} a_1 & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \ldots & a_{mn} \end{pmatrix}$$

where a_1 divides a_{ij}, $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, n$.

(b) Prove that there is a relations matrix of the form

$$\begin{pmatrix} a_1 & 0 & \dots & 0 \\ 0 & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & a_{m2} & \dots & a_{mn} \end{pmatrix}$$

where a_1 divides all the entries.

(c) Let a_2 be a g.c.d. of all the entries except the element a_1 in the relations matrix in (b). Prove that there is a relations matrix of the form

$$\begin{pmatrix} a_1 & 0 & 0 & \dots & 0 \\ 0 & a_2 & 0 & \dots & 0 \\ 0 & 0 & a_{33} & \dots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & a_{m3} & \dots & a_{mn} \end{pmatrix}$$

where a_1 divides a_2 and a_2 divides all the other entries of the matrix.

(d) Prove that there is a relations matrix of the form $\begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix}$ where D is a diagonal matrix with nonzero entries $a_1, a_2, \dots, a_k$, $k \leq n$, satisfying

$$a_1 \mid a_2 \mid \cdots \mid a_k.$$

Conclude that

$$M \cong R/(a_1) \oplus R/(a_2) \oplus \cdots \oplus R/(a_k) \oplus R^{n-k}.$$

If n is not the minimal number of generators required for M then some of the initial elements $a_1, a_2, \dots$ above will be units, so the corresponding direct summands above will be 0. If we remove these irrelevant factors we have produced the invariant factors of the module M. Further, the image of the new generators for R^n corresponding to the direct summands above will then be a set of R-generators for the cyclic submodules of M in its invariant factor decomposition (note that the image in M of the generators corresponding to factors with a_i a unit will be 0). The *column* operations performed in the relations matrix reduction correspond to changing the basis used for R^n as described in Exercise 17:

(a) Interchanging the i^{th} column with the j^{th} column corresponds to interchanging the i^{th} and j^{th} elements in the basis for R^n.

(b) For any $a \in R$, adding a times the j^{th} column to the i^{th} column corresponds to subtracting a times the i^{th} basis element from the j^{th} basis element.

Keeping track of the column operations performed and changing the initial choice of generators for M in the same way therefore gives a set of R-generators for the cyclic submodules of M in its invariant factor decomposition.

This process is quite fast computationally once an initial set of generators for M and initial relations matrix are determined. The element a_1 is determined using the Euclidean Algorithm as the g.c.d. of the elements in the initial relations matrix. Using the row and column operations we can obtain the appropriate linear combination of the entries to produce this g.c.d. in the (1,1)-position of a new relations matrix. One then subtracts the appropriate multiple of the first column and first row to obtain a matrix as in Exercise 19(b), then iterates this process. Some examples of this procedure in a special case are given at the end of the following section.

20. Let R be an integral domain with quotient field F and let M be any R-module. Prove that the rank of M equals the dimension of the vector space $F \otimes_R M$ over F.

21. Prove that a finitely generated module over a P.I.D. is projective if and only if it is free.

22. Let R be a P.I.D. that is not a field. Prove that no finitely generated R-module is injective. [Use Exercise 4, Section 10.5 to consider torsion and free modules separately.]

12.2 THE RATIONAL CANONICAL FORM

We now apply our results on finitely generated modules in the special case where the P.I.D. is the ring $F[x]$ of polynomials in x with coefficients in a field F.

Let V be a finite dimensional vector space over F of dimension n and let T be a fixed linear transformation of V (i.e., from V to itself). As we saw in Chapter 10 we can consider V as an $F[x]$-module where the element x acts on V as the linear transformation T (and so any polynomial in x acts on V as the same polynomial in T). Since V has finite dimension over F by assumption, it is by definition finitely generated as an F-module, hence certainly finitely generated as an $F[x]$-module, so the classification theorems of the preceding section apply.

Any nonzero free $F[x]$-module (being isomorphic to a direct sum of copies of $F[x]$) is an infinite dimensional vector space over F, so if V has finite dimension over F then it must in fact be a torsion $F[x]$-module (i.e., its free rank is 0). It follows from the Fundamental Theorem that then V is isomorphic as an $F[x]$-module to the direct sum of cyclic, torsion $F[x]$-modules. We shall see that this decomposition of V will allow us to choose a basis for V with respect to which the matrix representation for the linear transformation T is in a specific simple form. When we use the invariant factor decomposition of V we obtain the *rational canonical form* for the matrix for T, which we analyze in this section. When we use the elementary divisor decomposition (and when F contains all the eigenvalues of T) we obtain the *Jordan canonical form*, considered in the following section and mentioned earlier as the matrix representing T which is as close to being a diagonal matrix as possible. The uniqueness portion of the Fundamental Theorem ensures that the rational and Jordan canonical forms are unique (which is why they are referred to as *canonical*).

One important use of these canonical forms is to classify the distinct linear transformations of V. In particular they allow us to determine when two matrices represent the same linear transformation, i.e., when two given $n \times n$ matrices are similar.

Note that this will be another instance where the structure of the space being acted upon (the invariant factor decomposition of V for example) is used to obtain significant information on the algebraic objects (in this case the linear transformations) which are acting. This will be considered in the case of *groups* acting on vector spaces in Chapter 18 (and goes under the name of Representation Theory of Groups).

Before describing the rational canonical form in detail we first introduce some linear algebra.

Definition.

(1) An element λ of F is called an *eigenvalue* of the linear transformation T if there is a nonzero vector $v \in V$ such that $T(v) = \lambda v$. In this situation v is called an *eigenvector* of T with corresponding eigenvalue λ.

(2) If A is an $n \times n$ matrix with coefficients in F, an element λ is called an *eigenvalue* of A with corresponding eigenvector v if v is a nonzero $n \times 1$ column vector such that $Av = \lambda v$.

(3) If λ is an eigenvalue of the linear transformation T, the set $\{v \in V \mid T(v) = \lambda v\}$ is called the *eigenspace* of T corresponding to the eigenvalue λ. Similarly, if λ is an eigenvalue of the $n \times n$ matrix A, the set of $n \times 1$ matrices v with $Av = \lambda v$ is called the *eigenspace* of A corresponding to the eigenvalue λ.

Note that if we fix a basis $\mathcal{B}$ of V then any linear transformation T of V has an associated $n \times n$ matrix A. Conversely, if A is any $n \times n$ matrix then the map T defined by $T(v) = Av$ for $v \in V$, where the v on the right is the $n \times 1$ vector consisting of the coordinates of v with respect to the fixed basis $\mathcal{B}$ of V, is a linear transformation of V. Then v is an eigenvector of T with corresponding eigenvalue λ if and only if the coordinate vector of v with respect to $\mathcal{B}$ is an eigenvector of A with eigenvalue λ. In other words, the eigenvalues for the linear transformation T are the same as the eigenvalues for the matrix A of T with respect to any fixed basis for V.

Definition. The determinant of a linear transformation from V to V is the determinant of any matrix representing the linear transformation (note that this does not depend on the choice of the basis used).

Proposition 12. The following are equivalent:

(1) λ is an eigenvalue of T

(2) $\lambda I - T$ is a singular linear transformation of V

(3) $\det(\lambda I - T) = 0$.

Proof: Since λ is an eigenvalue of T with corresponding eigenvector v if and only if v is a nonzero vector in the kernel of $\lambda I - T$, it follows that (1) and (2) are equivalent.

(2) and (3) are equivalent by our results on determinants.

Definition. Let x be an indeterminate over F. The polynomial $\det(xI - T)$ is called the *characteristic polynomial* of T and will be denoted $c_T(x)$. If A is an $n \times n$ matrix with coefficients in F, $\det(xI - A)$ is called the *characteristic polynomial* of A and will be denoted $c_A(x)$.

It is easy to see by expanding the determinant that the characteristic polynomial of either T or A is a monic polynomial of degree $n = \dim V$. Proposition 12 says that the set of eigenvalues of T (or A) is precisely the set of roots of the characteristic polynomial of T (of A, respectively). In particular, T has at most n distinct eigenvalues.

We have seen that V considered as a module over $F[x]$ via the linear transformation T is a torsion $F[x]$-module. Let $m(x) \in F[x]$ be the unique monic polynomial generating the annihilator of V in $F[x]$. Equivalently, $m(x)$ is the unique monic polynomial of minimal degree annihilating V (i.e., such that $m(T)$ is the 0 linear transformation), and if $f(x) \in F[x]$ is any polynomial annihilating V, $m(x)$ divides $f(x)$. Since the ring of all $n \times n$ matrices over F is isomorphic to the collection of all linear transformations of V to itself (an isomorphism is obtained by choosing a basis for V), it follows that for

any $n \times n$ matrix A over F there is similarly a unique monic polynomial of minimal degree with $m(A)$ the zero matrix.

Definition. The unique monic polynomial which generates the ideal $\mathrm{Ann}(V)$ in $F[x]$ is called the *minimal polynomial* of T and will be denoted $m_T(x)$. The unique monic polynomial of smallest degree which when evaluated at the matrix A is the zero matrix is called the *minimal polynomial* of A and will be denoted $m_A(x)$.

It is easy to see (cf. Exercise 5) that the degrees of these minimal polynomials are at most n^2 where n is the dimension of V. We shall shortly prove that the minimal polynomial for T is a divisor of the characteristic polynomial for T (this is the *Cayley–Hamilton Theorem*), and similarly for A, so in fact the degrees of these polynomials are at most n.

We now describe the *rational canonical form* of the linear transformation T (respectively, of the $n \times n$ matrix A). By Theorem 5 we have an isomorphism

$$V \cong F[x]/(a_1(x)) \oplus F[x]/(a_2(x)) \oplus \cdots \oplus F[x]/(a_m(x)) \qquad (12.1)$$

of $F[x]$-modules where $a_1(x), a_2(x), \ldots, a_m(x)$ are polynomials in $F[x]$ of degree at least one with the divisibility conditions

$$a_1(x) \mid a_2(x) \mid \cdots \mid a_m(x).$$

These invariant factors $a_i(x)$ are only determined up to a unit in $F[x]$ but since the units of $F[x]$ are precisely the nonzero elements of F (i.e., the nonzero constant polynomials), we may make these polynomials *unique* by stipulating that they be *monic*.

Since the annihilator of V is the ideal $(a_m(x))$ (part (3) of Theorem 5), we immediately obtain:

Proposition 13. The minimal polynomial $m_T(x)$ is the largest invariant factor of V. All the invariant factors of V divide $m_T(x)$.

We shall see below how to calculate not only the minimal polynomial for T but also the other invariant factors.

We now choose a basis for each of the direct summands for V in the decomposition (1) above for which the matrix for T is quite simple. Recall that the linear transformation T acting on the left side of (1) is the element x acting by multiplication on each of the factors on the right side of the isomorphism in (1).

We have seen in the example following Proposition 1 of Chapter 11 that the elements $1, \bar{x}, \bar{x}^2, \ldots, \bar{x}^{k-1}$ give a basis for the vector space $F[x]/(a(x))$ where $a(x) = x^k + b_{k-1}x^{k-1} + \cdots + b_1 x + b_0$ is any monic polynomial in $F[x]$ and $\bar{x} = x \bmod (a(x))$. With respect to this basis the linear transformation of multiplication by x acts in a simple manner:

$$x: \quad \begin{array}{ll} 1 & \mapsto \bar{x} \\ \bar{x} & \mapsto \bar{x}^2 \\ \bar{x}^2 & \mapsto \bar{x}^3 \\ & \vdots \\ \bar{x}^{k-2} & \mapsto \bar{x}^{k-1} \\ \bar{x}^{k-1} & \mapsto \bar{x}^k = -b_0 - b_1\bar{x} - \cdots - b_{k-1}\bar{x}^{k-1} \end{array}$$

where the last equality is because $\bar{x}^k + b_{k-1}\bar{x}^{k-1} + \cdots + b_1\bar{x} + b_0 = 0$ since $a(\bar{x}) = 0$ in $F[x]/(a(x))$. With respect to this basis, the matrix for multiplication by x is therefore

$$\begin{pmatrix} 0 & 0 & \dots & \dots & \dots & -b_0 \\ 1 & 0 & \dots & \dots & \dots & -b_1 \\ 0 & 1 & \dots & \dots & \dots & -b_2 \\ 0 & 0 & \ddots & & & \vdots \\ \vdots & \vdots & & \ddots & & \vdots \\ 0 & 0 & \dots & \dots & 1 & -b_{k-1} \end{pmatrix}.$$

Such matrices are given a name:

Definition. Let $a(x) = x^k + b_{k-1}x^{k-1} + \cdots + b_1x + b_0$ be any monic polynomial in $F[x]$. The *companion matrix* of $a(x)$ is the $k \times k$ matrix with 1's down the first subdiagonal, $-b_0, -b_1, \ldots, -b_{k-1}$ down the last column and zeros elsewhere. The companion matrix of $a(x)$ will be denoted by $\mathcal{C}_{a(x)}$.

We apply this to each of the cyclic modules on the right side of (1) above and let $\mathcal{B}_i$ be the elements of V corresponding to the basis chosen above for the cyclic factor $F[x]/(a_i(x))$ under the isomorphism in (1). Then by definition the linear transformation T acts on $\mathcal{B}_i$ by the companion matrix for $a_i(x)$ since we have seen that this is how multiplication by x acts. The union $\mathcal{B}$ of the $\mathcal{B}_i$'s gives a basis for V since the sum on the right of (1) is direct and with respect to this basis the linear transformation T has as matrix the *direct sum* of the companion matrices for the invariant factors, i.e.,

$$\begin{pmatrix} \mathcal{C}_{a_1(x)} & & & \\ & \mathcal{C}_{a_2(x)} & & \\ & & \ddots & \\ & & & \mathcal{C}_{a_m(x)} \end{pmatrix}. \tag{12.2}$$

Notice that this matrix is uniquely determined from the invariant factors of the $F[x]$-module V and, by Theorem 9, the list of invariant factors uniquely determines the module V up to isomorphism as an $F[x]$-module.

Definition.

(1) A matrix is said to be in *rational canonical form* if it is the direct sum of companion matrices for monic polynomials $a_1(x), \ldots, a_m(x)$ of degree at least one with $a_1(x) \mid a_2(x) \mid \cdots \mid a_m(x)$. The polynomials $a_i(x)$ are called the *invariant factors* of the matrix. Such a matrix is also said to be a *block diagonal* matrix with blocks the companion matrices for the $a_i(x)$.

(2) A *rational canonical form* for a linear transformation T is a matrix representing T which is in rational canonical form.

We have seen that any linear transformation T has a rational canonical form. We now see that this rational canonical form is unique (hence is called *the* rational canonical form for T). To see this note that the process we used to determine the matrix of T

from the direct sum decomposition is reversible. Suppose $b_1(x), b_2(x), \ldots, b_t(x)$ are monic polynomials in $F[x]$ of degree at least one such that $b_i(x) \mid b_{i+1}(x)$ for all i and suppose for some basis $\mathcal{E}$ of V, that the matrix of T with respect to the basis $\mathcal{E}$ is the direct sum of the companion matrices of the $b_i(x)$. Then V must be a direct sum of T-stable subspaces D_i, one for each $b_i(x)$ in such a way that the matrix of T on each D_i is the companion matrix of $b_i(x)$. Let $\mathcal{E}_i$ be the corresponding (ordered) basis of D_i (so $\mathcal{E}$ is the union of the $\mathcal{E}_i$) and let e_i be the first basis element in $\mathcal{E}_i$. Then it is easy to see that D_i is a cyclic $F[x]$-module with generator e_i and that the annihilator of D_i is $b_i(x)$. Thus the torsion $F[x]$-module V decomposes into a direct sum of cyclic $F[x]$-modules in two ways, both of which satisfy the conditions of Theorem 5, i.e., both of which give lists of invariant factors. Since the invariant factors are unique by Theorem 9, $a_i(x)$ and $b_i(x)$ must differ by a unit factor in $F[x]$ and since the polynomials are monic by assumption, we must have $a_i(x) = b_i(x)$ for all i. This proves the following result:

Theorem 14. *(Rational Canonical Form for Linear Transformations)* Let V be a finite dimensional vector space over the field F and let T be a linear transformation of V.

(1) There is a basis for V with respect to which the matrix for T is in rational canonical form, i.e., is a block diagonal matrix whose diagonal blocks are the companion matrices for monic polynomials $a_1(x), a_2(x), \ldots, a_m(x)$ of degree at least one with $a_1(x) \mid a_2(x) \mid \cdots \mid a_m(x)$.

(2) The rational canonical form for T is unique.

The use of the word *rational* is to indicate that this canonical form is calculated entirely within the field F and exists for any linear transformation T. This is not the case for the Jordan canonical form (considered later), which only exists if the field F contains the eigenvalues for T (cf. also the remarks following Corollary 18).

The following result translates the notion of similar linear transformations (i.e., the same linear transformation up to a change of basis) into the language of modules and relates this notion to rational canonical forms.

Theorem 15. Let S and T be linear transformations of V. Then the following are equivalent:

(1) S and T are similar linear transformations

(2) the $F[x]$-modules obtained from V via S and via T are isomorphic $F[x]$-modules

(3) S and T have the same rational canonical form.

Proof: [(1) implies (2)] Assume there is a nonsingular linear transformation U such that $S = UTU^{-1}$. The vector space isomorphism $U : V \to V$ is also an $F[x]$-module homomorphism, where x acts on the first V via T and on the second via S, since for example $U(xv) = U(Tv) = UT(v) = SU(v) = x(Uv)$. Hence this is an $F[x]$-module isomorphism of the two modules in (2).

[(2) implies (3)] Assume (2) holds and denote by V_1 the vector space V made into an $F[x]$-module via S and denote by V_2 the space V made into an $F[x]$-module via T. Since $V_1 \cong V_2$ as $F[x]$-modules they have the same list of invariant factors. Thus S and T have a common rational canonical form.

[(3) implies (1)] Assume (3) holds. Since S and T have the same matrix representation with respect to some choice of (possibly different) bases of V by assumption, they are, up to a change of basis, the same linear transformation of V, hence are similar.

Let A be any $n \times n$ matrix with entries from F. Let V be an n-dimensional vector space over F. Recall we can then *define* a linear transformation T on V by choosing a basis for V and setting $T(v) = Av$ where v on the right hand side means the $n \times 1$ column vector of coordinates of v with respect to our chosen basis (this is just the usual identification of linear transformations with matrices). Then (of course) the matrix for this T with respect to this basis is the given matrix A. Put another way, any $n \times n$ matrix A with entries from the field F arises as the matrix for some linear transformation T of an n-dimensional vector space.

This dictionary between linear transformations of vector spaces and matrices allows us to state our previous two results in the language of matrices:

Theorem 16. *(Rational Canonical Form for Matrices)* Let A be an $n \times n$ matrix over the field F.

(1) The matrix A is similar to a matrix in rational canonical form, i.e., there is an invertible $n \times n$ matrix P over F such that $P^{-1}AP$ is a block diagonal matrix whose diagonal blocks are the companion matrices for monic polynomials $a_1(x), a_2(x), \ldots, a_m(x)$ of degree at least one with $a_1(x) \mid a_2(x) \mid \cdots \mid a_m(x)$.

(2) The rational canonical form for A is unique.

Definition. The *invariant factors* of an $n \times n$ matrix over a field F are the invariant factors of its rational canonical form.

Theorem 17. Let A and B be $n \times n$ matrices over the field F. Then A and B are similar if and only if A and B have the same rational canonical form.

If A is a matrix with entries from a field F and F is a subfield of a larger field K then we may also consider A as a matrix over K. The next result shows that the rational canonical form for A and questions of similarity do not depend on which field contains the entries of A.

Corollary 18. Let A and B be two $n \times n$ matrices over a field F and suppose F is a subfield of the field K.

(1) The rational canonical form of A is the same whether it is computed over K or over F. The minimal and characteristic polynomials and the invariant factors of A are the same whether A is considered as a matrix over F or as a matrix over K.

(2) The matrices A and B are similar over K if and only if they are similar over F, i.e., there exists an invertible $n \times n$ matrix P with entries from K such that $B = P^{-1}AP$ if and only if there exists an (in general different) invertible $n \times n$ matrix Q with entries from F such that $B = Q^{-1}AQ$.

Proof: (1) Let M be the rational canonical form of A when computed over the smaller field F. Since M satisfies the conditions in the definition of the rational canonical form over K, the uniqueness of the rational canonical form implies that M is also

the rational canonical form of A over K. Hence the invariant factors of A are the same whether A is viewed over F or over K. In particular, since the minimal polynomial is the largest invariant factor of A it also does not depend on the field over which A is viewed. It is clear from the determinant definition of the characteristic polynomial of A that this polynomial depends only on the entries of A (we shall see shortly that the characteristic polynomial is the product of all the invariant factors for A, which will give an alternate proof of this result).

(2) If A and B are similar over the smaller field F they are clearly similar over K. Conversely, if A and B are similar over K, they have the same rational canonical form over K. By (1) they have the same rational canonical form over F, hence are similar over F by Theorem 17.

This corollary asserts in particular that the rational canonical form for an $n \times n$ matrix A is an $n \times n$ matrix with entries in the smallest field containing the entries of A. Further, this canonical form is the same matrix even if we allow conjugation of A by nonsingular matrices whose entries come from larger fields. This explains the terminology of *rational* canonical form.

The next proposition gives the connection between the characteristic polynomial of a matrix (or of a linear transformation) and its invariant factors and is quite useful for determining these invariant factors (particularly for matrices of small size).

Lemma 19. Let $a(x) \in F[x]$ be any monic polynomial.

(1) The characteristic polynomial of the companion matrix of $a(x)$ is $a(x)$.

(2) If M is the block diagonal matrix

$$M = \begin{pmatrix} A_1 & 0 & \dots & 0 \\ 0 & A_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & A_k \end{pmatrix},$$

given by the direct sum of matrices $A_1, A_2, \dots, A_k$ then the characteristic polynomial of M is the product of the characteristic polynomials of $A_1, A_2, \dots, A_k$.

Proof: These are both straightforward exercises.

Proposition 20. Let A be an $n \times n$ matrix over the field F.

(1) The characteristic polynomial of A is the product of all the invariant factors of A.

(2) *(The Cayley–Hamilton Theorem)* The minimal polynomial of A divides the characteristic polynomial of A.

(3) The characteristic polynomial of A divides some power of the minimal polynomial of A. In particular these polynomials have the same roots, not counting multiplicities.

The same statements are true if the matrix A is replaced by a linear transformation T of an n-dimensional vector space over F.

Proof: Let B be the rational canonical form of A. By the previous lemma the block diagonal form of B shows that the characteristic polynomial of B is the product of the characteristic polynomials of the companion matrices of the invariant factors of A. By the first part of the lemma above, the characteristic polynomial of the companion matrix $C_{a(x)}$ for $a(x)$ is just $a(x)$, which implies that the characteristic polynomial for B is the product of the invariant factors of A. Since A and B are similar, they have the same characteristic polynomial, which proves (1). Assertion (2) is immediate from (1) since the minimal polynomial for A is the largest invariant factor of A. The fact that all the invariant factors divide the largest one immediately implies (3). The final assertion is clear from the dictionary between linear transformations of vector spaces and matrices.

Note that part (2) of the proposition is the assertion that the matrix A satisfies its own characteristic polynomial, i.e., $c_A(A) = 0$ as matrices, which is the usual formulation for the Cayley–Hamilton Theorem. Note also that it implies the degree of the minimal polynomial for A has degree at most n, a result mentioned before.

The relations in Proposition 20 are frequently quite useful in the determination of the invariant factors for a matrix A, particularly for matrices of small degree (cf. Exercises 3 and 4 and the examples). The following result (which relies on Exercises 16 to 19 in the previous section and whose proof we outline in the exercises) computes the invariant factors in general.

Let A be an $n \times n$ matrix over the field F. Then $xI - A$ is an $n \times n$ matrix with entries in $F[x]$. The three operations
(a) interchanging two rows or columns
(b) adding a multiple (in $F[x]$) of one row or column to another
(c) multiplying any row or column by a unit in $F[x]$, i.e., by a nonzero element in F,
are called *elementary row and column operations.*

Theorem 21. Let A be an $n \times n$ matrix over the field F. Using the three elementary row and column operations above, the $n \times n$ matrix $xI - A$ with entries from $F[x]$ can be put into the diagonal form (called the *Smith Normal Form* for A)

$$\begin{pmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & & a_1(x) & & & \\ & & & & a_2(x) & & \\ & & & & & \ddots & \\ & & & & & & a_m(x) \end{pmatrix}$$

with monic nonzero elements $a_1(x), a_2(x), \ldots, a_m(x)$ of $F[x]$ with degrees at least one and satisfying $a_1(x) \mid a_2(x) \mid \cdots \mid a_m(x)$. The elements $a_1(x), \ldots, a_m(x)$ are the invariant factors of A.

Proof: cf. the exercises.

Invariant Factor Decomposition Algorithm: Converting to Rational Canonical Form

As mentioned in the exercises near the end of the previous section, keeping track of the operations necessary to diagonalize $xI - A$ will explicitly give a matrix P such that $P^{-1}AP$ is in rational canonical form. Equivalently, if V is a given $F[x]$-module with vector space basis $[e_1, e_2, \ldots, e_n]$, then P defines the change of basis giving the Invariant Factor Decomposition of V into a direct sum of cyclic $F[x]$-modules. In particular, if A is the matrix of the linear transformation T of the $F[x]$-module V defined by x (i.e., $T(e_j) = xe_j = \sum_{i=1}^{n} a_{ij}e_i$ where $A = (a_{ij})$), then the matrix P defines the change of basis for V with respect to which the matrix for T is in rational canonical form.

We first describe the algorithm in the general context of determining the Invariant Factor Decomposition of a given $F[x]$-module V with vector space basis $[e_1, e_2, \ldots, e_n]$ (the proof is outlined in the exercises). We then describe the algorithm to convert a given $n \times n$ matrix A to rational canonical form (in which reference to an underlying vector space and associated linear transformation are suppressed).

Explicit numerical examples of this algorithm are given in Examples 2 and 3 following.

Invariant Factor Decomposition Algorithm

Let V be an $F[x]$-module with vector space basis $[e_1, e_2, \ldots, e_n]$ (so in particular these elements are generators for V as an $F[x]$-module). Let T be the linear transformation of V to itself defined by x and let A be the $n \times n$ matrix associated to T and this choice of basis for V, i.e.,

$$T(e_j) = xe_j = \sum_{i=1}^{n} a_{ij}e_i \qquad \text{where} \qquad A = (a_{ij}).$$

(1) Use the following three elementary row and column operations to diagonalize the matrix $xI - A$ over $F[x]$, keeping track of the *row* operations used:
- **(a)** interchange two rows or columns (which will be denoted by $R_i \leftrightarrow R_j$ for the interchange of the i^{th} and j^{th} rows and similarly by $C_i \leftrightarrow C_j$ for columns),
- **(b)** add a multiple (in $F[x]$) of one row or column to another (which will be denoted by $R_i + p(x)R_j \mapsto R_i$ if $p(x)$ times the j^{th} row is added to the i^{th} row, and similarly by $C_i + p(x)C_j \mapsto C_i$ for columns),
- **(c)** multiply any row or column by a unit in $F[x]$, i.e., by a nonzero element in F (which will be denoted by uR_i if the i^{th} row is multiplied by $u \in F^{\times}$, and similarly by uC_i for columns).

(2) Beginning with the $F[x]$-module generators $[e_1, e_2, \ldots, e_n]$, for each row operation used in (1), change the set of generators by the following rules:
- **(a)** If the i^{th} row is interchanged with the j^{th} row then interchange the i^{th} and j^{th} generators.
- **(b)** If $p(x)$ times the j^{th} row is added to the i^{th} row then subtract $p(x)$ times the i^{th} generator from the j^{th} generator (note the indices).

(c) If the i^{th} row is multiplied by the unit $u \in F$ then divide the i^{th} generator by u.

(3) When $xI - A$ has been diagonalized to the form in Theorem 21 the generators $[e_1, e_2, \ldots, e_n]$ for V will be in the form of $F[x]$-linear combinations of $e_1, e_2, \ldots, e_n$. Use $xe_j = T(e_j) = \sum_{i=1}^{n} a_{ij}e_i$ to write these elements as F-linear combinations of $e_1, e_2, \ldots, e_n$. When $xI - A$ has been diagonalized, the first $n - m$ of these linear combinations are 0 (providing a useful numerical check on the computations) and the remaining m linear combinations are nonzero, i.e., the generators for V are in the form $[0, \ldots, 0, f_1, \ldots, f_m]$ corresponding precisely to the diagonal elements in Theorem 21. The elements $f_1, \ldots, f_m$ are a set of $F[x]$-module generators for the cyclic factors in the invariant factor decomposition of V (with annihilators $(a_1(x)), \ldots, (a_m(x))$, respectively):

$$V = F[x]\, f_1 \oplus F[x]\, f_2 \oplus \ldots \oplus F[x]\, f_m,$$

$$F[x]\, f_i \cong F[x]/(a_i(x)) \qquad i = 1, 2, \ldots, m,$$

giving the Invariant Factor Decomposition of the $F[x]$-module V.

(4) The corresponding *vector space* basis for each cyclic factor of V is then given by the elements $f_i, Tf_i, T^2 f_i, \ldots, T^{\deg a_i(x)-1} f_i$.

(5) Write the k^{th} element of the vector space basis computed in (4) in terms of the original vector space basis $[e_1, e_2, \ldots, e_n]$ and use the coordinates for the k^{th} column of an $n \times n$ matrix P. Then $P^{-1}AP$ is in rational canonical form (with diagonal blocks the companion matrices for the $a_i(x)$). This is the matrix for the linear transformation T with respect to the vector space basis in (4).

We now describe the algorithm to convert a given $n \times n$ matrix A to rational canonical form, i.e., to determine an $n \times n$ matrix P so that $P^{-1}AP$ is in rational canonical form. This is nothing more than the algorithm above applied to the vector space $V = F^n$ of $n \times 1$ column vectors with standard basis $[e_1, e_2, \ldots, e_n]$ (where e_i is the column vector with 1 in the i^{th} position and 0's elsewhere) and T is the linear transformation defined by A and this choice of basis. Explicit reference to this underlying vector space and associated linear transformation are suppressed, so the algorithm is purely matrix theoretic.

Converting an $n \times n$ Matrix to Rational Canonical Form

Let A be an $n \times n$ matrix with entries in the field F.

(1) Use the following three elementary row and column operations to diagonalize the matrix $xI - A$ over $F[x]$, keeping track of the *row* operations used:

(a) interchange two rows or columns (which will be denoted by $R_i \leftrightarrow R_j$ for the interchange of the i^{th} and j^{th} rows and similarly by $C_i \leftrightarrow C_j$ for columns),

(b) add a multiple (in $F[x]$) of one row or column to another (which will be denoted by $R_i + p(x)R_j \mapsto R_i$ if $p(x)$ times the j^{th} row is added to the i^{th} row, and similarly by $C_i + p(x)C_j \mapsto C_i$ for columns),

(c) multiply any row or column by a unit in $F[x]$, i.e., by a nonzero element in F (which will be denoted by uR_i if the i^{th} row is multiplied by $u \in F^\times$, and similarly by uC_i for columns).

Define $d_1, \ldots, d_m$ to be the degrees of the monic nonconstant polynomials $a_1(x), \ldots, a_m(x)$ appearing on the diagonal, respectively.

(2) Beginning with the $n \times n$ identity matrix P', for each row operation used in (1), change the matrix P' by the following rules:

(a) If $R_i \leftrightarrow R_j$ then interchange the i^{th} and j^{th} columns of P' (i.e., $C_i \leftrightarrow C_j$ for P').

(b) If $R_i + p(x)R_j \mapsto R_i$ then subtract the product of the matrix $p(A)$ times the i^{th} column of P' from the j^{th} column of P' (i.e., $C_j - p(A)C_i \mapsto C_j$ for P' — note the indices).

(c) If uR_i then divide the elements of the i^{th} column of P' by u (i.e., $u^{-1}C_i$ for P').

(3) When $xI - A$ has been diagonalized to the form in Theorem 21 the first $n - m$ columns of the matrix P' are 0 (providing a useful numerical check on the computations) and the remaining m columns of P' are nonzero. For each $i = 1, 2, \ldots, m$, multiply the i^{th} nonzero column of P' successively by $A^0 = I, A^1, A^2, \ldots, A^{d_i - 1}$, where d_i is the integer in (1) above and use the resulting column vectors (in this order) as the next d_i columns of an $n \times n$ matrix P. Then $P^{-1}AP$ is in rational canonical form (whose diagonal blocks are the companion matrices for the polynomials $a_1(x), \ldots, a_m(x)$ in (1)).

In the theory of canonical forms for linear transformations (or matrices) the characteristic polynomial plays the role of the order of a finite abelian group and the minimal polynomial plays the role of the exponent (after all, they are the same invariants, one for modules over the Principal Ideal Domain $\mathbb{Z}$ and the other for modules over the Principal Ideal Domain $F[x]$) so we can solve problems directly analogous to those we considered for finite abelian groups in Chapter 5. In particular, this includes the following:

(A) determine the rational canonical form of a given matrix (analogous to decomposing a finite abelian group as a direct product of cyclic groups)

(B) determine whether two given matrices are similar (analogous to determining whether two given finite abelian groups are isomorphic)

(C) determine all similarity classes of matrices over F with a given characteristic polynomial (analogous to determining all abelian groups of a given order)

(D) determine all similarity classes of $n \times n$ matrices over F with a given minimal polynomial (analogous to determining all abelian groups of rank at most n of a given exponent).

Examples

(1) We find the rational canonical forms of the following matrices over $\mathbb{Q}$ and determine if they are similar:

$$A = \begin{pmatrix} 2 & -2 & 14 \\ 0 & 3 & -7 \\ 0 & 0 & 2 \end{pmatrix} \quad B = \begin{pmatrix} 0 & -4 & 85 \\ 1 & 4 & -30 \\ 0 & 0 & 3 \end{pmatrix} \quad C = \begin{pmatrix} 2 & 2 & 1 \\ 0 & 2 & -1 \\ 0 & 0 & 3 \end{pmatrix}.$$

A direct computation shows that all three of these matrices have the same characteristic polynomial: $c_A(x) = c_B(x) = c_C(x) = (x-2)^2(x-3)$. Since the minimal and char-

acteristic polynomials have the same roots, the only possibilities for the minimal polynomials are $(x-2)(x-3)$ or $(x-2)^2(x-3)$. We quickly find that $(A-2I)(A-3I) = 0$, $(B-2I)(B-3I) \neq 0$ (the 1,1-entry is nonzero) and $(C-2I)(C-3I) \neq 0$ (the 1,2-entry is nonzero). It follows that

$$m_A(x) = (x-2)(x-3), \quad m_B(x) = m_C(x) = (x-2)^2(x-3).$$

It follows immediately that there are no additional invariant factors for B and C. Since the invariant factors for A divide the minimal polynomial and have product the characteristic polynomial, we see that A has for invariant factors the polynomials $x-2$, $(x-2)(x-3) = x^2-5x+6$. (For 2×2 and 3×3 matrices the determination of the characteristic and minimal polynomials determines all the invariant factors, cf. Exercises 3 and 4.) We conclude that B and C are similar and neither is similar to A. The rational canonical forms are (note $(x-2)^2(x-3) = x^3-7x^2+16x-12$)

$$\begin{pmatrix} 2 & 0 & 0 \\ 0 & 0 & -6 \\ 0 & 1 & 5 \end{pmatrix} \quad \begin{pmatrix} 0 & 0 & 12 \\ 1 & 0 & -16 \\ 0 & 1 & 7 \end{pmatrix} \quad \begin{pmatrix} 0 & 0 & 12 \\ 1 & 0 & -16 \\ 0 & 1 & 7 \end{pmatrix}.$$

(2) In the example above the rational canonical forms were obtained simply by determining the characteristic and minimal polynomials for the matrices. As mentioned, this is sufficient for 2×2 and 3×3 matrices since this information is sufficient to determine all of the invariant factors. For larger matrices, however, this is in general not sufficient (cf. the next example) and more work is required to determine the invariant factors. In this example we again compute the rational canonical form for the matrix A in Example 1 following the two algorithms outlined above. While this is computationally more difficult for this small matrix (as will be apparent), it has the advantage even in this case that it also explicitly computes a matrix P with $P^{-1}AP$ in rational canonical form.

I. (*Invariant Factor Decomposition*) We use row and column operations (in $\mathbb{Q}[x]$) to reduce the matrix

$$xI - A = \begin{pmatrix} x-2 & 2 & -14 \\ 0 & x-3 & 7 \\ 0 & 0 & x-2 \end{pmatrix}$$

to diagonal form. As in the invariant factor decomposition algorithm, we shall use the notation $R_i \leftrightarrow R_j$ to denote the interchange of the i^{th} and j^{th} rows, $R_i + aR_j \mapsto R_i$ if a times the j^{th} row is added to the i^{th} row, simply uR_i if the i^{th} row is multiplied by u (and similarly for columns, using C instead of R). Note also that the first two operations we perform below are rather *ad hoc* and were chosen simply to have integers everywhere in the computation:

$$\begin{pmatrix} x-2 & 2 & -14 \\ 0 & x-3 & 7 \\ 0 & 0 & x-2 \end{pmatrix} \underset{R_1+R_2 \mapsto R_1}{\longrightarrow} \begin{pmatrix} x-2 & x-1 & -7 \\ 0 & x-3 & 7 \\ 0 & 0 & x-2 \end{pmatrix} \longrightarrow$$

$$\underset{C_1-C_2 \mapsto C_1}{\longrightarrow} \begin{pmatrix} -1 & x-1 & -7 \\ -x+3 & x-3 & 7 \\ 0 & 0 & x-2 \end{pmatrix} \underset{-R_1}{\longrightarrow} \begin{pmatrix} 1 & -x+1 & 7 \\ -x+3 & x-3 & 7 \\ 0 & 0 & x-2 \end{pmatrix} \longrightarrow$$

$$\xrightarrow[R_2+(x-3)R_1 \mapsto R_2]{} \begin{pmatrix} 1 & -x+1 & 7 \\ 0 & -x^2+5x-6 & 7(x-2) \\ 0 & 0 & x-2 \end{pmatrix} \xrightarrow[C_2+(x-1)C_1 \mapsto C_2]{} \begin{pmatrix} 1 & 0 & 7 \\ 0 & -x^2+5x-6 & 7(x-2) \\ 0 & 0 & x-2 \end{pmatrix} \longrightarrow$$

$$\xrightarrow[C_3-7C_1 \mapsto C_3]{} \begin{pmatrix} 1 & 0 & 0 \\ 0 & -x^2+5x-6 & 7(x-2) \\ 0 & 0 & x-2 \end{pmatrix} \xrightarrow[-C_2]{} \begin{pmatrix} 1 & 0 & 0 \\ 0 & x^2-5x+6 & 7(x-2) \\ 0 & 0 & x-2 \end{pmatrix} \longrightarrow$$

$$\xrightarrow[R_2-7R_3 \mapsto R_2]{} \begin{pmatrix} 1 & 0 & 0 \\ 0 & x^2-5x+6 & 0 \\ 0 & 0 & x-2 \end{pmatrix} \xrightarrow[C_2 \leftrightarrow C_3]{R_2 \leftrightarrow R_3} \begin{pmatrix} 1 & 0 & 0 \\ 0 & x-2 & 0 \\ 0 & 0 & x^2-5x+6 \end{pmatrix}.$$

This determines the invariant factors $x-2, x^2-5x+6$ for this matrix, which we determined in Example 1 above. Let now V be a 3-dimensional vector space over $\mathbb{Q}$ with basis e_1, e_2, e_3 and let T be the corresponding linear transformation (which defines the action of x on V), i.e.,

$$\begin{aligned} xe_1 &= T(e_1) = 2e_1 \\ xe_2 &= T(e_2) = -2e_1 + 3e_2 \\ xe_3 &= T(e_3) = 14e_1 - 7e_2 + 2e_3. \end{aligned}$$

The row operations used in the reduction above were

$$R_1 + R_2 \mapsto R_1,\ -R_1,\ R_2 + (x-3)R_1 \mapsto R_2,\ R_2 - 7R_3 \mapsto R_2,\ R_2 \leftrightarrow R_3.$$

Starting with the basis $[e_1, e_2, e_3]$ for V and changing it according to the rules given in the text, we obtain

$$\begin{aligned} [e_1, e_2, e_3] &\longrightarrow [e_1, e_2-e_1, e_3] \longrightarrow [-e_1, e_2-e_1, e_3] \\ &\longrightarrow [-e_1-(x-3)(e_2-e_1), e_2-e_1, e_3] \\ &\longrightarrow [-e_1-(x-3)(e_2-e_1), e_2-e_1, e_3+7(e_2-e_1)] \\ &\longrightarrow [-e_1-(x-3)(e_2-e_1), e_3+7(e_2-e_1), e_2-e_1]. \end{aligned}$$

Using the formulas above for the action of x, we see that these last elements are the elements $[0, -7e_1 + 7e_2 + e_3, -e_1 + e_2]$ of V corresponding to the elements $1, x-2$ and x^2-5x+6 in the diagonalized form of $xI - A$, respectively. The elements $f_1 = -7e_1 + 7e_2 + e_3$ and $f_2 = -e_1 + e_2$ are therefore $\mathbb{Q}[x]$-module generators for the two cyclic factors of V in its invariant factor decomposition as a $\mathbb{Q}[x]$-module. The corresponding $\mathbb{Q}$-vector space bases for these two factors are then f_1 and $f_2, xf_2 = Tf_2$, i.e., $-7e_1+7e_2+e_3$ and $-e_1+e_2$, $T(-e_1+e_2) = -4e_1+3e_2$. Then the matrix

$$P = \begin{pmatrix} -7 & -1 & -4 \\ 7 & 1 & 3 \\ 1 & 0 & 0 \end{pmatrix}$$

conjugates A into its rational canonical form:

$$P^{-1}AP = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 0 & -6 \\ 0 & 1 & 5 \end{pmatrix},$$

as one easily checks.

II. (*Converting A Directly to Rational Canonical Form*) We use the row operations involved in the diagonalization of $xI - A$ to determine the matrix P' of the algorithm above:

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \underset{\substack{C_2 - C_1 \\ \mapsto C_2}}{\longrightarrow} \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \underset{-C_1}{\longrightarrow} \begin{pmatrix} -1 & -1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \rightarrow$$

$$\underset{\substack{C_1 - (A-3I)C_2 \\ \mapsto C_1}}{\longrightarrow} \begin{pmatrix} 0 & -1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \underset{\substack{C_3 + 7C_2 \\ \mapsto C_3}}{\longrightarrow} \begin{pmatrix} 0 & -1 & -7 \\ 0 & 1 & 7 \\ 0 & 0 & 1 \end{pmatrix} \underset{C_2 \leftrightarrow C_3}{\longrightarrow} \begin{pmatrix} 0 & -7 & -1 \\ 0 & 7 & 1 \\ 0 & 1 & 0 \end{pmatrix} = P'.$$

Here we have $d_1 = 1$ and $d_2 = 2$, corresponding to the second and third nonzero columns of P', respectively. The columns of P are therefore given by

$$\begin{pmatrix} -7 \\ 7 \\ 1 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}, \quad A\begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} = \begin{pmatrix} -4 \\ 3 \\ 0 \end{pmatrix},$$

respectively, which again gives the matrix P above.

(3) For the 3×3 matrix A it was not necessary to perform the lengthy calculations above merely to determine the rational canonical form (equivalently, the invariant factors), as we saw in Example 1. For $n \times n$ matrices with $n \geq 4$, however, the computation of the characteristic and minimal polynomials is in general not sufficient for the determination of all the invariant factors, so the more extensive calculations of the previous example may become necessary. For example, consider the matrix

$$D = \begin{pmatrix} 1 & 2 & -4 & 4 \\ 2 & -1 & 4 & -8 \\ 1 & 0 & 1 & -2 \\ 0 & 1 & -2 & 3 \end{pmatrix}.$$

A short computation shows that the characteristic polynomial of D is $(x-1)^4$. The possible minimal polynomials are then $x-1$, $(x-1)^2$, $(x-1)^3$ and $(x-1)^4$. Clearly $D - I \neq 0$ and another short computation shows that $(D-I)^2 = 0$, so the minimal polynomial for D is $(x-1)^2$. There are then two possible sets of invariant factors:

$$x-1, x-1, (x-1)^2 \qquad \text{and} \qquad (x-1)^2, (x-1)^2.$$

To determine the invariant factors for D we apply the procedure of the previous example to the 4×4 matrix

$$xI - D = \begin{pmatrix} x-1 & -2 & 4 & -4 \\ -2 & x+1 & -4 & 8 \\ -1 & 0 & x-1 & 2 \\ 0 & -1 & 2 & x-3 \end{pmatrix}.$$

The diagonal matrix obtained from this matrix by elementary row and column operations is the matrix

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & (x-1)^2 & 0 \\ 0 & 0 & 0 & (x-1)^2 \end{pmatrix},$$

which shows that the invariant factors for D are $(x-1)^2$, $(x-1)^2$ (one series of elementary row and column operations which diagonalize $xI - D$ are $R_1 \leftrightarrow R_3$, $-R_1$,

$R_2+2R_1 \mapsto R_2$, $R_3-(x-1)R_1 \mapsto R_3$, $C_3+(x-1)C_1 \mapsto C_3$, $C_4+2C_1 \mapsto C_4$, $R_2 \leftrightarrow R_4$, $-R_2$, $R_3+2R_2 \mapsto R_3$, $R_4-(x+1)R_2 \mapsto R_4$, $C_3+2C_2 \mapsto C_3$, $C_4+(x-3)C_2 \mapsto C_4$).

I. (*Invariant Factor Decomposition*) If e_1, e_2, e_3, e_4 is a basis for V in this case, then using the row operations in this diagonalization as in the previous example we see that the generators of V corresponding to the factors above are $(x-1)e_1 - 2e_2 - e_3 = 0$, $-2e_1+(x+1)e_2-e_4 = 0, e_1, e_2$. Hence a vector space basis for the two direct factors in the invariant decomposition of V in this case is given by e_1, Te_1 and e_2, Te_2 where T is the linear transformation defined by D, i.e., $e_1, e_1+2e_2+e_3$ and $e_2, 2e_1-e_2+e_4$. The corresponding matrix P relating these bases is

$$P = \begin{pmatrix} 1 & 1 & 0 & 2 \\ 0 & 2 & 1 & -1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

so that $P^{-1}DP$ is in rational canonical form:

$$P^{-1}DP = \begin{pmatrix} 0 & -1 & 0 & 0 \\ 1 & 2 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 2 \end{pmatrix}$$

as can easily be checked.

II. (*Converting D Directly to Rational Canonical Form*) As in Example 2 we determine the matrix P' of the algorithm from the row operations used in the diagonalization of $xI - D$:

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \xrightarrow[C_1 \leftrightarrow C_3]{} \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \xrightarrow[-C_1]{} \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \rightarrow$$

$$\xrightarrow[\substack{C_1-2C_2 \\ \mapsto C_1}]{} \begin{pmatrix} 0 & 0 & 1 & 0 \\ -2 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \xrightarrow[\substack{C_1+(D-I)C_3 \\ \mapsto C_1}]{} \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \xrightarrow[C_2 \leftrightarrow C_4]{} \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \rightarrow$$

$$\xrightarrow[-C_2]{} \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{pmatrix} \xrightarrow[\substack{C_2-2C_3 \\ \mapsto C_2}]{} \begin{pmatrix} 0 & -2 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{pmatrix} \xrightarrow[\substack{C_2+(D+I)C_4 \\ \mapsto C_2}]{} \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} = P'.$$

Here we have $d_1 = 2$ and $d_2 = 2$, corresponding to the third and fourth nonzero columns of P'. The columns of P are therefore given by

$$\begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad D\begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 1 \\ 0 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \quad D\begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 2 \\ -1 \\ 0 \\ 1 \end{pmatrix},$$

respectively, which again gives the matrix P above.

(4) In this example we determine all similarity classes of matrices A with entries from $\mathbb{Q}$ with characteristic polynomial $(x^4-1)(x^2-1)$. First note that any matrix with a degree