RINGS, MODULES AND LINEAR ALGEBRA

N. P. STRICKLAND

1. Introduction

At the end of the course, you should be able to prove the following:

1. Let G be an Abelian group of order p^3, where p is prime. Then G is isomorphic to either Z_{p^3}, $\mathbb{Z}_{p^2} \times \mathbb{Z}_p$ or $\mathbb{Z}_p \times \mathbb{Z}_p \times \mathbb{Z}_p$.
2. Let A be a 3×3 matrix of rational numbers satisfying $(A + I)^3 = 0$. Then A is conjugate to one of the following three matrices:

$$\begin{pmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & -1 \end{pmatrix} \qquad \begin{pmatrix} -1 & 1 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix} \qquad \begin{pmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix}$$

   (In other words, there is an invertible matrix P such that PAP^{-1} is one of the three matrices listed.)
3. Let $f\colon \mathbb{R} \to \mathbb{C}$ satisfy the differential equation $f'''' + f = 0$. Then $f(x) = ae^x + be^{-x} + ce^{ix} + de^{-ix}$ for some constants a, b, c, d.

The remarkable thing is that all these problems can be addressed using essentially the same ideas: the theory of modules over a Euclidean domain.

2. Rings and Fields

A *commutative ring* is a set R of things that can be added, negated and multiplied in a sensible way to get new elements of R. More precisely, we require that the following axioms be satisfied:

(a) If $a, b \in R$ then $a + b \in R$. [closure under addition]
(b) There is an element $0 \in R$ such that $a + 0 = a$ for all $a \in R$. [additive identity]
(c) For each element $a \in R$ there is an element $-a \in R$ such that $a + (-a) = 0$. [additive inverses]
(d) $a + (b + c) = (a + b) + c$ for all $a, b, c \in R$. [associativity of addition]
(e) $a + b = b + a$ for all $a, b \in R$. [commutativity of addition]
(f) If $a, b \in R$ then $ab \in R$. [closure under multiplication]
(g) There is an element $1 \in R$ such that $1a = a$ for all $a \in R$. [multiplicative identity]
(h) $a(bc) = (ab)c$ for all $a, b, c \in R$. [associativity of multiplication]
(i) $ab = ba$ for all $a, b \in R$. [commutativity of multiplication]
(j) $a(b + c) = ab + ac$ for all $a, b, c \in R$. [distributivity]

Strictly speaking, we should say that a ring is a set R *together with a definition of addition, negation and multiplication* such that the axioms hold. In all the examples that we will consider, there is a unique obvious way to define these operations.

We will not consider noncommutative rings in this course, so we will just use the word "ring" to mean "commutative ring". We will use without comment various standard consequences of the axioms, such as the facts that $-(-a) = a$, $0.a = 0$ and $(-1).a = -a$.

Definition 2.1. A ring R is an *integral domain* if $1 \neq 0$, and whenever $a, b \neq 0$ we also have $ab \neq 0$.

Date: February 11, 2002.

Definition 2.2. An element a in a ring R is *invertible* if there is an element $b \in R$ such that $ab = 1$. If so, then this element b is unique and we call it a^{-1}. A *field* is a ring K such that $1 \neq 0$ and every nonzero element is invertible. Every field is an integral domain.

Example 2.3. The set $\mathbb{Z}$ of integers is a ring. If we add, subtract or multiply any two integers, we get another integer, so axioms (a), (c) and (f) hold. The numbers 0 and 1 are integers so axioms (b) and (g) hold. The remaining axioms are familiar properties of addition, subtraction and multiplication. It is also clear that $\mathbb{Z}$ is an integral domain.

Example 2.4. The set $\mathbb{N}$ of natural numbers is not a ring, because $0 \notin \mathbb{N}$, so there is no additive identity, in other words axiom (b) does not hold. Moreover, if $n \in \mathbb{N}$ then $-n \notin \mathbb{N}$, so $\mathbb{N}$ does not have additive inverses.

Example 2.5. The set $2\mathbb{Z}$ of even integers is not a ring, because it does not contain the multiplicative identity element 1.

Example 2.6. The set $\mathbb{Q}$ of rational numbers, the set $\mathbb{R}$ of real numbers, and the set $\mathbb{C}$ of complex numbers, are all fields.

Example 2.7. The set $X = \mathbb{R} \setminus \mathbb{Q}$ of irrational numbers is not a ring. Indeed, we have $\pi \in X$ and $-\pi \in X$ but $0 = \pi + (-\pi) \notin X$, so X is not closed under addition. Moreover, $\sqrt{2} \in X$ but $\sqrt{2}.\sqrt{2} = 2 \notin X$, so X is not closed under multiplication. Even more obviously, X contains neither 0 nor 1, so it does not have an additive identity or a multiplicative identity.

Example 2.8. For any natural number n, the set $\mathbb{Z}_n = \{\overline{0}, \overline{1}, \dots, \overline{n-1}\}$ of integers modulo n is a ring. For example, we have

$$\begin{aligned} \mathbb{Z}_4 &= \{\overline{0}, \overline{1}, \overline{2}, \overline{3}\} \\ \overline{2} + \overline{3} &= \overline{5} = \overline{1} \\ \overline{1} - \overline{2} &= \overline{-1} = \overline{3} \\ \overline{2}\,.\,\overline{3} &= \overline{6} = \overline{2}. \end{aligned}$$

Note that $\mathbb{Z}_4$ is not an integral domain, because $\overline{2} \neq 0$ but $\overline{2}.\overline{2} = 0$. In general, it can be shown that when n is prime, $\mathbb{Z}_n$ is a field (and thus an integral domain), but when n is not prime, $\mathbb{Z}_n$ is neither a field nor an integral domain.

Example 2.9. We write $\mathbb{Z}[x]$ for the set of all polynomials with integer coefficients. For example, $7x^3 - 22x + 3$ and x^{1001} are elements of $\mathbb{Z}[x]$ but $(x+1)/(x-1)$ and $x^2 - 1/2$ and $x - x^{-1}$ are not. The general form of an element of $\mathbb{Z}[x]$ is $f(x) = a_0 + a_1 x + \dots + a_n x^n$ for some integer $n \geq 0$ and integers $a_0, \dots, a_n$. Integers are polynomials of degree zero, so $\mathbb{Z} \subseteq \mathbb{Z}[x]$. The usual operations of addition, multiplication and negation of polynomials make $\mathbb{Z}[x]$ into a ring.

More generally, given any ring R we can consider the ring $R[x]$ of polynomials with coefficients in R. For example, we can consider $f(x) = \overline{2}x^2 + \overline{3} \in \mathbb{Z}_6[x]$ and $g(x) = \overline{3}x + \overline{2} \in \mathbb{Z}_6[x]$ and we find that

$$\begin{aligned} f(x)g(x) &= \overline{6}x^3 + \overline{4}x^2 + \overline{9}x + \overline{6} \\ &= \overline{4}x^2 + \overline{3}x. \end{aligned}$$

This gives us rings $\mathbb{Z}[x] \subseteq \mathbb{Q}[x] \subseteq \mathbb{R}[x] \subseteq \mathbb{C}[x]$. We can also use more than one variable; for example, we have a ring $\mathbb{Q}[x, y, z]$ containing elements like $(x^2 + y^2 + z^2)/4$ or $1 + xyz$ (but not x/y or $\sqrt{2}x$ or e^{x+y}).

Example 2.10. Let D denote the operation of differentiation with respect to t, so $D(t^3) = 3t^2$, $D(\sin(t)) = \cos(t)$, $D^3(f(t)) = f'''(t)$ and so on. Using this we can build more complicated operators like $(D^2 + 2D + 3)(f(t)) = f''(t) + 2f'(t) + 3f(t)$ and so on. By a *differential operator* we mean an operation of the form $a_0 + a_1 D + \dots + a_n D^n$ for some $n \geq 0$ and some list of coefficients $a_i \in \mathbb{R}$. The set of differential operators is the polynomial ring $\mathbb{R}[D]$; its is essentially the same

as $\mathbb{R}[x]$ except for the notation, and the fact that the elements are interpreted as operators rather than functions.

Remark 2.11. Here we are only considering linear ordinary differential operators with constant coefficients. For more serious work on differential equations one needs to work with more general operators, which generally form noncommutative rings.

Example 2.12. Let p be a prime number. Let $\mathbb{Z}_{(p)}$ be the set of rational numbers x that can be written in the form a/b, where a and b are integers and b is not divisible by p. For example, $123/101$ and $-4/12 = (-1)/3$ and $8 = 8/1$ are elements of $\mathbb{Z}_{(2)}$, but $5/6$ and $1/8$ are not. As any integer n can be written as $n/1$ and 1 is not divisible by p, we see that $\mathbb{Z} \subseteq \mathbb{Z}_{(p)}$. Now suppose that $x, y \in \mathbb{Z}_{(p)}$, so we can write $x = a/b$ and $y = c/d$ for some integers a, b, c and d, where b and d are not divisible by p. As p is prime this means that bd is also not divisible by p. We have

$$\begin{aligned} x + y &= (ad + bc)/(bd) \\ xy &= (ac)/(bd) \\ -x &= -a/b. \end{aligned}$$

As $ad + bc$, ac, $-a$, b and bd are integers, and bd and b are not divisible by p, this means that $x + y$, xy and $-x$ lie in $\mathbb{Z}_{(p)}$. Thus $\mathbb{Z}_{(p)}$ is a subring of $\mathbb{Q}$, called the ring of integers localized at p. (There is a long story coming from algebraic geometry that explains why the word "localized" is appropriate.)

Example 2.13. We write $\mathbb{Z}[i]$ for the set of complex numbers of the form $a + bi$, where a and b are integers (possibly zero). Thus 7, $6 - 4i$ and $12i$ are elements of $\mathbb{Z}[i]$, but $2/3$ and $1 - i/5$ are not. Note that

$$\begin{aligned} (a + bi) + (c + di) &= (a + c) + (b + d)i \\ (a + bi)(c + di) &= (ac - bd) + (ad + bc)i \\ -(a + bi) &= (-a) + (-b)i. \end{aligned}$$

It follows easily that $\mathbb{Z}[i]$ is closed under addition, multiplication and negation, so it is a subring of $\mathbb{C}$. The elements of $\mathbb{Z}[i]$ are called *Gaussian integers*.

3. Modules

Definition 3.1. Let R be a ring. A *module* over R is a set M of things with a definition of $m + n$ for all $m, n \in M$ and a definition of am for all $a \in R$ and $m \in M$ such that the following axioms are satisfied:

(a) If $m, n \in M$ then $m + n \in M$. [closure under addition]
(b) There is an element $0 \in M$ such that $m + 0 = m$ for all $m \in M$. [additive identity]
(c) For each $m \in M$ there is an element $-m \in M$ such that $m + (-m) = 0$. [additive inverses]
(d) $m + (n + p) = (m + n) + p$ for all $m, n, p \in M$. [associativity of addition]
(e) $m + n = n + m$ for all $m, n \in M$. [commutativity of addition]
(f) If $a \in R$ and $m \in M$ then $am \in M$. [closure of M under multiplication by R]
(g) $1.m = m$ for all $m \in M$.
(h) $(ab)m = a(bm)$ for all $a, b \in R$ and $m \in M$. [associativity of multiplication]
(i) $(a + b)m = am + bm$ for all $a, b \in R$ and $m \in M$. [left distributivity of multiplication]
(j) $a(m + n) = am + an$ for all $a \in R$ and $m, n \in M$. [right distributivity of multiplication]

Remark 3.2. Note that axioms (a) to (e) say that M is in particular an Abelian group under addition.

Example 3.3. Let R be any ring, and let d be a natural number. We then write R^d for the set of d-tuples $(x_1, \dots, x_d)$ with $x_1, \dots, x_d \in R$. We make R^d into a module over R by defining

$$\begin{aligned} (x_1, \dots, x_d) + (y_1, \dots, y_d) &= (x_1 + y_1, \dots, x_d + y_d) \\ a(x_1, \dots, x_d) &= (ax_1, \dots, ax_d). \end{aligned}$$

It is straightforward to check that the axioms are satisfied. In particular, the case $d = 1$ says that we can regard R as a module over itself.

If R is a field, then an R-module is just a vector space over R. Modules are just the natural generalization of vector spaces defined over arbitrary rings rather than just fields. It is a basic fact of linear algebra that if K is a field and V is a vector space over K with a finite spanning set, then V is isomorphic to K^d for some integer d, called the *dimension* of V. The situation for modules over non-fields is more complicated; a module is usually not isomorphic to R^d for any d. The next simplest case after fields is when R is a Euclidean domain, and most of the course will be devoted to the study of modules over such rings.

Proposition 3.4. *A $\mathbb{Z}$-module is just an Abelian group. More precisely, if M is an Abelian group (with the group operation written as addition) then there is a unique way to define am for all $a \in \mathbb{Z}$ and $m \in M$ such that axioms (f) to (j) hold, making M a $\mathbb{Z}$-module.*

Sketch proof. Rather than giving a complete proof of this, we will give an outline of the argument with examples.

The basic idea is very simple. We just define

$$\begin{aligned} 3m &= m + m + m \\ -5m &= -(m + m + m + m + m) = (-m) + (-m) + (-m) + (-m) + (-m) \end{aligned}$$

and so on. This defines multiplication (of integers by group elements) in terms of addition and negation of group elements. We actually have no choice about these definitions if we want the axioms to be satisfied: as $3 = 1 + 1 + 1$, axiom (i) says we we must have $3m = (1 + 1 + 1)m = 1m + 1m + 1m$, and axiom (g) says that $1m = m$ so we must have $3m = m + m + m$, and so on.

We now need to check that axioms (f) to (j) are satisfied. Axioms (f) and (g) are immediate. The remaining axioms are easy to check when a and b are nonnegative: for example

$$\begin{aligned} 2(3m) &= 2(m + m + m) \\ &= (m + m + m) + (m + m + m) \\ &= m + m + m + m + m + m \\ &= (2 \times 3)m \\ 2m + 3m &= (m + m) + (m + m + m) \\ &= m + m + m + m + m \\ &= (2 + 3)m \\ 3(m + n) &= (m + n) + (m + n) + (m + n) \\ &= (m + m + m) + (n + n + n) \\ &= 3m + 3n. \end{aligned}$$

If we allow a or b to be negative then there are quite a few more cases to check depending on the various possible combinations of signs, but they are all quite straightforward. For example

$$\begin{aligned} 5m + (-3)m &= (m + m + m + m + m) + ((-m) + (-m) + (-m)) \\ &= m + m + (m + (-m)) + (m + (-m)) + (m + (-m)) \\ &= m + m \\ &= (5 + (-3))m. \end{aligned}$$

□

We will next give an example involving differential operators. To avoid annoying technicalities, it is best to restrict attention to functions that can be differentiated as many times as we like. We therefore introduce the following definition.

Definition 3.5. A function $f\colon \mathbb{R} \to \mathbb{R}$ is *smooth* if the n'th derivatives $f^{(n)}(t)$ are defined and continuous everywhere on $\mathbb{R}$ for all $n \geq 0$. In particular, the function $f = f^{(0)}$ itself must be defined and continuous everywhere.

For example, $\sin(t)$, $\cos(t)$, e^t, t^2 and so on are smooth. However, the functions $1/t$ and $\log(t)$ are not defined at $t = 0$, so they are not smooth. The function $f(t) = |t|$ is defined and continuous everywhere and $f'(t) = -1$ for $t < 0$ and $f'(t) = 1$ for $t > 0$ but $f'(0)$ is undefined so f is not smooth. Similarly, if $g(t) = t^{1/3}$ then $g'(t) = t^{-2/3}/3$, which is undefined at $t = 0$ so g is not smooth.

We write $C^\infty(\mathbb{R}, \mathbb{R})$ for the set of all smooth functions from $\mathbb{R}$ to $\mathbb{R}$. If f and g are smooth and a is constant then one can show that $f + g$ and af are smooth. It follows that $C^\infty(\mathbb{R}, \mathbb{R})$ is a vector space over $\mathbb{R}$. Similarly, the set $C^\infty(\mathbb{R}, \mathbb{C})$ of smooth functions from $\mathbb{R}$ to $\mathbb{C}$ is a vector space over $\mathbb{C}$.

Example 3.6. $C^\infty(\mathbb{R}, \mathbb{R})$ is a module over the ring $\mathbb{R}[D]$ of differential operators. For an operator $p(D) = a_0 + a_1 D + \ldots + a_n D^n$ and a smooth function $f(t)$, the product $p(D)f$ is defined by the usual rule

$$p(D)f = a_0 f + a_1 f' + a_2 f'' + \ldots + a_n f^{(n)}.$$

One can check the axioms directly, or use a more abstract approach discussed in the next section. Similarly, $C^\infty(\mathbb{R}, \mathbb{C})$ is a module over $\mathbb{C}[D]$.

Example 3.7.

$$\begin{aligned}(1 + D + D^2).(1 + t + t^2) &= (1 + t + t^2) + (1 + t + t^2)' + (1 + t + t^2)'' \\ &= (1 + t + t^2) + (1 + 2t) + (2) \\ &= 4 + 3t + t^2.\end{aligned}$$

Example 3.8. Consider the function $f(t) = t\sin(t)$; I claim that $(D^2 + 1)^2 f = 0$. Indeed, we have

$$\begin{aligned}f'(t) &= \sin(t) + t\cos(t) \\ f''(t) &= 2\cos(t) - t\sin(t) \\ ((D^2 + 1)f)(t) &= f(t) + f''(t) = 2\cos(t)\end{aligned}$$

We also have $\cos'(t) = -\sin(t)$ and so $\cos''(t) = -\cos(t)$ so $(D^2 + 1)\cos = 0$. It follows that $(D^2 + 1)^2 f = (D^2 + 1)(2\cos) = 0$.

Example 3.9. Consider a function of the form $f(t) = e^{\lambda t} + e^{\mu t}$. I claim that

$$(p(D)f)(t) = p(\lambda)e^{\lambda t} + p(\mu)e^{\mu t}.$$

Indeed, we have

$$\begin{aligned}f'(t) &= \lambda e^{\lambda t} + \mu e^{\mu t} \\ f''(t) &= \lambda^2 e^{\lambda t} + \mu^2 e^{\mu t}\end{aligned}$$

and more generally $f^{(k)}(t) = \lambda^k e^{\lambda t} + \mu^k e^{\mu t}$ (as one can easily check by induction). If $p(D) = a_0 + a_1 D + \ldots + a_m D^m$ then we have

$$\begin{aligned}(p(D)f)(t) &= a_0(e^{\lambda t} + e^{\mu t}) + a_1(\lambda e^{\lambda t} + \mu e^{\mu t}) + \ldots + a_m(\lambda^m e^{\lambda t} + \mu^m e^{\mu t}) \\ &= (a_0 + a_1\lambda + \ldots + a_m\lambda^m)e^{\lambda t} + (a_0 + a_1\mu + \ldots + a_m\mu^m)e^{\mu t} \\ &= p(\lambda)e^{\lambda t} + p(\mu)e^{\mu t}.\end{aligned}$$

4. Modules over polynomial rings

We next consider modules over $K[x]$, where K is a field. The upshot here is that the study of modules over $K[x]$ is essentially the same as the study of square matrices over K, or of endomorphisms of vector spaces over K.

We start with some comments about the process of "substituting a matrix into a polynomial". Let K be a field, and let A be an $n \times n$ matrix over K. Using the usual matrix multiplication we can define A^2, A^3 and so on; all of these are again $n \times n$ matrices over K. Thus, given a

polynomial $f(x) = a_0 + a_1x + \ldots + a_dx^d \in K[x]$ we can define another $n \times n$ matrix $f(A)$ by $f(A) = a_0I + a_1A + \ldots + a_dA^d$.

Example 4.1. If $A = \left(\begin{smallmatrix} 1 & 2 \\ 3 & 4 \end{smallmatrix}\right)$ and $f(x) = 7 + 6x + 5x^2$ then $A^2 = \left(\begin{smallmatrix} 7 & 10 \\ 15 & 22 \end{smallmatrix}\right)$ and so

$$\begin{aligned} f(A) &= 7\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right) + 6\left(\begin{smallmatrix} 1 & 2 \\ 3 & 4 \end{smallmatrix}\right) + 5\left(\begin{smallmatrix} 7 & 10 \\ 15 & 22 \end{smallmatrix}\right) \\ &= \left(\begin{smallmatrix} 7 & 0 \\ 0 & 7 \end{smallmatrix}\right) + \left(\begin{smallmatrix} 6 & 12 \\ 18 & 24 \end{smallmatrix}\right) + \left(\begin{smallmatrix} 35 & 50 \\ 75 & 110 \end{smallmatrix}\right) \\ &= \left(\begin{smallmatrix} 48 & 62 \\ 93 & 141 \end{smallmatrix}\right). \end{aligned}$$

Example 4.2. Consider a diagonal matrix $A = \left(\begin{smallmatrix} \lambda & 0 \\ 0 & \mu \end{smallmatrix}\right)$. Then

$$A^2 = \begin{pmatrix} \lambda & 0 \\ 0 & \mu \end{pmatrix}\begin{pmatrix} \lambda & 0 \\ 0 & \mu \end{pmatrix} = \begin{pmatrix} \lambda^2 & 0 \\ 0 & \mu^2 \end{pmatrix},$$

and more generally it is not hard to see that

$$A^k = \begin{pmatrix} \lambda^k & 0 \\ 0 & \mu^k \end{pmatrix}.$$

(Exercise: prove this by induction.) It follows that

$$\begin{aligned} f(A) &= a_0\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + a_1\begin{pmatrix} \lambda & 0 \\ 0 & \mu \end{pmatrix} + \ldots + a_d\begin{pmatrix} \lambda^d & 0 \\ 0 & \mu^d \end{pmatrix} \\ &= \begin{pmatrix} a_0 + a_1\lambda + \ldots a_d\lambda^d & 0 \\ 0 & a_0 + a_1\mu + \ldots a_d\mu^d \end{pmatrix} \\ &= \begin{pmatrix} f(\lambda) & 0 \\ 0 & f(\mu) \end{pmatrix}. \end{aligned}$$

More generally, if A is an $n \times n$ matrix with entries $\lambda_1, \ldots, \lambda_n$ on the diagonal and zeros elsewhere, then $f(A)$ has entries $f(\lambda_1), \ldots, f(\lambda_n)$ on the diagonal and zeros elsewhere.

Example 4.3. Consider the matrix $A = \left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$. It is easy to check that

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & k \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & k+1 \\ 0 & 1 \end{pmatrix},$$

and thus that $A^k = \left(\begin{smallmatrix} 1 & k \\ 0 & 1 \end{smallmatrix}\right)$ for all k. It follows that

$$\begin{aligned} f(A) &= a_0\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + a_1\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} + \ldots + a_d\begin{pmatrix} 1 & d \\ 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} a_0 + \ldots + a_d & a_1 + 2a_2 + \ldots + da_d \\ 0 & a_0 + \ldots + a_d \end{pmatrix}. \end{aligned}$$

Note that $f(1) = a_0 + \ldots + a_d$. Note also that the derivative $f'(x)$ is given by $f'(x) = a_1 + 2a_2x + \ldots + da_dx^{d-1}$, so that $f'(1) = a_1 + 2a_2 + \ldots + da_d$. We can thus rewrite the above result as

$$f(A) = \begin{pmatrix} f(1) & f'(1) \\ 0 & f(1) \end{pmatrix}.$$

We next need to check that some things work out as they "ought" to when we substitute matrices into polynomials. (Recall that matrix multiplication is noncommutative, there are nonzero matrices whose square is zero, and numerous other funny things can happen; so we need to be on our guard.)

Proposition 4.4. *Let A be an $n \times n$ matrix over a field K. Then for any two polynomials $f, g \in K[x]$ we have*

$$\begin{aligned} (f+g)(A) &= f(A) + g(A) \\ (fg)(A) &= f(A)g(A). \end{aligned}$$

Proof. Suppose that $f(x) = \sum_i a_i x^i$ and $g(x) = \sum_j b_j x^j$. Then $(f+g)(x) = \sum_i c_i x^i$ where $c_i = a_i + b_i$, and

$$(fg)(x) = (\sum_i a_i x^i)(\sum_j b_j x^j) = \sum_{i,j} a_i b_j x^{i+j} = \sum_k d_k x^k,$$

where $d_k = \sum_{i=0}^{k} a_i b_{k-i}$.

Thus

$$\begin{aligned}(f+g)(A) &= \sum_i c_i A^i \\ &= \sum_i (a_i A^i + b_i A^i) \\ &= \sum_i a_i A^i + \sum_i b_i A^i \\ &= f(A) + g(A).\end{aligned}$$

Similarly

$$\begin{aligned}(fg)(A) &= \sum_k d_k A^k \\ &= \sum_k \sum_{i=0}^{k} a_i b_{k-i} A^k \\ &= \sum_k \sum_{i=0}^{k} (a_i A^i)(b_{k-i} A^{k-i}) \\ &= \sum_i \sum_j (a_i A^i)(b_j A^j) \\ &= \sum_i a_i A^i \sum_j b_j A^j \\ &= f(A)g(A).\end{aligned}$$

□

We are now ready to construct some modules over $K[x]$.

Construction 4.5. Let A be an $n \times n$ matrix over a field K; we will use this to define a module M_A over $K[x]$. The elements of M_A are just the vectors $v = (v_1, \ldots, v_n)$ of length n over K, so $M_A = K^n$ as a set. Addition and subtraction of vectors is defined in the usual way. All that is left is to define the product fv for $f \in K[x]$ and $v \in K^n$, which we do by the formula $fv = f(A)v$. Here $f(A)$ is an $n \times n$ matrix, so the right hand side is defined by the ordinary multiplication of vectors by matrices.

We need to check the module axioms. Axioms (a) to (e) only involve addition and negation so they are clear. Axiom (f) is also clear because fv is certainly a vector in K^n. If $f(x)$ is constant polynomial 1, then $f(A)$ is the identity matrix, so $fv = Iv = v$ for all v; this gives axiom (g). For axiom (h) we recall that $(fg)(A) = f(A)g(A)$ so

$$\begin{aligned}(fg)v &= (fg)(A)v \\ &= f(A)g(A)v \\ &= f(A)(gv) \\ &= f(gv).\end{aligned}$$

Similarly, axiom (i) follows from the fact that $(f+g)(A) = f(A) + g(A)$. Finally, axiom (j) is clear, because $B(v+w) = Bv + Bw$ for any matrix B and any vectors v and w.

Remark 4.6. Let A and B be two different $n \times n$ matrices. Then M_A and M_B have the same elements but the multiplication rules in M_A and M_B are different, so M_A and M_B are different modules.

Example 4.7. Let A be the matrix $\left(\begin{smallmatrix} 2 & 0 \\ 0 & 3 \end{smallmatrix}\right)$ over $\mathbb{Q}$, so that $f(A) = \left(\begin{smallmatrix} f(2) & 0 \\ 0 & f(3) \end{smallmatrix}\right)$ (by Example 4.2). Then $M_A = \mathbb{Q}^2$, with the multiplication rule $f.(s,t) = (f(2)s, f(3)t)$. For example, if $g(x) = x^2 - 6$ then $g(2) = -2$ and $g(3) = 3$ so we have

$$(x^2 - 6)(10, 11) = (-2 \times 10, 3 \times 11) = (-20, 33).$$

Example 4.8. Let A be the matrix $\left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$ over $\mathbb{Q}$, so that $f(A) = \left(\begin{smallmatrix} f(1) & f'(1) \\ 0 & f(1) \end{smallmatrix}\right)$ (by Example 4.3). Then M_A is the set $\mathbb{Q}^2$ with the group operation $f.(s,t) = (f(1)s + f'(1)t, f(1)t)$. For example, if $g(x) = x^2 - 6$ then $g(1) = -5$ and $g'(1) = 2$ so we have

$$(x^2 - 6)(10, 11) = (-5 \times 10 + 2 \times 11, -5 \times 11) = (-28, -55).$$

Example 4.9. The simplest examples are where A is just a 1×1 matrix, or in other words just an element $\lambda \in K$. The module M_λ is just a copy of K, with the multiplication rule $f.a = f(\lambda)a$. For example, the polynomial $f(x) = 1 + x + x^2$ satisfies $f(2) = 7$, so in the module M_2 over $\mathbb{Q}[x]$ we have $f.6 = 7 \times 6 = 42$.

There is a well-known correspondence between matrices and endomorphisms, and for many purposes it is more natural to use the latter. Let V be a vector space over a field K, and let ϕ be an endomorphism of V (in other words, a linear map from V to itself). Then we can define $\phi^2(v) = \phi(\phi(v))$ to get a new endomorphism of V, and similarly we can define ϕ^k for all $k \geq 0$. More generally, for any polynomial $f(x) = a_0 + a_1 x + \ldots + a_d x^d \in K[x]$ we can define an endomorphism $f(\phi)$ by

$$f(\phi)(v) = a_0 v + a_1 \phi(v) + \ldots + a_d \phi^d(v).$$

We can then make V into a module over $K[x]$ by defining $fv = f(\phi)(v)$.

Example 4.10. We can define a map $\partial\colon C^\infty(\mathbb{R}, \mathbb{R}) \to C^\infty(\mathbb{R}, \mathbb{R})$ by $\partial(f) = f'$. As $(f+g)' = f' + g'$ and $(cf)' = cf'$ for constant c, we see that ∂ is an $\mathbb{R}$-linear endomorphism of $C^\infty(\mathbb{R}, \mathbb{R})$, making $C^\infty(\mathbb{R}, \mathbb{R})$ into a module over $\mathbb{R}[x]$. The multiplication rule is as follows: if $p(x) = \sum_i a_i x^i$ and $f(t) \in C^\infty(\mathbb{R}, \mathbb{R})$ then

$$\begin{aligned} p.f &= a_0 \partial^0(f) + a_1 \partial^1(f) + a_2 \partial^2(f) + \ldots \\ &= a_0 f + a_1 f' + a_2 f'' + \ldots . \end{aligned}$$

Thus, this example is just the same as Example 2.10, with a slightly different viewpoint and less natural notation.

We explained above how a vector space V over K with an endomorphism ϕ can be regarded as a $K[x]$-module. We conclude this section by showing that *every* $K[x]$-module arises in this way.

Indeed, let M be a module over $K[x]$. As mentioned previously, axioms (a) to (e) say that M is an Abelian group under addition. Also, if $a \in K$ then we can regard a as a constant polynomial, so am is defined for all $m \in M$. As M is a module over $K[x]$, axioms (f) to (j) are valid for all polynomials a and b, so certainly they are valid for the special case of constant polynomials. Thus, we can regard M as a module over K. A module over a field is the same thing as a vector space, so M is a vector space over K.

Next, if $m \in M$ then xm is another element of M, so we can define a function $\phi\colon M \to M$ by $\phi(m) = xm$. I claim that this is a K-linear endomorphism. Indeed, for any $m, n \in M$ we have $x(m+n) = xm + xn$ by the right distributivity law, which means that $\phi(m+n) = \phi(m) + \phi(n)$. Moreover, for $a \in K$ we have $ax = xa$, so

$$a\phi(m) = a(xm) = (ax)m = (xa)m = x(am) = \phi(am)$$

(using axiom (h) twice). This shows that ϕ is linear, as claimed. Now consider a polynomial $f(x) = \sum_i a_i x^i \in K[x]$. I claim that $fm = \sum_i a_i \phi^i(m) = f(\phi)(m)$ for all $m \in M$. Indeed, we

have

$$(x^2)m = x(xm) = x\phi(m) = \phi(\phi(m)) = \phi^2(m)$$
$$(x^3)m = x(x^2m) = x\phi^2(m) = \phi(\phi^2(m)) = \phi^3(m).$$

Extending this by induction, we see that $x^k m = \phi^k(m)$ for all k. Thus

$$\begin{aligned} fm &= (\sum_i a_i x^i)m \\ &= \sum_i a_i x^i m \\ &= \sum_i a_i \phi^i(m) \\ &= f(\phi)(m). \end{aligned}$$

Thus, the module structure is obtained from the endomorphism ϕ in the way considered previously.

5. General module theory

Let R be a ring.

Definition 5.1. Let M be an R-module. A *submodule* of M is a subset $N \subseteq M$ such that

(a) $0 \in N$
(b) If $n, m \in N$ then $n + m \in N$ (ie N is closed under addition)
(c) If $n \in N$ and $a \in R$ then $an \in N$ (ie N is closed under multiplication by elements of R).

Note that if N is a submodule and $n \in N$ then $-n = (-1)n \in N$, so N is closed under negation and thus is a subgroup of M under addition. It is easy to see that N can itself be considered as an R-module.

Example 5.2. If R is a field, then modules are just the same as vector spaces, and submodules are just the same as vector subspaces.

Example 5.3. If $R = \mathbb{Z}$, then modules are just the same as Abelian groups, and submodules are just the same as subgroups.

Example 5.4. If M is a module over any ring R, it is clear that $\{0\}$ and M itself are submodules of M.

Example 5.5. Let V be a vector space over a field K, equipped with a K-linear endomorphism $\phi\colon V \to V$. We regard V as a $K[x]$-module in the usual way. We say that a subset $W \subseteq V$ is *stable under* ϕ if $\phi(w) \in W$ for all $w \in W$ (or more briefly, if $\phi(W) \subseteq W$).

I claim that a subset $W \subseteq V$ is a $K[x]$-submodule if and only if it is a vector subspace and is stable under ϕ. Indeed, suppose that W is a submodule. Then it is certainly closed under addition and under multiplication by constant polynomials (ie elements of K) so it is a vector subspace. Also, it is closed under multiplication by x, so for $w \in W$ we have $\phi(w) = xw \in W$; this shows that W is stable under ϕ, as claimed.

Conversely, suppose that W is a vector subspace and is stable under ϕ. Clearly W is closed under addition. For any $w \in W$ we have $\phi(w) \in W$. Thus $\phi^2(w) = \phi(\phi(w)) = \phi(\text{an element of } W) =$ another element of W , so $\phi^2(W) \subseteq W$. Thus $\phi^3(w) = \phi(\phi^2(w)) = \phi(\text{an element of } W) =$ another element of W , so $\phi^3(W) \subseteq W$, and so on, so $\phi^k(w) \in W$ for all $k \geq 0$. Now consider a polynomial $f(x) = a_0 + \ldots + a_d x^d \in K[x]$. We then have $fw = \sum_i a_i \phi^i(w)$. The vectors $\phi^i(w)$ lie in W, the coefficients a_i lie in K, and W is a vector subspace of V, so we see that $\sum_i a_i \phi^i(w) \in W$. Thus $fw \in W$ for all $w \in W$ and $f \in K[x]$, so W is a submodule of V.

Example 5.6. Let A be the matrix $\begin{pmatrix} 0 & -6 \\ 1 & 5 \end{pmatrix}$ over $\mathbb{Q}$, and use this to make $\mathbb{Q}^2$ into a module over $\mathbb{Q}[x]$. Put $W_0 = \{(u, v) \in \mathbb{Q}^2 \mid u = -3v\}$ and $W_1 = \{(u, v) \in \mathbb{Q}^2 \mid u = -4v\}$. A typical

element of W_0 has the form $(-3v, v)$ and we have

$$\begin{pmatrix} 0 & -6 \\ 1 & 5 \end{pmatrix} \begin{pmatrix} -3v \\ v \end{pmatrix} = \begin{pmatrix} -6v \\ 2v \end{pmatrix},$$

which also lies in W_0. Thus W_0 is stable under A and thus is a submodule of $\mathbb{Q}^2$.

However, W_1 is not a submodule. Indeed, the vector $(-4, 1)$ lies in W_1 but

$$\begin{pmatrix} 0 & -6 \\ 1 & 5 \end{pmatrix} \begin{pmatrix} -4 \\ 1 \end{pmatrix} = \begin{pmatrix} -6 \\ 1 \end{pmatrix},$$

which does not lie in W_1.

Example 5.7. Suppose that $\lambda, \mu \in K$ and $\lambda \neq \mu$. Define $\phi\colon K^2 \to K^2$ by $\phi(u, v) = (\lambda u, \mu v)$, and use this to make K^2 into a module over $K[x]$. Define

$$L = \{(u, 0) \mid u \in K\} \subset K^2$$
$$M = \{(0, v) \mid v \in K\} \subset K^2.$$

I claim that L and M are $K[x]$-submodules of K^2, and moreover that the only submodules are $\{0\}$, L, M and K^2 itself.

It is clear that L and M are vector subspaces of K^2. Moreover we have $\phi(u, 0) = (\lambda u, 0) \in L$, so L is stable under ϕ and thus is a submodule. Similarly $\phi(0, v) = (0, \mu v) \in M$, so M is a submodule. It is trivial to check that $\{0\}$ and K^2 are subspaces of K^2.

Now let W be any submodule of K^2. Then W is also a vector subspace, with $0 \leq \dim(W) \leq \dim(K^2) = 2$. If $\dim(W) = 0$ then clearly $W = 0$, and if $\dim(W) = 2$ then clearly $W = K^2$. We can thus assume that $\dim(W) = 1$, so $W = K.(u, v)$ for some vector $(u, v) \neq (0, 0)$. Because W is a $K[x]$-submodule, we know that $(\lambda u, \mu v) \in W$, but $W = K.(u, v)$ so $(\lambda u, \mu v) = \nu.(u, v)$ for some $\nu \in K$, so $(\lambda - \nu)u = (\mu - \nu)v = 0$. If $u \neq 0$ then we deduce that $\nu = \lambda$ so $(\mu - \lambda)v = 0$, but $\mu \neq \lambda$ so $v = 0$. This means that $W = K.(u, 0) = L$. Similarly, if $v \neq 0$ we deduce that $W = M$.

Example 5.8. The set $\mathbb{R}[t]$ of polynomial functions is a vector subspace of the space $C^\infty(\mathbb{R}, \mathbb{R})$ of all smooth functions from $\mathbb{R}$ to $\mathbb{R}$. Moreover if $f \in \mathbb{R}[t]$ then the derivative of f is again a polynomial, in other words $\partial(f) = f' \in \mathbb{R}[t]$. This means that the subspace $\mathbb{R}[t]$ is stable under the endomorphism ∂, so it is an $\mathbb{R}[D]$-submodule of $C^\infty(\mathbb{R}, \mathbb{R})$.

Example 5.9. Let W be the space of functions of the form $f(t) = a\cos(t) + b\sin(t)$ (with $a, b \in \mathbb{R}$). Because $\partial(a\cos(t) + b\sin(t)) = b\cos(t) - a\sin(t)$, we see that W is stable under ∂. It is thus an $\mathbb{R}[D]$-submodule of $C^\infty(\mathbb{R}, \mathbb{R})$.

Remark 5.10. Suppose that N_0 and N_1 are two submodules of an R-module M. I claim that $N_0 \cap N_1$ is again a submodule. Indeed, as $0 \in N_0$ and $0 \in N_1$ we have $0 \in N_0 \cap N_1$. Suppose that $n, m \in N_0 \cap N_1$. As $n, m \in N_0$ and N_0 is a submodule we have $n + m \in N_0$. As $n, m \in N_1$ and N_1 is a submodule we have $n + m \in N_1$. Thus $n + m \in N_0 \cap N_1$. Now suppose that $a \in R$. As N_0 is a submodule and $n \in N_0$ we have $an \in N_0$. As N_1 is a submodule and $n \in N_1$ we also have $an \in N_1$, so $an \in N_0 \cap N_1$. This shows that $N_0 \cap N_1$ is a submodule, as claimed.

Definition 5.11. Suppose that N_0 and N_1 are two submodules of an R-module M. We define $N_0 + N_1$ to be the set of elements $x \in M$ that can be written in the form $x = n_0 + n_1$ for some $n_0 \in N_0$ and $n_1 \in N_1$. I claim that this is a submodule of M. Indeed, suppose that $x, y \in N_0 + N_1$, so we can write $x = n_0 + n_1$ and $y = m_0 + m_1$ with $n_0, m_0 \in N_0$ and $n_1, m_1 \in N_1$. Then $x + y$ can be written as $(n_0 + m_0) + (n_1 + m_1)$, with $n_0 + m_0 \in N_0$ and $n_1 + m_1 \in N_1$, so $x + y \in N_0 + N_1$. Similarly, if $a \in R$ then $an_0 \in N_0$ and $an_1 \in N_1$ so $ax = an_0 + an_1 \in N_0 + N_1$. This shows that $N_0 + N_1$ is closed under addition and under multiplication by R, so it is a submodule as claimed.

Definition 5.12. Let N_0 and N_1 be R-modules. We define $N_0 \oplus N_1$ to be the set of pairs (n_0, n_1) with $n_0 \in N_0$ and $n_1 \in N_1$. We make this set into an R-module by defining

$$(n_0, n_1) + (m_0, m_1) = (n_0 + m_0, n_1 + m_1)$$
$$a(n_0, n_1) = (an_0, an_1).$$

(It is a longish but straightforward exercise to check that the axioms are satisfied.) This R-module is called the *external direct sum* of N_0 and N_1.

Example 5.13. The group $\mathbb{Z}_2$ has elements $\overline{0}$ and $\overline{1}$, and the group $\mathbb{Z}_3$ has elements $\overline{0}$, $\overline{1}$ and $\overline{2}$. Thus, the group $\mathbb{Z}_2 \oplus \mathbb{Z}_3$ has elements $(\overline{0},\overline{0})$, $(\overline{0},\overline{1})$, $(\overline{0},\overline{2})$, $(\overline{1},\overline{0})$, $(\overline{1},\overline{1})$ and $(\overline{1},\overline{2})$. To illustrate the addition law, we have $(\overline{1},\overline{2}) + (\overline{1},\overline{2}) = (\overline{2},\overline{4})$. The first component is to be interpreted as an element of $\mathbb{Z}_2$, so $\overline{2} = \overline{0}$. The second component is to be interpreted as an element of $\mathbb{Z}_3$, so $\overline{4} = \overline{1}$. Thus $(\overline{1},\overline{2}) + (\overline{1},\overline{2}) = (\overline{0},\overline{1})$.

Example 5.14. An element of $R^n \oplus R^m$ is a pair (u, v) with $u \in R^n$ and $v \in R^m$, or in other words a list $(u_1, \dots, u_n, v_1, \dots, v_m)$ where each u_i and v_j is an element of R. Thus, $R^n \oplus R^m = R^{n+m}$.

The next example relies on the following definition:

Definition 5.15. Let A and B be matrices over a field K, of sizes $p \times q$ and $n \times m$. The *block sum* of A and B is the matrix $\left(\begin{array}{c|c} A & 0_{n\times q} \\ \hline 0_{p\times m} & B \end{array}\right)$, of size $(p+n) \times (q+m)$. This is denoted by $A \oplus B$. For example, if $A = \left(\begin{smallmatrix} 1 & 2 \\ 3 & 4 \end{smallmatrix}\right)$ and $B = \left(\begin{smallmatrix} 5 & 6 \\ 7 & 8 \end{smallmatrix}\right)$ then the block sum of A and B is

$$A \oplus B = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \oplus \begin{pmatrix} 5 & 6 \\ 7 & 8 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 0 & 0 \\ 3 & 4 & 0 & 0 \\ 0 & 0 & 5 & 6 \\ 0 & 0 & 7 & 8 \end{pmatrix}.$$

Note that an element $w \in R^{p+n}$ can be written as $w = (u, v)$ with $u \in R^p$ and $v \in R^n$, and we have

$$(A \oplus B)w = \left(\begin{array}{c|c} A & 0 \\ \hline 0 & B \end{array}\right) \left(\frac{u}{v}\right) = \left(\frac{Au}{Bv}\right).$$

Example 5.16. Let A and B be square matrices over a field K, of sizes n and m say. We then have modules M_A and M_B over $K[x]$. The elements of $M_A \oplus M_B$ are pairs $w = (u, v)$ with $u \in K^n$ and $v \in K^m$, or equivalently they are elements of K^{n+m}. The module structure is given by the rule $x(u, v) = (xu, xv) = (Au, Bv)$, or in other words $xw = (A \oplus B)w$. Thus $M_A \oplus M_B = M_{A\oplus B}$.

Definition 5.17. Let M be an R-module, and let N_0 and N_1 be submodules. We say that M is the *internal direct sum* of N_0 and N_1 if $N_0 + N_1 = M$ and $N_0 \cap N_1 = \{0\}$.

Remark 5.18. We can define a function $\sigma\colon N_0 \oplus N_1 \to M$ by $\sigma(n_0, n_1) = n_0 + n_1$. When we have defined homomorphisms and isomorphisms of modules, we will see that σ is always a homomorphism, and that σ is an isomorphism if and only if M is the internal direct sum of N_0 and N_1. This is the precise sense in which internal direct sums are "the same" as external ones.

Example 5.19. In example 5.7 we see that K^2 is the internal direct sum of L and M.

Example 5.20. Consider the Abelian group $M = \mathbb{Z}_{12}$ as a module over $\mathbb{Z}$. Put $N_0 = \{\overline{0}, \overline{3}, \overline{6}, \overline{9}\}$ and $N_1 = \{\overline{0}, \overline{4}, \overline{8}\}$. It is easy to see that N_0 and N_1 are subgroups, and obviously $N_0 \cap N_1 = \{\overline{0}\}$. I claim that we also have $N_0 + N_1 = M$. Indeed, we have $\overline{1} = \overline{9} + \overline{4} \in N_0 + N_1$ and $N_0 + N_1$ is a submodule so for any $a \in \mathbb{Z}$ we have $\overline{a} = a.\overline{1} \in N_0 + N_1$, as required. Thus M is the internal direct sum of N_0 and N_1.

Example 5.21. Let V be the space of functions $f \in C^\infty(\mathbb{R}, \mathbb{R})$ that satisfy $f'' = f$. This is a vector space closed under differentiation, so it is an $\mathbb{R}[D]$-submodule of $C^\infty(\mathbb{R}, \mathbb{R})$. Put $W_0 = \{f \mid f' = f\}$ and $W_1 = \{f \mid f' = -f\}$. These are also vector spaces closed under differentiation, so they are $\mathbb{R}[D]$-submodules of $C^\infty(\mathbb{R}, \mathbb{R})$. If $f \in W_1$ then $f'' = (-f)' = -(-f) = f$, so $f \in V$. This shows that $W_1 \subseteq V$ and similarly $W_0 \subseteq V$, so W_0 and W_1 are submodules of V.

I claim that V is the direct sum of W_0 and W_1. One way to see this is just to solve the differential equations. We find that V consists of all functions of the form $ae^t + be^{-t}$, that W_0 consists of all functions of the form ae^t, and that W_1 consists of all functions of the form be^{-t}, and the claim is clear from this.

We can also prove the claim without solving the differential equations explicitly. Indeed, if $f \in W_0 \cap W_1$ then $f = f'$ (because $f \in W_0$) and $f' = -f$ (because $f \in W_1$) so $f = -f$, so $f = 0$. This shows that $W_0 \cap W_1 = \{0\}$. Next, suppose that $g \in V$, so $g'' = g$. Put $g_0 = (g + g')/2$ and $g_1 = (g - g')/2$. We find that $g_0' = (g' + g'')/2 = (g' + g)/2 = g_0$, so $g_0 \in W_0$. Similarly, we have $g_1' = (g' - g'')/2 = (g' - g)/2 = -g_1$, so $g_1 \in W_1$. As $g = g_0 + g_1$ it follows that $g \in W_0 + W_1$, and we conclude that $V = W_0 + W_1$ as required.

Definition 5.22. Let M be a module over a ring R, and let $m_1, \dots, m_r$ be elements of M. Let N be the set of elements $x \in M$ that can be written in the form $x = a_1 m_1 + \dots + a_r m_r$ for some $a_1, \dots, a_r \in R$. I claim that this is a submodule of M. Indeed, if $x, y \in N$ then we have $x = \sum_i a_i m_i$ and $y = \sum_i b_i m_i$ for some $a_1, \dots, a_r, b_1, \dots, b_r \in R$. We then have $x + y = \sum_i (a_i + b_i) m_i$ so $x + y \in N$; this shows that N is closed under addition. Similarly, if $c \in R$ we have $cx = \sum_i (ca_i) m_i \in N$, so N is closed under multiplication by R, so it is a submodule as claimed.

We call N *the submodule generated by* $\{m_1, \dots, m_r\}$. In particular, we say that M *is generated by* $\{m_1, \dots, m_r\}$ if $N = M$, or equivalently if every element $x \in M$ can be written in the form $a_1 m_1 + \dots + a_r m_r$. We say that M is *finitely generated* if there is some finite list of elements that generates M.

Definition 5.23. We say that an R-module M is *cyclic* if there is a single element $m \in M$ that generates M, which means that every element $x \in M$ can be written in the form $x = am$ for some $a \in R$.

Example 5.24. The module R^d is clearly generated by the standard basis elements $e_1 = (1, 0, \dots, 0)$, $e_2 = (0, 1, 0, \dots, 0)$ and so on. In particular it is finitely generated. It is not cyclic unless $d = 1$.

Example 5.25. Let M be a finite Abelian group, considered as a $\mathbb{Z}$-module. Let the elements of M be $m_1, \dots, m_d$. Then any element $m \in M$ is equal to m_i for some i, so certainly it can be expressed in the form $\sum_i a_i m_i$ (for example, $m_2 = 0.m_1 + 1.m_2 + 0.m_3 + \dots + 0.m_d$). Thus, M is finitely generated as a $\mathbb{Z}$-module.

Example 5.26. Let W_2 be the space of functions of the form $f(t) = a + bt + ct^2$, considered as a module over $\mathbb{R}[D]$ in the usual way. In particular, the function $g(t) = t^2$ gives an element of W_2. I claim that W_2 is generated by g, and thus is cyclic. Indeed, we have $g'(t)/2 = t$ and $g''(t)/2 = 1$. It follows that for any function $f(t) = a + bt + ct^2$, we have $(c + (b/2)D + (a/2)D^2)g = cg + (b/2)g' + (a/2)g'' = f$, so $f \in \mathbb{R}[D]g$. This proves that $\mathbb{R}[D]g = W_2$ as required.

It is not hard to extend this method to show that the space W_d of polynomials of degree at most d is also a cyclic module over $\mathbb{R}[D]$ generated by the function $g(t) = t^d$.

Example 5.27. Consider $\mathbb{R}[x]$ as a module over $\mathbb{R}$; I claim it is not finitely generated. Indeed, suppose we have a finite list $f_1, \dots, f_n$ of elements of $\mathbb{R}[x]$. Let d_i be the degree of the polynomial f_i, and put $d = \max(d_1, \dots, d_n)$. Then each of the polynomials f_i only involves the powers $1, x, x^2, \dots, x^d$, so any polynomial of the form $a_1 f_1 + \dots + a_n f_n$ (with $a_1, \dots, a_n \in \mathbb{R}$) also involves only these powers. This means that x^{d+1} cannot be written in the form $a_1 f_1 + \dots + a_n f_n$, so the elements $f_1, \dots, f_n$ do not generate $\mathbb{R}[x]$ as a module over $\mathbb{R}$.

6. Homomorphisms

Definition 6.1. Let M and N be modules over a ring R. An *R-module homomorphism* (or just *homomorphism*) from M to N is a function $\alpha\colon M \to N$ such that

(a) $\alpha(m_0 + m_1) = \alpha(m_0) + \alpha(m_1)$ for all $m_0, m_1 \in M$.
(b) $\alpha(am) = a\alpha(m)$ for all $a \in R$ and $m \in M$.

Note that this implies that $\alpha(0) = \alpha(0.0) = 0\alpha(0) = 0$ and $\alpha(-m) = \alpha((-1).m) = (-1).\alpha(m) = -\alpha(m)$.

An *isomorphism* is a homomorphism which is also a bijection.

Remark 6.2. Let $\alpha\colon M \to N$ be an isomorphism. As α is a bijection, there is an inverse function $\alpha^{-1}\colon N \to M$ such that $\alpha(\alpha^{-1}(n)) = n$ for all $n \in N$ and $\alpha^{-1}(\alpha(m)) = m$ for all

$m \in M$. I claim that α^{-1} is also a homomorphism. To see this, suppose that $n_0, n_1 \in N$. We then have elements $\alpha^{-1}(n_0)$ and $\alpha^{-1}(n_1)$ in M. As α is a homomorphism, we have $\alpha(\alpha^{-1}(n_0) + \alpha^{-1}(n_1)) = \alpha(\alpha^{-1}(n_0)) + \alpha(\alpha^{-1}(n_1)) = n_0 + n_1$. We can apply α^{-1} to this equation to get $\alpha^{-1}(\alpha(\alpha^{-1}(n_0)) + \alpha(\alpha^{-1}(n_1))) = \alpha^{-1}(n_0 + n_1)$. Because $\alpha^{-1}(\alpha(m)) = m$ for all m, the left hand side is just $\alpha^{-1}(n_0) + \alpha^{-1}(n_1)$. We thus have $\alpha^{-1}(n_0) + \alpha^{-1}(n_1) = \alpha^{-1}(n_0 + n_1)$, showing that α^{-1} respects addition.

Similarly, suppose that $n \in N$ and $a \in R$. As α respects multiplication by R, we have $\alpha(a\alpha^{-1}(n)) = a\alpha(\alpha^{-1}(n)) = an$. By applying α^{-1} to this equation we get $\alpha^{-1}(\alpha(a\alpha^{-1}(n))) = \alpha^{-1}(an)$. The left hand side is just $a\alpha^{-1}(n)$, so we have $a\alpha^{-1}(n) = \alpha^{-1}(an)$, completing the proof that α^{-1} is a homomorphism.

Example 6.3. Let R be any ring. Define $\tau\colon R^2 \to R^2$, $\sigma\colon R^3 \to R$ and $\delta\colon R^2 \to R^3$ by

$$\begin{aligned}\tau(u,v) &= (v,u)\\ \sigma(x,y,z) &= x+y+z\\ \delta(u,v) &= (u, v-u, -v).\end{aligned}$$

It is easy to check that these are all homomorphisms. For example, we have

$$\begin{aligned}\delta(u_0,v_0) + \delta(u_1,v_1) &= (u_0, v_0-u_0, -v_0) + (u_1, v_1-u_1, -v_1)\\ &= (u_0+u_1, v_0+v_1-u_0-u_1, -v_0-v_1)\\ &= \delta(u_0+u_1, v_0+v_1)\\ &= \delta((u_0,v_0)+(u_1,v_1))\end{aligned}$$

and

$$\begin{aligned}a\delta(u,v) &= a.(u, v-u, -v)\\ &= (au, av-au, -av)\\ &= \delta(au, av)\\ &= \delta(a.(u,v)),\end{aligned}$$

so δ is a homomorphism.

Example 6.4. I would like to define two homomorphisms $\alpha, \beta\colon \mathbb{Z}_3 \to \mathbb{Z}_{12}$ by $\alpha(\overline{m}) = \overline{4m}$ and $\beta(\overline{m}) = \overline{5m}$. There is a potential problem with this kind of definition, which means that the definition of β is actually invalid, although it turns out that α is OK. Consider the element $x = \overline{1} \in \mathbb{Z}_3$, which can also be described as $x = \overline{4}$. Using the description $x = \overline{1}$ we get $\beta(x) = \overline{5} \in \mathbb{Z}_{12}$. Using the description $x = \overline{4}$ we get $\beta(x) = \overline{20} \in \mathbb{Z}_{12}$. As $20 \neq 5 \pmod{12}$, the elements $\overline{5}$ and $\overline{20}$ in $\mathbb{Z}_{12}$ are not the same, so our definition of β is not self-consistent.

However, this problem does not occur with α. To see why, suppose we describe an element $y \in \mathbb{Z}_3$ in two different ways, say $y = \overline{n} = \overline{m}$. As $\overline{n} = \overline{m}$ in $\mathbb{Z}_3$, we have $n = m \pmod 3$, so $n = m + 3k$ for some integer k. This means that $4n = 4m + 12k$, so $\overline{4n} = \overline{4m}$ in $\mathbb{Z}_{12}$. This means that we get the same answer for $\alpha(y)$ no matter which description we use, so α is a well-defined function from $\mathbb{Z}_3$ to $\mathbb{Z}_{12}$.

We also have $\alpha(\overline{n}) + \alpha(\overline{m}) = \overline{4n} + \overline{4m} = \overline{4(n+m)} = \alpha(\overline{n} + \overline{m})$, so α is a homomorphism of groups. It follows easily that it is also a homomorphism of $\mathbb{Z}$-modules.

The above example generalizes as follows:

Proposition 6.5. *Let p, q and r be integers such that $p, q > 0$ and pr is divisible by q. Then there is a unique homomorphism $\alpha\colon \mathbb{Z}_p \to \mathbb{Z}_q$ such that $\alpha(\overline{m}) = \overline{rm}$ for all $m \in \mathbb{Z}$.*

Proof. By assumption we have $pr = qs$ for some integer s. If $\overline{n} = \overline{m}$ in $\mathbb{Z}_p$ then $m = n + pk$ for some k, so $rm = rn + rpk = rn + qsk$, so $rm = rn \pmod q$, so $\overline{rm} = \overline{rn}$ in $\mathbb{Z}_q$. This shows that α is well-defined. We also have $\alpha(\overline{n}) + \alpha(\overline{m}) = \overline{rn} + \overline{rm} = \overline{r(n+m)} = \alpha(\overline{n} + \overline{m})$, so α is a homomorphism. □

Example 6.6. Because 6×5 is divisible by 15, there is a unique homomorphism $\alpha\colon \mathbb{Z}_6 \to \mathbb{Z}_{15}$ such that $\alpha(\overline{m}) = \overline{5m}$ for all m.

Example 6.7. Let N be a module over a ring R, and let $n_1, \dots, n_d$ be a list of elements of N. We define a function $\alpha\colon R^d \to N$ by

$$\alpha(u_1, \dots, u_d) = u_1 n_1 + \dots + u_d n_d.$$

I claim that this is a homomorphism. Indeed, we have

$$\begin{aligned}\alpha((u_1, \dots, u_d) + (v_1, \dots, v_d)) &= \alpha(u_1 + v_1, \dots, u_d + v_d)\\ &= (u_1 + v_1)n_1 + \dots + (u_d + v_d)n_d\\ &= (u_1 n_1 + \dots + u_d n_d) + (v_1 n_1 + \dots + v_d n_d)\\ &= \alpha(u_1, \dots, u_d) + \alpha(v_1, \dots, v_d),\end{aligned}$$

and

$$\begin{aligned}\alpha(a(u_1, \dots, u_d)) &= \alpha(au_1, \dots, au_d)\\ &= au_1 n_1 + \dots + au_d n_d\\ &= a(u_1 n_1 + \dots + u_d n_d)\\ &= a\alpha(u_1, \dots, u_d)\end{aligned}$$

as required.

Next, I claim that every homomorphism $\beta\colon R^d \to N$ is of the form just described. To see this, let e_k be the element of R^d given by $e_k = (0, \dots, 0, 1, 0, \dots, 0)$, with the 1 in the k'th place. Put $n_k = \beta(e_k) \in N$. Any element $u = (u_1, \dots, u_d)$ in R^d can be written as $u = u_1 e_1 + \dots + u_d e_d$. For example, in the case $d = 3$ we have

$$(u_1, u_2, u_3) = u_1(1, 0, 0) + u_2(0, 1, 0) + u_3(0, 0, 1) = u_1 e_1 + u_2 e_2 + u_3 e_3.$$

As β is a homomorphism, we have

$$\begin{aligned}\beta(u_1, \dots, u_d) &= \beta(u_1 e_1 + \dots + u_d e_d)\\ &= u_1\beta(e_1) + \dots + u_d\beta(e_d)\\ &= u_1 n_1 + \dots + u_d n_d,\end{aligned}$$

so β has the form described previously.

We summarize the above discussion as follows:

Proposition 6.8. *For any list $(n_1, \dots, n_d)$ of elements of N we have a homomorphism $\alpha\colon R^d \to N$ given by $\alpha(u) = \sum_i u_i n_i$. Conversely, every homomorphism $\alpha\colon R^d \to N$ arises in this way from a unique list $(n_1, \dots, n_d)$.* □

Example 6.9. We now consider homomorphisms $\alpha\colon R^d \to R^e$. By the previous example, α corresponds to a list $u_1, \dots, u_d$, where each u_i is an element of R^e, or in other words a vector of length e. We can construct a matrix A (with d columns and e rows) whose columns are the vectors $u_1, \dots, u_d$, and we find that $\alpha(x) = Ax$ for all $x \in R^d$.

Consider for example the homomorphism $\delta\colon R^2 \to R^3$ given by $\delta(s, t) = (s, t - s, -t)$. We have

$$\begin{aligned}u_1 &= \delta(e_1) = \delta(1, 0) = (1, -1, 0)\\ u_2 &= \delta(e_2) = \delta(0, 1) = (0, 1, -1)\\ A &= \begin{pmatrix} 1 & 0\\ -1 & 1\\ 0 & -1\end{pmatrix}\end{aligned}$$

so

$$A\begin{pmatrix} s\\ t\end{pmatrix} = \begin{pmatrix} 1 & 0\\ -1 & 1\\ 0 & -1\end{pmatrix}\begin{pmatrix} s\\ t\end{pmatrix} = \begin{pmatrix} s\\ t - s\\ -t\end{pmatrix} = \alpha(s, t)$$

as claimed.

We summarize the above discussion as follows:

Proposition 6.10. *Homomorphisms from R^d to R^e are essentially the same as $d \times e$ matrices over R.* □

We next consider homomorphisms of modules over polynomial rings.

Proposition 6.11. *Let V and W be vector spaces over a field K, and let $\phi\colon V \to V$ and $\psi\colon W \to W$ be K-linear maps. We use these to make V and W into modules over $K[x]$ in the usual way, so that $xv = \phi(v)$ for $v \in V$ and $xw = \psi(w)$ for $w \in W$. Then the $K[x]$-module homomorphisms from V to W are precisely the K-linear maps $\gamma\colon V \to W$ such that $\psi\gamma = \gamma\phi$, or in other words $\psi(\gamma(v)) = \gamma(\phi(v))$ for all $v \in V$.*

Proof. Let $\gamma\colon V \to W$ be a $K[x]$-module homomorphism, so $\gamma(v_0 + v_1) = \gamma(v_0) + \gamma(v_1)$ for all $v_0, v_1 \in V$, and $\gamma(av) = a\gamma(v)$ for all $a \in K[x]$ and $v \in V$. By taking a to be a constant polynomial, we see that $\gamma(av) = a\gamma(v)$ for all $a \in K$, so γ is a K-linear map. Now take $a = x$ instead. As $v \in V$ we have $xv = \phi(v)$, and as $\gamma(v) \in W$ we have $x\gamma(v) = \psi(\gamma(v))$. Thus, the equation $\gamma(xv) = x\gamma(v)$ becomes $\gamma(\phi(v)) = \psi(\gamma(v))$, as required.

Conversely, suppose we have a K-linear map $\gamma\colon V \to W$ satisfying $\gamma\phi = \psi\gamma$. It follows that

$$\gamma\phi^2 = (\gamma\phi)\phi = (\psi\gamma)\phi = \psi(\gamma\phi) = \psi^2\gamma.$$

This can be extended by induction to show that $\gamma\phi^k = \psi^k\gamma$ for all $k \geq 0$. Thus, for any polynomial $p(x) = \sum_i a_i x^i$ we have

$$\gamma p(\phi) = \sum_i a_i \gamma\phi^i = \sum_i a_i \psi^i \gamma = p(\psi)\gamma,$$

so $\gamma(p(x)v) = p(x)\gamma(v)$, so γ is a $K[x]$-module homomorphism. □

Remark 6.12. We can indicate where the maps γ, ϕ and ψ go by a diagram as follows:

$$\begin{array}{ccc} V & \xrightarrow{\gamma} & W \\ {\scriptstyle\phi}\downarrow & & \downarrow{\scriptstyle\psi} \\ V & \xrightarrow[\gamma]{} & W. \end{array}$$

The condition $\psi\gamma = \gamma\phi$ means that the two way round the square are the same. This is usually expressed by saying that the diagram *commutes*.

Remark 6.13. The proposition can be restated as follows in terms of matrices: If A and B are square matrices of size n and m over a field K, then the $K[x]$-module homomorphisms from M_A to M_B are the $n \times m$ matrices C over K such that $CA = BC$. Here the matrices A, B and C correspond to the linear maps ϕ, ψ and γ respectively.

Remark 6.14. If matrices A and B are chosen randomly, then usually the only matrix C satisfying $CA = BC$ will be the zero matrix. Of course, most of the examples in these notes are chosen specially so that there are some nonzero solutions.

Example 6.15. Put $A = \left(\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right)$ and $B = \left(\begin{smallmatrix} 0 & 2 \\ 2 & 0 \end{smallmatrix}\right)$. The homomorphisms from M_A to M_B are then the matrices $C = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ for which $CA = BC$, or in other words

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

or equivalently

$$\begin{aligned} a + b &= 2c \\ a + b &= 2d \\ c + d &= 2a \\ c + d &= 2b. \end{aligned}$$

Solving these equations gives $a = b = c = d$. Thus, the homomorphisms from M_A to M_B are precisely the matrices of the form $\left(\begin{smallmatrix} a & a \\ a & a \end{smallmatrix}\right)$ for some $a \in K$.

Example 6.16. Define $\phi\colon \mathbb{R}^2 \to \mathbb{R}^2$ by $\phi(u,v) = (v,u)$, and define $\psi\colon \mathbb{R}^3 \to \mathbb{R}^3$ by $\psi(u,v,w) = (v,w,u)$. We use these to make $\mathbb{R}^2$ and $\mathbb{R}^3$ into modules over $\mathbb{R}[x]$ as usual. Define $\gamma\colon \mathbb{R}^2 \to \mathbb{R}^3$ by $\gamma(u,v) = (u+v, u+v, u+v)/2$. I claim that this is an $\mathbb{R}[x]$-module homomorphism. It is clearly $\mathbb{R}$-linear, so it is enough to check that $\gamma\phi = \psi\gamma$. We have $\gamma\phi(u,v) = \gamma(v,u) = (v+u, v+u, v+u)/2$. We also have $\gamma(u,v) = (u+v, u+v, u+v)/2$ and $\psi(w,w,w) = (w,w,w)$ for any w, so $\psi\gamma(u,v) = (u+v, u+v, u+v)/2$, which is the same as $\gamma\phi(u,v)$, as claimed.

Next, I claim that any other $\mathbb{R}[x]$-module homomorphism $\beta\colon \mathbb{R}^2 \to \mathbb{R}^3$ is actually a multiple of γ. To see this, note that $\beta(1,1) \in \mathbb{R}^3$, so $\beta(1,1) = (\lambda,\mu,\nu)$ for some $\lambda,\mu,\nu \in \mathbb{R}$. As β is a homomorphism we must have $\beta\phi(1,1) = \psi\beta(1,1) = \psi(\lambda,\mu,\nu) = (\mu,\nu,\lambda)$. On the other hand, we have $\phi(1,1) = (1,1)$ so $\beta\phi(1,1) = \beta(1,1) = (\lambda,\mu,\nu)$. Thus $(\lambda,\mu,\nu) = (\mu,\nu,\lambda)$, so $\nu = \mu = \lambda$. This means that $\beta(1,1) = (\lambda,\lambda,\lambda) = \lambda.(1,1,1)$.

We next claim that $\beta(1,-1) = (0,0,0)$. To see this, note that $\beta\phi^3 = \psi^3\beta$. Moreover $\psi^3(u,v,w) = (u,v,w)$ for all $(u,v,w) \in \mathbb{R}^3$, and $\phi^3(1,-1) = (-1,1) = -(1,-1)$, so the equation $\psi^3\beta(1,-1) = \beta\phi^3(1,-1)$ becomes $\beta(1,-1) = \beta(-1,1) = -\beta(1,-1)$. By rearranging we see that $\beta(1,-1) = 0$ as claimed.

Now consider an arbitrary element $(u,v) \in \mathbb{R}^2$. We can write this as

$$(u,v) = \frac{u+v}{2}(1,1) + \frac{u-v}{2}(1,-1),$$

so

$$\begin{aligned}\beta(u,v) &= \frac{u+v}{2}\beta(1,1) + \frac{u-v}{2}\beta(1,-1)\\ &= \frac{u+v}{2}(\lambda,\lambda,\lambda) + \frac{u-v}{2}(0,0,0)\\ &= \lambda\gamma(u,v).\end{aligned}$$

This proves that $\beta = \lambda\gamma$, as claimed.

Example 6.17. Define $\tau\colon C^\infty(\mathbb{R},\mathbb{R}) \to C^\infty(\mathbb{R},\mathbb{R})$ by $\tau(f)(t) = f(t+1)$. Thus if $g(t) = 3t$ and $h(t) = \sin(2\pi t)$ and $k(t) = 2^t$ then $\tau(g) = g + 3$ and $\tau(h) = h$ and $\tau(k) = 2k$. It is clear that τ is an $\mathbb{R}$-linear map. By the chain rule we have

$$\frac{d}{dt}f(t+1) = f'(t+1)\frac{d}{dt}(t+1) = f'(t+1),$$

so $\partial(\tau(f)) = \tau(\partial(f))$, so τ is a homomorphism of $\mathbb{R}[D]$-modules.

Proposition 6.18. *Let A and B be $n \times n$ matrices over a field K. Then M_A is isomorphic to M_B if and only if A is conjugate to B, in other words there is an invertible $n \times n$ matrix P such that $PAP^{-1} = B$.*

Proof. A homomorphism from M_A to M_B is an $n \times n$ matrix P such that $PA = BP$. Such a homomorphism is an isomorphism if and only if P is invertible, and if so, the condition $PA = BP$ is equivalent to the condition $PAP^{-1} = B$. □

Corollary 6.19. *Suppose that A is a diagonalizable $n \times n$ matrix over K. Then there exists an invertible matrix P such that $D := PAP^{-1}$ is a diagonal matrix, with diagonal entries $\lambda_1, \dots, \lambda_n$ say. Then M_A is isomorphic to M_D and thus to $M_{\lambda_1} \oplus \dots \oplus M_{\lambda_n}$.* □

Definition 6.20. Let $\alpha\colon M \to N$ be a homomorphism of modules over a ring R. We define the kernel and image of α by

$$\begin{aligned}\ker(\alpha) &= \{m \in M \mid \alpha(m) = 0\}\\ \text{image}(\alpha) &= \{n \in N \mid n = \alpha(m) \text{ for some } m \in M\}.\end{aligned}$$

Proposition 6.21. $\ker(\alpha)$ *is a submodule of M and* $\text{image}(\alpha)$ *is a submodule of N.*

Proof. Suppose that $m_0, m_1 \in \ker(\alpha)$ and that $a \in R$. We then have $\alpha(m_0) = \alpha(m_1) = 0$ and so

$$\alpha(m_0 + m_1) = \alpha(m_0) + \alpha(m_1) = 0 + 0 = 0$$
$$\alpha(am_0) = a\alpha(m_0) = a.0 = 0,$$

so $m_0 + m_1 \in \ker(\alpha)$ and $am_0 \in \ker(\alpha)$. This shows that $\ker(\alpha)$ is a submodule of M.

Now suppose that $n_0, n_1 \in \text{image}(\alpha)$ and $a \in R$. We then have $n_0 = \alpha(m_0)$ for some $m_0 \in M$ and $n_1 = \alpha(m_1)$ for some $m_1 \in M$. It follows that $\alpha(m_0 + m_1) = n_0 + n_1$, so $n_0 + n_1$ is α(something), so $n_0 + n_1 \in \text{image}(\alpha)$. Similarly, we have $an_0 = \alpha(am_0)$, so $an_0 \in \text{image}(\alpha)$. This shows that $\text{image}(\alpha)$ is a submodule of N. □

Example 6.22. Define $\alpha\colon \mathbb{Z} \to \mathbb{Z}_{12}$ by $\alpha(n) = \overline{4n}$. We then have

$$\begin{array}{lll} \alpha(0) = \overline{0} & \alpha(1) = \overline{4} & \alpha(2) = \overline{8} \\ \alpha(3) = \overline{12} = \overline{0} & \alpha(4) = \overline{16} = \overline{4} & \alpha(5) = \overline{20} = \overline{8} \\ \alpha(6) = \overline{24} = \overline{0} & \alpha(7) = \overline{28} = \overline{4} & \alpha(8) = \overline{32} = \overline{8} \end{array}$$

and everything repeats with period three. It follows that the only elements of $\mathbb{Z}_{12}$ that can be written in the form $\alpha(n)$ are $\overline{0}$, $\overline{4}$ and $\overline{8}$, so $\text{image}(\alpha) = \{\overline{0}, \overline{4}, \overline{8}\}$. It also follows that $\alpha(n) = \overline{0}$ if and only if n is divisible by 3, so $\ker(\alpha) = \{n \in \mathbb{Z} \mid n = 0 \pmod 3\} = 3\mathbb{Z}$.

Example 6.23. Define $\sigma\colon R^3 \to R$ by $\sigma(x, y, z) = x + y + z$ and $\delta\colon R^2 \to R^3$ by $\delta(u, v) = (u, v - u, -v)$. I claim that $\text{image}(\delta) = \ker(\sigma)$.

To see this, first suppose that $(x, y, z) \in \text{image}(\delta)$. This means that $(x, y, z) = \delta(u, v) = (u, v - u, -v)$ for some u, v. It follows that $\sigma(x, y, z) = x + y + z = u + (v - u) + (-v) = 0$, so $(x, y, z) \in \ker(\sigma)$. This proves that $\text{image}(\delta) \subseteq \ker(\sigma)$.

Conversely, suppose that $(x, y, z) \in \ker(\sigma)$. This means that $x+y+z = 0$, so $-(x+y) = z$. From this it follows that $(x, y, z) = (x, (x + y) - x, -(x + y)) = \delta(x, x + y)$, so (x, y, z) is δ(something), so $(x, y, z) \in \text{image}(\delta)$. This proves that $\ker(\sigma) \subseteq \text{image}(\delta)$ and thus that $\ker(\sigma) = \text{image}(\delta)$, as claimed.

Remark 6.24. Suppose we have modules L, M, N and homomorphisms $\alpha\colon L \to M$ and $\beta\colon M \to N$ such that $\text{image}(\alpha) = \ker(\beta)$. We then say that the sequence $L \xrightarrow{\alpha} M \xrightarrow{\beta} N$ is *exact*; for example, the sequence $R^2 \xrightarrow{\delta} R^3 \xrightarrow{\sigma} R$ in the above example is exact. This is a very important concept elsewhere in the theory of modules, although we will make little use of it in this course.

Proposition 6.25. *Let $\alpha\colon M \to N$ be a homomorphism of R-modules. Then*

(a) *α is injective if and only if* $\ker(\alpha) = \{0\}$.
(b) *α is surjective if and only if* $\text{image}(\alpha) = N$.
(c) *α is an isomorphism if and only if* $\ker(\alpha) = \{0\}$ *and* $\text{image}(\alpha) = N$.

Proof. First, suppose that $\ker(\alpha) = \{0\}$. If $\alpha(m_0) = \alpha(m_1)$ then $\alpha(m_0 - m_1) = \alpha(m_0) - \alpha(m_1) = 0$, so $m_0 - m_1 \in \ker(\alpha) = \{0\}$, so $m_0 - m_1 = 0$, so $m_0 = m_1$. This proves that α is injective.

Conversely, suppose that α is injective. If $m \in \ker(\alpha)$ then $\alpha(m) = 0$, so $\alpha(m) = \alpha(0)$, and as α is injective this means that $m = 0$. Thus $\ker(\alpha) = \{0\}$. This completes the proof of (a).

For (b), note that $\text{image}(\alpha)$ is the set of things in N that can be written in the form $\alpha(m)$ for some m. Thus $\text{image}(\alpha) = M$ if and only if *every* element in N can be written as $\alpha(m)$ for some m, and this is precisely the definition of surjectivity.

Finally, an isomorphism is just a bijective homomorphism. It is standard that a function is a bijection if and only if it is both injective and surjective, so (c) follows immediately from (a) and (b). □

7. Factor modules

Let M be an R-module for some ring R, and let N be a submodule. We next define the factor module M/N.

For any element $m \in M$ we define the set $m + N = \{m + n \mid n \in N\}$, which is a subset of M. A *coset* of N in M is a subset $C \subseteq M$ that can be written in the form $C = m + N$ for some m. We write M/N for the set of all such cosets.

Example 7.1. Take $R = \mathbb{Z}$ and $M = \mathbb{Z}$ and

$$\begin{aligned} N = 3\mathbb{Z} &= \{n \in \mathbb{Z} \mid n = 0 \pmod 3\} \\ &= \{\dots, -9, -6, -3, 0, 3, 6, 9, \dots\}. \end{aligned}$$

Consider the following three sets:

$$\begin{aligned} A &= \{\dots, -10, -7, -4, -1, 2, 5, 8, \dots\} \\ B &= \{\dots, -9, -6, -3, 0, 3, 6, 9, \dots\} \\ C &= \{\dots, -8, -5, -2, 1, 4, 7, 10, \dots\}. \end{aligned}$$

The set A can be described as $-7 + N$ or as $-1 + N$ or as $26 + N$, so it is a coset. Similarly, B can be described as $0 + N$ or $999 + N$ and C can be described as $-8 + N$ or $1 + N$, so A, B and C are all cosets. In fact, they are the only cosets, so $\mathbb{Z}/N = \{A, B, C\}$.

Proposition 7.2. *Let R, M and N be as above, and suppose that $m_0, m_1 \in M$. Then the following are equivalent*

(a) $m_0 + N = m_1 + N$
(b) $m_0 - m_1 \in N$
(c) $m_0 \in m_1 + N$
(d) $m_1 \in m_0 + N$
(e) $(m_0 + N) \cap (m_1 + N)$ *is nonempty.*

Proof. If $m_0 \in m_1 + N$ then $m_0 = m_1 + n$ for some $n \in N$ so $m_0 - m_1 = n \in N$. Conversely, if $m_0 - m_1 \in N$ then the equation $m_0 = m_1 + (m_0 - m_1)$ shows that $m_0 \in m_1 + N$, so statements (b) and (c) are equivalent. Similarly, (b) and (d) are equivalent.

If (c) holds then m_0 lies in both $m_0 + N$ and $m_1 + N$, so $(m_0 + N) \cap (m_1 + N) \neq \emptyset$, so (e) holds.

Conversely suppose that (e) holds, so there is an element x lying in both $m_0 + N$ and $m_1 + N$. This means that $x = m_0 + n_0$ for some $n_0 \in N$ and $x = m_1 + n_1$ for some $n_1 \in N$, so $m_0 = x - n_0 = m_1 + (n_1 - n_0)$. This shows that $m_0 \in m_1 + N$, so (c) holds. We now see that (b), (c), (d) and (e) are all equivalent to each other.

If (a) holds then it is clear that (c) holds and thus that (b),(d) and (e) also hold. Conversely, suppose that (b) holds. If $x \in m_0 + N$ then $x = m_0 + n$ for some $n \in N$, so $x = m_1 + ((m_0 - m_1) + n)$ and $(m_0 - m_1) + n \in N$ so $x \in m_1 + N$. This shows that $m_0 + N \subseteq m_1 + N$, and a similar argument shows that $m_1 + N \subseteq m_0 + N$, so (a) holds. This now shows that all five statements are equivalent to each other. □

We next want to define addition of cosets. Given two cosets C_0 and C_1 we choose $m_0, m_1 \in M$ such that $C_0 = m_0 + N$ and $C_1 = m_1 + N$, and then we would like to define $C_0 + C_1 = (m_0 + m_1) + N$. Similarly, if $a \in R$ we would like to define $aC_0 = (am_0) + N$. There is a potential problem here: suppose we chose a different description of the same coset C_0 (say as $m_0' + N$) and a different description of C_1 (say as $m_1' + N$). This gives an apparently different answer for $C_0 + C_1$: before we had $(m_0 + m_1) + N$, now we have $(m_0' + m_1') + N$. If these were genuinely different cosets then our definition of addition would be ambiguous and thus invalid. However, we will show that these are simply different descriptions of the same coset, so our definition is unambiguous after all.

Lemma 7.3. *If $m_0 + N = m_0' + N$ and $m_1 + N = m_1' + N$ then $(m_0 + m_1) + N = (m_0' + m_1') + N$ and $(am_0) + N = (am_0') + N$.*

Proof. By Proposition 7.2 we have $m_0 - m_0' \in N$ and $m_1 - m_1' \in N$. As N is closed under addition this means that $m_0 - m_0' + m_1 - m_1' \in N$, or in other words $(m_0 + m_1) - (m_0' + m_1') \in N$, so $(m_0 + m_1) + N = (m_0' + m_1') + N$ as claimed. Similarly, as $m_0 - m_0' \in N$ and N is a submodule, we have $am_0 - am_0' = a(m_0 - m_0') \in N$, so $(am_0) + N = (am_0') + N$ as claimed. □

Corollary 7.4. *We can unambiguously define addition of cosets and multiplication of cosets by elements of R, using the formulae*

$$\begin{aligned} (m_0 + N) + (m_1 + N) &= (m_0 + m_1) + N \\ a(m + N) &= (am) + N. \end{aligned}$$

Notation 7.5. If there is no ambiguity about which submodule N is intended, we will write $\overline{m}$ for $m + N$.

Proposition 7.6. *The definitions in Corollary 7.4 make the set M/N into an R-module. Moreover, the function $\pi\colon M \to M/N$ defined by $\pi(m) = \overline{m} = m + N$ is an R-module homomorphism.*

Proof. We need to check all the axioms in Definition 3.1. All the proofs follow the same pattern, so we will do only two of them. We first consider axiom (d), which says that addition is associative. Let A, B, C be any three cosets; we must show that $A + (B + C) = (A + B) + C$. We can choose $a, b, c \in M$ such that $A = a + N$, $B = b + N$ and $C = c + N$. From the definition of addition in M/N we have $B + C = (b + c) + N$ and thus $A + (B + C) = (a + (b + c)) + N$. Similarly, we have $(A + B) + C = ((a + b) + c) + N$. As addition is associative in the module M that we started with, we have $(a+b)+c = a+(b+c)$ so $((a+b)+c)+N = (a+(b+c))+N$, so $(A+B)+C = A+(B+C)$ as required.

We next check axiom (j) (which says that multiplication is right-distributive). Let A and B be elements of M/N, and let r be an element of R; we must show that $r(A + B) = rA + rB$. Choose a and b such that $A = a + N$ and $B = b + N$. We then have $A + B = (a + b) + N$ so $r(A + B) = r((a + b) + N) = (r(a + b)) + N = (ra + rb) + N = (ra + N) + (rb + N) = r(a + N) + r(b + N) = rA + rB$, as required.

The very definition of addition and multiplication in M/N says that $\pi(m_0) + \pi(m_1) = \pi(m_0 + m_1)$ and $a\pi(m) = \pi(am)$, so π is an R-module homomorphism. □

Theorem 7.7 (The first isomorphism theorem)**.** *Let $\alpha\colon M \to N$ be a homomorphism of R-modules. Put $K = \ker(\alpha) \subseteq M$ and $L = \text{image}(\alpha) \subseteq N$. Then there is an isomorphism $\overline{\alpha}\colon M/K \to L$ such that $\overline{\alpha}(m + K) = \alpha(m)$ for all $m \in M$.*

Proof. Let C be a coset in M/K. We can choose an element $m \in M$ such that $C = m + K$, and clearly $\alpha(m) \in \text{image}(\alpha) = L$. We would like to define $\overline{\alpha}(C) = \alpha(m)$, but we need to check that this is well-defined. Suppose we describe the same coset C in a different way, say as $C = m' + K$. This gives an apparently different answer for $\overline{\alpha}(C)$: before we had $\alpha(m)$, now we have $\alpha(m')$. As $m' + K = m + K$ we have $m' - m \in K = \ker(\alpha)$, which means that $\alpha(m') - \alpha(m) = \alpha(m' - m) = 0$, so $\alpha(m') = \alpha(m)$, so our two answers are actually the same. Thus, we have a well-defined function $\overline{\alpha}\colon M/K \to L$ satisfying $\overline{\alpha}(m + K) = \alpha(m)$ for all m.

Next, we have

$$\begin{aligned}\overline{\alpha}((m_0 + K) + (m_1 + K)) &= \overline{\alpha}((m_0 + m_1) + K)\\ &= \alpha(m_0 + m_1)\\ &= \alpha(m_0) + \alpha(m_1)\\ &= \overline{\alpha}(m_0 + K) + \overline{\alpha}(m_1 + K)\end{aligned}$$

and

$$\begin{aligned}\overline{\alpha}(a(m + K)) &= \overline{\alpha}((am) + K)\\ &= \alpha(am)\\ &= a\alpha(m)\\ &= a\overline{\alpha}(m + K),\end{aligned}$$

so $\overline{\alpha}$ is a homomorphism.

We next show that $\overline{\alpha}\colon M/K \to L$ is surjective. As L was defined as the image of α, any element $n \in L$ has the form $n = \alpha(m)$ for some $m \in M$. This means that $n = \overline{\alpha}(m + K)$, so n is in the image of $\overline{\alpha}$. As this is true for every element of L, the homomorphism $\overline{\alpha}$ is surjective.

Finally, we show that $\overline{\alpha}$ is injective. Suppose we have cosets C_0, C_1 with $\overline{\alpha}(C_0) = \overline{\alpha}(C_1)$. Choose $m_0, m_1 \in M$ such that $C_0 = m_0 + K$ and $C_1 = m_1 + K$. Then the equation $\overline{\alpha}(C_0) = \overline{\alpha}(C_1)$ means that $\alpha(m_0) = \alpha(m_1)$, so $\alpha(m_0 - m_1) = 0$, so $m_0 - m_1 \in \ker(\alpha) = K$. As $m_0 - m_1 \in K$ we have $m_0 + K = m_1 + K$ or in other words $C_0 = C_1$. This proves that $\overline{\alpha}$ is injective as well as surjective, so it is an isomorphism. □

Example 7.8. Define $\alpha\colon \mathbb{Z}^2 \to \mathbb{Z}^2$ by $\alpha(u,v) = (u+v, u+v)$, and put $K = \ker(\alpha)$ and $L = \text{image}(\alpha)$. Clearly $\alpha(u,v) = (0,0)$ if and only if $v = -u$, so

$$K = \{(u,v) \in \mathbb{Z}^2 \mid v = -u\} = \{(t,-t) \mid t \in \mathbb{Z}\}.$$

Next, note that that $\alpha(u,v)$ is always of the form (r,r) for some r. Conversely, any vector of the form (r,r) can be written as $\alpha(r,0)$, so it lies in the image of α. Thus

$$L = \{(x,y) \in \mathbb{Z}^2 \mid x = y\} = \{(r,r) \mid r \in \mathbb{Z}\}.$$

Consider the set

$$C = \{\dots, (-3,5), (-2,4), (-1,3), (0,2), (1,1), (2,0), (3,-1), \dots\}.$$

This can be described as $C = (1,1) + K$ or $C = (-3,5) + K$ or $C = (-999, 1001) + K$, so it is a coset of K, or in other words an element of the group $\mathbb{Z}^2/K$. We have

$$\overline{\alpha}(C) = \alpha(1,1) = \alpha(-3,5) = \alpha(-999,1001) = (2,2).$$

Proposition 7.9. *Let M be a cyclic R-module. Then $M \simeq R/I$ for some submodule I of R.*

Proof. Choose an element m that generates M. Define a homomorphism $\alpha\colon R \to M$ by $\alpha(a) = am$. As m generates M, this homomorphism is surjective. Put $I = \ker(\alpha)$, which is a submodule of R. The First Isomorphism Theorem now tells us that $R/I \simeq M$. □

8. Ideals and factor rings

Definition 8.1. An *ideal* in a ring R is a subset $I \subseteq R$ such that

(a) $0 \in I$
(b) If $b, c \in I$ then $b + c \in I$
(c) If $a \in R$ and $b \in I$ then $ab \in I$.

Remark 8.2. We remarked earlier that R can be regarded as a module over itself. By comparing the above definition with Definition 5.1 we see that ideals are just the same as submodules of R.

Example 8.3. Let I be the set of even integers; then I is an ideal in $\mathbb{Z}$. Indeed, 0 is even so (a) holds; the sum of two even integers is even so (b) holds; and the product of an even integer with any other integer is still even so (c) holds.

Example 8.4. Put $R = \mathbb{Z}[x]$ and $I = \{f \in \mathbb{Z}[x] \mid f(1) = 0\}$. For example $x^{10} - x \in I$ and $(x-3)(x-2)(x-1) \in I$ but $x + 7 \notin I$ (because $1 + 7 \neq 0$) and $\frac{1}{2}x^2 - x + \frac{1}{2} \notin I$ (because the coefficients are not integers, so $\frac{1}{2}x^2 - x + \frac{1}{2} \notin \mathbb{Z}[x]$). Clearly the zero polynomial is an element of I, so (a) holds. If $f, g \in I$ then $f(1) = g(1) = 0$ so $(f + g)(1) = f(1) + g(1) = 0 + 0 = 0$, so $f + g \in I$; thus (b) holds. If $f \in I$ and g is any polynomial then $(gf)(1) = g(1)f(1) = g(1).0 = 0$ so $gf \in I$; thus (c) holds. This shows that I is an ideal in R.

Example 8.5. Put $R = \mathbb{Z}[x]$ and $I = \{$ constant polynomials $\} \subseteq R$. Then I is *not* an ideal. Axioms (a) and (b) certainly hold. Moreover, if a and b are elements of I then $ab \in I$ also. However, axiom (c) says more than this: it says that if $b \in I$ and a is *any* element of R, not necessarily in I, then ab must be in I. However, $x \in R$ and $1 \in I$ but $x.1 \notin I$ so axiom (c) is violated.

Example 8.6. Let K be a field. It is easy to see that $\{0\}$ and K itself are ideals in K; I claim that these are the only ideals. Indeed, let I be an ideal in K. If $I \neq \{0\}$ then we have some nonzero element $b \in I$. For any element $c \in K$ we have $cb^{-1} \in K$ and $b \in I$ so axiom (c) tells us that $(cb^{-1})b \in I$, or in other words $c \in I$. This means that $I = K$, as required.

Example 8.7. Let R be any ring, and let x be any element of R. Define $Rx = \{ux \mid u \in R\}$. I claim that this is an ideal (called the *principal ideal* generated by x). First, we have $0 = 0.x \in Rx$, so axiom (a) holds. Second, if $a, b \in Rx$ then there exist u, v such that $a = ux$ and $b = vx$ so $a + b = (u + v)x$, so $a + b \in Rx$. This shows that (b) holds. Finally, if $a \in R$ and $b \in Rx$ then $b = vx$ for some $v \in R$ so $ab = (av)x \in Rx$, so (c) holds.

In example 8.3, the ideal I is just $\mathbb{Z}.2$. In example 8.4, the ideal I is just $\mathbb{Z}[x].(x-1)$.

Now let I be an ideal in a ring R. Recall that I is an R-submodule of R, so we can define the R-module $R/I = \{x + I \mid x \in R\}$ as before. In the case where I is the principal ideal Ra for some $a \in R$, we will generally write R/a rather than R/Ra.

We next show that R/I can itself be regarded as a ring.

Lemma 8.8. *If $a + I = a' + I$ and $b + I = b' + I$ then $ab + I = a'b' + I$.*

Proof. As $a + I = a' + I$, the element $u := a - a'$ lies in I. Similarly, the element $v := b - b'$ lies in I. We have $a = a' + u$ and $b = b' + v$ so

$$ab - a'b' = (a' + u)(b' + v) - a'b' = a'v + b'u + uv = (a' + u)v + b'u.$$

As $a' + u \in R$ and $v \in I$, axiom (c) says that $(a' + u)v \in I$. As $b' \in R$ and $u \in I$, axiom (c) also says that $b'u \in I$. As I is closed under addition, this means that $(a' + u)v + b'u \in I$, so $ab - a'b' \in I$, so $ab + I = a'b' + I$ as required. □

It follows that we can define multiplication of cosets unambiguously by $(a + I)(b + I) = ab + I$. By the method of Proposition 7.6 we see that this makes R/I into a ring.

Definition 8.9. Let R_0 and R_1 be rings. A *ring homomorphism* from R_0 to R_1 is a function $\alpha\colon R_0 \to R_1$ such that

(a) $\alpha(a + b) = \alpha(a) + \alpha(b)$ for all $a, b \in R_0$

(b) $\alpha(1) = 1$
(c) $\alpha(ab) = \alpha(a)\alpha(b)$ for all $a, b \in R_0$.

One can check that a ring homomorphism automatically satisfies $\alpha(0) = 0$ and $\alpha(-a) = -\alpha(a)$. Moreover, if a is invertible in R_0 then $\alpha(a)$ is invertible in R_1 with $\alpha(a)^{-1} = \alpha(a^{-1})$. We say that α is an *isomorphism* if it is a bijection as well as a homomorphism. If so, one can check that the inverse function $\alpha^{-1}\colon R_1 \to R_0$ is also a ring homomorphism.

If we define $\pi\colon R \to R/I$ by $\pi(a) = a + I$, we find that $\pi(a+b) = \pi(a) + \pi(b)$ and also $\pi(1) = 1$ and $\pi(ab) = \pi(a)\pi(b)$, in other words π is a homomorphism of rings.

The following result is the *First Isomorphism Theorem for Rings.*

Theorem 8.10. *Let $\alpha\colon R_0 \to R_1$ be a homomorphism of rings. Then* $\ker(\alpha)$ *is an ideal in R_0 and* $\operatorname{image}(\alpha)$ *is a subring of R_1. Moreover, there is a ring isomorphism $\overline{\alpha}\colon R_0/\ker(\alpha) \simeq \operatorname{image}(\alpha)$ given by $\overline{\alpha}(a + \ker(\alpha)) = \alpha(a)$. In particular, if α is surjective then $R_0/\ker(\alpha) \simeq R_1$.*

Proof. Put $K = \ker(\alpha)$ and $R_2 = \operatorname{image}(\alpha)$. If $a, b \in K$ then $\alpha(a) = \alpha(b) = 0$ so $\alpha(a + b) = \alpha(a) + \alpha(b) = 0$ so $a + b \in K$. Also, if c is any element of R_0 then $\alpha(ca) = \alpha(c)\alpha(a) = \alpha(c).0 = 0$, so $ca \in K$. This shows that K is an ideal in R_0.

As $\alpha(0) = 0$ and $\alpha(1) = 1$ we see that $0, 1 \in \operatorname{image}(\alpha) = R_2$. If $u, v \in R_2$ then we have $u = \alpha(a)$ and $v = \alpha(b)$ for some $a, b \in R_0$. Thus $u + v = \alpha(a + b)$ and $-u = \alpha(-a)$ and $uv = \alpha(ab)$, so $u + v, -u, uv \in R_2$. This shows that R_2 is a subring of R_1.

Just as in the proof of the first isomorphism theorem for modules, we have a well-defined bijection $\overline{\alpha}\colon R_0/K \to R_2$ given by $\overline{\alpha}(a + K) = \alpha(a)$. We then have

$$\overline{\alpha}((a + K)(b + K)) = \overline{\alpha}(ab + K) = \alpha(ab) = \alpha(a)\alpha(b) = \overline{\alpha}(a + K)\overline{\alpha}(b + K),$$

and similarly $\overline{\alpha}(1 + K) = 1$ and $\overline{\alpha}((a + K) + (b + K)) = \overline{\alpha}(a + K) + \overline{\alpha}(b + K)$, so $\overline{\alpha}$ is a ring homomorphism. □

Example 8.11. For any $n > 0$ we put $n\mathbb{Z} = \{nk \mid k \in \mathbb{Z}\} = \{m \in \mathbb{Z} \mid m = 0 \pmod n\}$. This is an ideal in $\mathbb{Z}$, so we have a factor ring $\mathbb{Z}/n\mathbb{Z}$. Note that $a + n\mathbb{Z} = b + n\mathbb{Z}$ iff $a - b \in n\mathbb{Z}$ iff $a = b \pmod n$. Using this, we see that $\mathbb{Z}/n\mathbb{Z}$ is just the usual ring $\mathbb{Z}_n$ of residue classes modulo n.

Example 8.12. Let K be a field, and λ an element of K. Put $I = \{f \in K[x] \mid f(\lambda) = 0\}$, which is an ideal in the ring $R := K[x]$. To see this, define a function $\alpha\colon K[x] \to K$ by $\alpha(f) = f(\lambda)$. Clearly

$$\begin{aligned} \alpha(f + g) &= (f + g)(\lambda) = f(\lambda) + g(\lambda) = \alpha(f) + \alpha(g) \\ \alpha(fg) &= (fg)(\lambda) = f(\lambda)g(\lambda) = \alpha(f)\alpha(g) \\ \alpha(1) &= 1 \end{aligned}$$

so α is a ring homomorphism. If $c \in K$ then we can regard c as a constant polynomial and we find that $\alpha(c) = c$; this shows that α is surjective. It is clear that $\ker(\alpha) = I$, so $K[x]/I \simeq K$ by the First Isomorphism Theorem.

It is also a standard fact that $f(\lambda) = 0$ iff f is divisible by $x - \lambda$, so I is just the principal ideal $K[x].(x - \lambda)$, so we have shown that $K[x]/(x - \lambda) \simeq K$.

Example 8.13. Let I be the principal ideal $\mathbb{R}[x](x^2 + 1)$ in $\mathbb{R}[x]$. I claim that $\mathbb{R}[x]/(x^2 + 1) = \mathbb{R}[x]/I$ is isomorphic to $\mathbb{C}$. The basic point is just that $\overline{x}^2 + 1 = \overline{x^2 + 1} = \overline{0}$, so $\overline{x}^2 = -\overline{1}$, so $\overline{x}$ is a square root of -1.

To give a formal proof, we consider the function $\alpha\colon \mathbb{R}[x] \to \mathbb{C}$ defined by $\alpha(f) = f(i)$. Just as in the last example we find that α is a ring homomorphism. Any complex number z can be written in the form $a + bi$ for some $a, b \in \mathbb{R}$ and we find that $\alpha(a + bx) = a + bi = z$, so α is surjective. Thus, if we put $I = \ker(\alpha) = \{f \in \mathbb{R}[x] \mid f(i) = 0\}$ we have $\mathbb{R}[x]/I \simeq \mathbb{C}$.

As $i^2 + 1 = 0$ we have $x^2 + 1 \in I$ and so $\mathbb{R}[x](x^2 + 1) \subseteq I$. Conversely, suppose that $f \in I$, so $f(i) = 0$. We can divide $f(x)$ by $x^2 + 1$ to get $f(x) = (x^2 + 1)q(x) + a + bx$ for some polynomial $q(x) \in \mathbb{R}[x]$ and $a, b \in \mathbb{R}$. We then have $0 = f(i) = q(i)(i^2 + 1) + a + bi = a + bi$, and by comparing real and imaginary parts we see that $a = b = 0$. This means that $f(x) = q(x)(x^2 + 1)$, so $f(x) \in \mathbb{R}[x].(x^2 + 1)$. This shows that $I = \mathbb{R}[x](x^2 + 1)$, and so $\mathbb{R}[x]/(x^2 + 1) = \mathbb{R}[x]/I \simeq \mathbb{C}$.

The next example relies on the following result.

Lemma 8.14. *Let p be a prime number and $k \geq 0$. If a is not divisible by p then $\overline{a}$ is invertible in $\mathbb{Z}_{p^k}$.*

Proof. Let d be the greatest common divisor of a and p^k. This is in particular a divisor of p^k, so it must be of the form p^j for some $j \leq k$. It must also be a divisor of a, which is impossible if $j > 0$, because a is not divisible by p. We must therefore have $j = 0$ or in other words $d = 1$. As $(a, p^k) = 1$ we have $ab + p^k c = 1$ for some integers b, c. This means that $ab = 1 \pmod{p^k}$ or in other words $\overline{a}\overline{b} = \overline{1}$. This shows that $\overline{a}$ is invertible (with inverse $\overline{b}$) as required. □

Example 8.15. Let p be a prime, and consider the ring $\mathbb{Z}_{(p)}$ as in example 2.12. Let I be the principal ideal $\mathbb{Z}_{(p)}.p^k$ for some $k \geq 0$. I claim that $\mathbb{Z}_{(p)}/p^k$ is isomorphic to $\mathbb{Z}_{p^k} = \mathbb{Z}/p^k$. To see this, we must define a homomorphism $\rho\colon \mathbb{Z}_{(p)} \to \mathbb{Z}/p^k$. Any element of $x \in \mathbb{Z}_{(p)}$ can be written as $x = a/b$ where $a, b \in \mathbb{Z}$ and $b \neq 0 \pmod{p}$. This means that $\overline{b}$ is invertible in $\mathbb{Z}/p^k$, so we have an element $\overline{a}\overline{b}^{-1} \in \mathbb{Z}/p^k$. Now suppose we describe x in a different way, say as $x = c/d$ with $c, d \in \mathbb{Z}$ and $d \neq 0 \pmod{p}$. Then $a/b = c/d$ so $ad = bc$ so $\overline{a}\overline{d} = \overline{b}\overline{c}$. As $\overline{b}$ and $\overline{d}$ are invertible we can divide through by them to get $\overline{a}\overline{b}^{-1} = \overline{c}\overline{d}^{-1}$. Thus, we can unambiguously define a function $\rho\colon \mathbb{Z}_{(p)} \to \mathbb{Z}/p^k$ by $\rho(a/b) = \overline{a}\overline{b}^{-1}$. It is easy to check that this is a ring homomorphism.

For any element $y \in \mathbb{Z}/p^k$ we can write $y = \overline{a}$ for some $a \in \{0, 1, \dots, p^k - 1\}$. Any of these numbers a can be regarded as an element of $\mathbb{Z}_{(p)}$ and then we have $\rho(a) = \overline{a} = y$. This shows that $\text{image}(\rho) = \mathbb{Z}/p^k$.

Next, suppose that $x \in \ker(\rho)$. Then $x = a/b$ for some $a, b \in \mathbb{Z}$ with $b \neq 0 \pmod{p}$ and $\overline{a}\overline{b}^{-1} = \rho(a/b) = 0$ in $\mathbb{Z}/p^k$. We can multiply this equation by $\overline{b}$ to see that $\overline{a} = 0$ in $\mathbb{Z}/p^k$, so $a = 0 \pmod{p^k}$, so $a = p^k c$ for some integer c. If we define $y = c/b$ we find that $y \in \mathbb{Z}_{(p)}$ and $x = p^k y$ so $x \in p^k \mathbb{Z}_{(p)}$. Conversely, if $x \in p^k \mathbb{Z}_{(p)}$ then $x = p^k y$ for some $y \in \mathbb{Z}_{(p)}$ so $\rho(y) \in \mathbb{Z}/p^k$ and $\rho(x) = p^k \rho(y)$. However, it is easy to see that $p^k z = 0$ for all $z \in \mathbb{Z}/p^k$, so $\rho(x) = 0$ so $x \in \ker(\rho)$.

The first isomorphism theorem now gives us an isomorphism $\overline{\rho}\colon \mathbb{Z}_{(p)}/\ker(\rho) \to \text{image}(\rho)$, or equivalently $\overline{\rho}\colon \mathbb{Z}_{(p)}/p^k \to \mathbb{Z}/p^k$.

Proposition 8.16. *Modules over R/I are the same thing as modules over R with the property that $am = 0$ for all $a \in I$ and $m \in M$.*

Proof. Let M be a module over R/I. To make M into an R-module, we need to define am for each $a \in R$ and $m \in M$. Note that $\pi(a) = (a + I) \in R/I$, and M is a module over R/I, so $\pi(a)m$ is already defined. We can thus define am to be $\pi(a)m$. It is easy to check that the axioms are satisfied; for example, we have

$$(a + b)m = \pi(a + b)m = (\pi(a) + \pi(b))m = \pi(a)m + \pi(b)m = am + bm,$$

so multiplication is left distributive. Thus M is a module over R/I. If $a \in I$ then $\pi(a) = 0$ so $am = \pi(a)m = 0$ as required.

Conversely, suppose that M is a module over R with the property that $am = 0$ for all $a \in I$ and $m \in M$. To make M into an R/I-module, we must define Am for each coset $A \in R/I$ and each $m \in M$. We would like to do this by writing A in the form $a + I$ for some $a \in R$ and defining Am to be the same as am. This raises the usual problem of ambiguity, but if $a + I = a' + I$ then $a - a' \in I$, so $(a - a')m = 0$ (by our assumption on M) so $am = a'm$. Thus, we have an unambiguous definition of Am for $A \in R/I$ and $m \in M$. It is again straightforward to check that the module axioms are satisfied. For example, if $A, B \in R/I$ we can choose $a, b \in R$ such that $A = a + I$ and $b = b + I$. We then have $A + B = (a + b) + I$ so

$$(A + B)m = ((a + b) + I)m = (a + b)m = am + bm = (a + I)m + (b + I)m = Am + Bm,$$

so multiplication is left distributive. Thus M is a module over R/I, as required. □

Example 8.17. Consider the Abelian group $V = \{0, a, b, c\}$ with addition table as follows:

$$a + a = b + b = c + c = 0$$
$$a + b = c$$
$$b + c = a$$
$$c + a = b.$$

Like any Abelian group, this can be regarded as a $\mathbb{Z}$-module. As $a + a = b + b = c + c = 0$, we see that $2v = 0$ for all $v \in V$, and thus that $nv = 0$ for all $n \in 2\mathbb{Z}$. It follows that V can be regarded as a module over the ring $\mathbb{Z}_2$.

Example 8.18. Let M be an Abelian group of order d, considered as a module over $\mathbb{Z}$ as usual. By Lagrange's theorem, if $m \in M$ then the order of m divides d. As we are using additive notation, this just means that $dm = 0$. It follows that $am = 0$ for all $a \in d\mathbb{Z}$, so M can be regarded as a module over $\mathbb{Z}_d$.

Example 8.19. Let W be the space of functions of the form $f(t) = a\cos(t) + b\sin(t)$ (with $a, b \in \mathbb{R}$). As in Example 5.9, we can regard this as a module over $\mathbb{R}[D]$. If f is as above then $f'(t) = -a\sin(t) + b\cos(t)$ and so $f''(t) = -a\cos(t) - b\sin(t) = -f(t)$, so $(D^2 + 1)f = f'' + f = 0$. If we let I be the principal ideal $(D^2 + 1)\mathbb{R}[D]$ we find that $p(D)f = 0$ for all $p(D) \in I$ and $f \in W$, so W can be regarded as a module over the ring $\mathbb{R}[D]/I = \mathbb{R}[D]/(D^2 + 1)$.

9. Euclidean domains

We next consider Euclidean domains, which are a particular kind of commutative ring. The idea is to generalize the following two facts:

1. If n and m are integers with $m \neq 0$ then we can divide n by m to get a quotient q with remainder r. We then have $n = mq + r$ and $|r| < |m|$.
2. If f and g are polynomials over $\mathbb{C}$ with $g \neq 0$ then we can divide f by g to get a quotient q with remainder r. We then have $f = gq + r$ and the degree of r is less than the degree of g.

Given a ring R and elements $a, b \in R$ with $b \neq 0$, we would like to do a similar kind of division to get an equation $a = bq + r$ where the "size" of r is less than the "size" of b. If $R = \mathbb{Z}$ then "size" means absolute value, and if $R = \mathbb{C}[x]$ then "size" means degree. For a general ring there may not be a suitable notion of size, so there may not be a useful division algorithm. A suitable notion of size is called a *Euclidean valuation*; the formal definition is as follows.

Definition 9.1. A *Euclidean valuation* on a ring R is a function $\nu(a)$ defined for all nonzero elements a of R such that

(a) $\nu(a)$ is a nonnegative integer whenever $a \neq 0$.
(b) If $a, b \in R$ and $b \neq 0$ then there are elements $q, r \in R$ such that $a = bq + r$ and either $r = 0$ or $\nu(r) < \nu(b)$.
(c) If $a, b \in R$ and $a \neq 0$ and $b \neq 0$ then $ab \neq 0$ and $\nu(a) \leq \nu(ab)$.

A *Euclidean domain* is a ring R for which there exists a Euclidean valuation.

Remark 9.2. The first part of condition (c) says that the product of two nonzero elements is nonzero, or in other words that R is an integral domain.

Example 9.3. The function $\nu(a) = |a|$ is a Euclidean valuation on $\mathbb{Z}$. Indeed, conditions (a) and (c) are clear, and (b) is just the ordinary division algorithm for integers.

Example 9.4. If K is a field then we can define a Euclidean valuation ν on $K[x]$ by $\nu(f) = \deg(f) =$ the degree of f. Indeed, conditions (a) and (c) are clear, and (b) is just the ordinary division algorithm for polynomials.

Example 9.5. The function $\nu(x + iy) = |x + iy|^2 = x^2 + y^2$ defines a Euclidean valuation on $\mathbb{Z}[i]$. Indeed, condition (a) is clear. For condition (b), suppose that $a = x + iy$ and $b = u + iv$ for some integers u, v, x, y. As $b \neq 0$ we can consider the complex number $a/b = s + it$ say, where $s, t \in \mathbb{R}$. Let s_0 be the closest integer to s, so $|s - s_0| \leq 1/2$. (If s has the form $m + 1/2$ for some integer

m then we could take $s_0 = m$ or $s_0 = m+1$; it doesn't matter which.) Similarly, we let t_0 be the closest integer to t, so $|t - t_0| \leq 1/2$. We put $q = s_0 + it_0$ and $r = a - qb$ so that $a = qb + r$. We find that

$$|a/b - q|^2 = |(s - s_0) + (t - t_0)i|^2 = (s - s_0)^2 + (t - t_0)^2 \leq 1/4 + 1/4 < 1,$$

so $|r|^2 = |a - qb|^2 = |b|^2|a/b - q|^2 < |b|^2$, or in other words $\nu(r) < \nu(b)$ as required.

Finally, for condition (c), we have $\nu(ab) = |ab|^2 = |a|^2|b|^2$. It is clear that $|b|^2$ is a nonnegative integer, and as $b \neq 0$ we have $|b|^2 \neq 0$ so $|b|^2 \geq 1$ so $\nu(ab) \geq \nu(a)$.

Example 9.6. We next define a Euclidean valuation on $\mathbb{Z}_{(p)}$. Any nonzero element $a \in \mathbb{Z}_{(p)}$ has the form u/w, where u and w are integers and w is not divisible by p. If we divide u by p as many times as possible we end up with an equation $u = p^t v$ where v is not divisible by p. We then define $\nu(a) = \nu(p^t v/w) = t$. For example, if $p = 3$ we have $\nu(567/13) = \nu(3^4 \times 7/13) = 4$.

Now suppose we have some other element $b \in \mathbb{Z}_{(p)}$ with $b \neq 0$. We can write this in the form $b = p^s x/y$ where p does not divide x or y, so $\nu(b) = s$. We then have $ab = p^{t+s}(vx)/(wy)$, from which it is not hard to see that $\nu(ab) = t + s = \nu(a) + \nu(b)$. Given this, condition (c) is clear.

Condition (b) is also satisfied, in a rather trivial way. If $t < s$ then we just put $q = 0$ and $r = a$ and we have $a = qb + r$ with $\nu(r) < \nu(b)$. On the other hand, if $t \geq s$ then the number $q := a/b = p^{t-s}(vy)/(wx)$ lies in $\mathbb{Z}_{(p)}$ and $a = bq$ so we can take $r = 0$.

For the rest of this section, we let R denote a Euclidean domain, with Euclidean valuation ν say.

Theorem 9.7. *Every ideal I in R is principal.*

Proof. Let I be an ideal; we must find an element $b \in R$ such that $I = Rb$. First, if $I = \{0\}$ then $I = R0$ as required; so we may assume that $I \neq \{0\}$. Each nonzero element $b \in I$ has a valuation $\nu(b) \in \mathbb{Z}$ with $\nu(b) \geq 0$. Choose such an element for which $\nu(b)$ is as small as possible. By assumption $b \in I$ and I is an ideal so $Rb \subseteq I$.

Conversely, suppose that $a \in I$. By axiom (b), then there are elements $q, r \in R$ such that $a = bq + r$ and either $r = 0$ or $\nu(r) < \nu(b)$. Note that $r = a - bq$ and $a, b \in I$ so $r \in I$. If r were a nonzero element of I with $\nu(r) < \nu(b)$, this would contradict our choice of b. We must therefore have $r = 0$ instead, so $a = qb$, so $a \in Rb$. This proves that $I \subseteq Rb$ and thus that $I = Rb$. □

Now suppose we have two elements $a, b \in R$. Then $Ra + Rb$ is an ideal in R, so by the theorem there must be an element d such that $Ra + Rb = Rd$. This raises the question of how to find d explicitly.

The first thing to note is that $Ra + Rb$ is *not* the same as $R(a + b)$. For example, take $R = \mathbb{Z}$ and $a = 3$ and $b = 2$. We can write any number n as $n = n \times 3 + (-n) \times 2$ so $n \in \mathbb{Z}.3 + \mathbb{Z}.2$, which shows that $\mathbb{Z}.3 + \mathbb{Z}.2 = \mathbb{Z}$. However, $\mathbb{Z}.(2 + 3)$ consists only of the multiples of 5, so it is not the same.

Definition 9.8. Let a and b be elements of R. We say that a is *divisible* by b if and only if there is an element $c \in R$ such that $a = bc$, or equivalently if $a \in Rb$, or equivalently if $Ra \subseteq Rb$. We also write $b|a$ if a is divisible by b.

Definition 9.9. Let a and b be elements of R. A *common divisor* of a and b is an element $d \in R$ such that a and b are both divisible by d. A *greatest common divisor* (or *gcd*) of a and b is an element d such that

(i) d is a common divisor of a and b; and
(ii) if d' is any other common divisor then d is divisible by d'.

Proposition 9.10. *We have $Ra + Rb = Rd$ if and only if d is a gcd of a and b. If so, then d can be written as $xa + yb$ for some $x, y \in R$.*

Proof. First suppose that $Ra + Rb = Rd$. We can write a as $1a + 0b$, so $a \in Ra + Rb = Rd$, so a is divisible by d. Similarly b is divisible by d, so d is a common divisor of a and b.

Next, it is clear that $d \in Rd$, so $d \in Ra + Rb$, so $d = xa + yb$ for some $x, y \in R$.

Finally, suppose that d' is another common divisor of a and b. Then $a = u'd'$ and $b = v'd'$ for some $u', v' \in R$. This means that $d = xa + yb = xu'd' + yv'd' = (xu' + yv')d'$, so d is divisible by d'. This proves that d is a gcd of a and b.

Conversely, suppose that d is a gcd of a and b. We know from Theorem 9.7 that $Ra + Rb = Rc$ for some $c \in R$, and we know from the first half of this proof that c must also be a gcd of a and b. As d is a greatest common divisor and c is a common divisor, we see that d is divisible by c, so $Rd \subseteq Rc$. As c is a greatest common divisor and d is a common divisor, we see that c is divisible by d, so $Rc \subseteq Rd$, so $Rd = Rc = Ra + Rb$ as required. □

We now explain how to actually find the gcd of two elements $a, b \in R$. The answer is to use the Euclidean algorithm, just as in $\mathbb{Z}$ or $\mathbb{R}[x]$. We may assume that $a, b \neq 0$ (otherwise the problem is trivial). We then define $a_0 = a$ and $b_0 = b$. By the definition of a Euclidean valuation, we can find $p_0, a_1 \in R$ such that $a_0 = p_0 b_0 + a_1$ and either $a_1 = 0$ or $\nu(a_1) < \nu(b_0)$. (Informally, a_1 is the remainder when we divide a_0 by b_0.) Assuming that $a_1 \neq 0$, we can then find q_1, b_1 such that $b_0 = q_1 a_1 + b_1$ and either $b_1 = 0$ or $\nu(b_1) < \nu(a_1)$. Assuming that $b_1 \neq 0$ we can find p_1, a_2 such that $a_1 = p_1 b_1 + a_2$ and either $a_2 = 0$ or $\nu(a_2) < \nu(b_1)$. Assuming that $a_2 \neq 0$, we can then find q_2, b_2 such that $b_1 = q_2 a_2 + b_2$ and either $b_2 = 0$ or $\nu(b_2) < \nu(a_2)$. Continuing in this way, we get a sequence

$$\nu(b_0) > \nu(a_1) > \nu(b_1) > \nu(a_2) > \nu(b_2) > \dots .$$

Now, all these valuations are nonnegative integers so they cannot keep decreasing forever. Thus, after a finite number of steps we must end up with either $a_k = 0$ or $b_k = 0$, forcing the process to stop.

Suppose that the first term to be zero is a_k, so the elements $a_0, \dots, a_{k-1}$ and $b_0, \dots, b_{k-1}$ are all nonzero. I claim that $Ra + Rb = Rb_{k-1}$, so that b_{k-1} is a gcd of a and b.

To see this, put $I = Ra + Rb$ and $J = Rb_{k-1}$; we must show that $I = J$. Certainly $a_0, b_0 \in I$. Using the equation $a_1 = a_0 - p_0 b_0$ we deduce that $a_1 \in I$. Using the equation $b_1 = b_0 - q_1 a_1$ we deduce that $b_1 \in I$. Using the equation $a_2 = a_1 - p_1 b_1$ we deduce that $a_2 \in I$. Continuing in this way, we see that $a_i, b_i \in I$ for all i and in particular that $b_{k-1} \in I$, so $J \subseteq I$.

We now use a similar argument in the opposite direction. Clearly $b_{k-1} \in J$. We have $a_{k-1} = p_{k-1} b_{k-1} + a_k$ but $a_k = 0$ so $a_{k-1} = p_{k-1} b_{k-1} \in J$. We can now use the equation $b_{k-2} = q_{k-1} a_{k-1} + b_{k-1}$ to show that $b_{k-2} \in J$, and then use the equation $a_{k-2} = p_{k-2} b_{k-2} + a_{k-1}$ to show that $a_{k-2} \in J$. Working backwards in this way, we eventually find that $a_0, b_0 \in J$ or in other words $a, b \in J$. This implies that $ua + vb \in J$ for all $u, v \in R$, or in other words that $I = Ra + Rb \subseteq J$. We have already seen that $J \subseteq I$, so $I = J$ as required.

All this assumed that the first term to be zero was a_k. It could instead happen that the first term to be zero was b_k, in which case a very similar argument would show that $Ra + Rb = Ra_k$, so a_k is a gcd of a and b.

10. Factorization in Euclidean domains

Let R be a Euclidean domain. We say that an element $a \in R$ is a *nonunit* if it is not invertible in R. We say that an element is *reducible* if it can be written as a product of two nonunits. We say that it is *irreducible* if it is a nonunit but is not reducible.

Note that 0 is a nonunit and $0 = 0.0$ so 0 is a product of two nonunits and thus is reducible. For the rest of this section, we exclude the element 0 from consideration.

Example 10.1. Let p be a prime number. Then the only ways to factor p in $\mathbb{Z}$ are $p = 1.p = (-1).(-p) = p.1 = (-p).(-1)$. In each case, one of the factors is either 1 or -1 and thus is invertible in $\mathbb{Z}$. Thus, p is irreducible. Similarly, $-p$ is irreducible, but if $n > 0$ and n is composite then n and $-n$ are reducible.

Example 10.2. Consider the ring $\mathbb{C}[x]$. I claim that the irreducible elements are precisely the polynomials of the form $ax + b$ with $a \neq 0$. Firstly, polynomials of degree 0 are invertible (because we have excluded the zero polynomial from consideration). Thus, any nonunit has degree at least

1, so any reducible polynomial has degree at least 2. Thus when $a \neq 0$ the polynomial $ax + b$ is a nonunit but not reducible, so it is irreducible.

Next, let $f(x)$ be any polynomial of degree $d > 1$. By the Fundamental Theorem of Algebra, f has a root, in other words there is a complex number a such that $f(a) = 0$. This means that f is divisible by $x - a$, say $f(x) = (x - a)g(x)$ for some polynomial g. Note that g has degree $d - 1 > 0$, so both $g(x)$ and $x - a$ are nonunits, so f is reducible. This shows that the irreducibles are precisely the polynomials of degree exactly one, as claimed.

Example 10.3. Any element in $x \in \mathbb{Z}_{(p)}$ can be written as $x = p^v a/b$ where a and b are not divisible by p and so a/b is a unit in $\mathbb{Z}_{(p)}$. In other words, every element is a unit multiple of p^v for some $v \geq 0$. Using this we find that x is a unit iff $v = 0$, that x is irreducible iff $v = 1$, and that x is reducible iff $v > 1$.

Definition 10.4. We write $a \sim b$ if and only if there is an invertible element $u \in R$ such that $au = b$. If so, we say that a is a *unit multiple* or *associate* of b. It is not hard to see that $a \sim b$ if and only if $Ra = Rb$, and that the relation $\sim$ is an equivalence relation.

Note that in $\mathbb{Z}$, both 7 and -7 are irreducible, but they are unit multiples of each other so for many purposes it makes no sense to use both of them. It is thus traditional to ignore -7. Similarly, in $\mathbb{C}[x]$ both $x - 2$ and $2x - 4$ are irreducible but they are unit multiples of each other, so we usually ignore $2x - 4$. This leads us to make the following definitions.

Definition 10.5. A *complete set of irreducibles* in a Euclidean domain R is a set $\mathcal{P}$ of irreducibles such that for every irreducible p, there is a unique irreducible $p' \in \mathcal{P}$ such that $p' \sim p$.

Note that for any Euclidean domain, we can always choose a complete set of irreducibles. We simply divide the set of all irreducibles up into equivalence classes under the relation $\sim$, and we pick one irreducible from each equivalence class.

Example 10.6. The set of positive prime numbers is a complete set of irreducibles in $\mathbb{Z}$. The set $\{x - a \mid a \in \mathbb{C}\}$ is a complete set of irreducibles in $\mathbb{C}[x]$. The set $\{p\}$ (with just one element) is a complete set of irreducibles in $\mathbb{Z}_{(p)}$.

Proposition 10.7. *If p is irreducible in R then R/p is a field.*

Proof. As p is not a unit we see that 1 is not divisible by p, so $\overline{1} \neq \overline{0}$ in R/p.

Every nonzero element of R/p has the form $\overline{a} = a + Rp$, where $a \in R$ but $a \notin Rp$. We then have $Ra + Rp = Rd$, where d is a gcd of a and p. Thus d divides a and p, say $a = ud$ and $p = vd$. As p is irreducible, one of v and d must be a unit. If v is a unit we have $a = ud = uv^{-1}p$ so $a \in Rp$, contrary to assumption. Thus d must be a unit instead, so $1 \in Rd = Ra + Rp$, so $1 = xa + yp$ for some $x, y \in R$. This means that $\overline{xa} = \overline{1}$ in R/p, so $\overline{a}$ is invertible as required. □

Corollary 10.8. *If a and b are not divisible by p, then ab is not divisible by p.*

Proof. $a + Rp$ and $b + Rp$ are nonzero elements of the field R/p, so $ab + Rp = (a + Rp)(b + Rp)$ is nonzero, so p does not divide ab. □

Lemma 10.9. *If a and b are nonzero and b is a nonunit then $\nu(ab) > \nu(a)$.*

Proof. Axiom (c) for Euclidean valuations says that $\nu(ab) \geq \nu(a)$. It is thus enough to suppose that $\nu(ab) = \nu(a)$ and deduce a contradiction. We can divide a by ab with remainder to get $a = abq + r$ for some q, r with either $r = 0$ or $\nu(r) < \nu(ab) = \nu(a)$. Note that $a(1 - bq) = r$. As b is not a unit we cannot have $bq = 1$, so $1 - bq \neq 0$. We also have $a \neq 0$ so $r \neq 0$, so $\nu(r) < \nu(a)$. However, Axiom (c) also says that $\nu(a(1 - bq)) \geq \nu(a)$, giving the required contradiction. □

Lemma 10.10. *If $\nu(a) = 0$ then a is invertible.*

Proof. Divide 1 by a to get $1 = qa + r$ with $r = 0$ or $\nu(r) < \nu(a)$. As valuations are always nonnegative we cannot have $\nu(r) < \nu(a)$ so we must have $r = 0$, so $1 = qa$, so q is an inverse for a. □

Theorem 10.11. *Let R be a Euclidean domain, and let $\mathcal{P}$ be a complete set of irreducibles in R. Then any nonzero element $a \in R$ can be written in the form $a = up_1^{n_1} \dots p_r^{n_r}$, where $p_1, \dots, p_r$ are distinct irreducibles in $\mathcal{P}$ and $n_1, \dots, n_r \in \mathbb{N}$ and u is invertible. Moreover, this factorization is unique (except that it could be written in a different order, for example $3^2 \times 5^3 = 5^3 \times 3^2$).*

Proof. We first show that any nonzero element $a \in R$ can be written as a unit times a product of standard irreducibles, by induction on $\nu(a)$. If $\nu(a) = 0$ then a is a unit, which we think of as a unit times the product of the empty list of standard irreducibles. If $\nu(a) > 0$ then a is a nonunit. If a is irreducible then it has the form up, where u is a unit and p is a standard irreducible. Otherwise, a can be written as a product of two nonunits, say $a = bc$. By Lemma 10.9 we see that $\nu(b)$ and $\nu(c)$ are strictly less than $\nu(a)$. By induction we may assume that b and c can be written as units times products of standard irreducibles, and it follows that the same is true of a. By collecting factors together, we can write $a = up_1^{n_1} \dots p_r^{n_r}$, where $p_1, \dots, p_r$ are distinct irreducibles in $\mathcal{P}$ and $n_1, \dots, n_r \in \mathbb{N}$ and u is invertible.

Suppose we have another such factorization $a = vq_1^{m_1} \dots q_s^{m_s}$, where $q_1, \dots, q_s$ are distinct irreducibles in $\mathcal{P}$ and $m_1, \dots, m_s \in \mathbb{N}$ and v is invertible. I claim that $p_1 = q_j$ for some j. If not, then all the elements q_j would be indivisible by p_1, as would v, and Corollary 10.8 would tell us that $v \prod_j q_j^{m_j}$ is indivisible by p_1, so a is indivisible by p_1, which is clearly false. Thus $p_1 = q_j$ for some j, and the same argument shows that each p_i is a q_j, and similarly that each q_j is a p_i. Thus, after renumbering the q's if necessary, we may assume that $r = s$ and $p_i = q_i$ for all i.

Now suppose that $n_1 \leq m_1$. Put $b = up_2^{n_2} \dots p_r^{n_r}$ and $c = vp_1^{m_1 - n_1} p_2^{m_2} \dots p_r^{m_r}$. We then have $p_1^{n_1} b = a$ and $p_1^{n_1} c = a$ and $p_1^{n_1} \neq 0$ so $b = c$. Using Corollary 10.8 again we see that b is not divisible by p_1 so c is not divisible by p_1 so $m_1 - n_1 = 0$ so $m_1 = n_1$. A similar argument works if $n_1 \geq m_1$, so $n_1 = m_1$ in all cases. The same method shows that $n_i = m_i$ for all i, and given this, it is clear that $u = v$ as well. □

11. Finite free modules over a Euclidean domain

Throughout this section, we let R denote a Euclidean domain.

Definition 11.1. A *finite free module* over R is an R-module M that is isomorphic to R^d for some nonnegative integer d.

To understand this definition more explicitly, we introduce the notion of a basis.

Definition 11.2. Let M be an R-module. We say that a list $\{m_1, \dots, m_d\}$ of elements of M is a *basis* if for every element $m \in M$ there is precisely one list $(u_1, \dots, u_d) \in R^d$ such that $m = u_1 m_1 + \dots + u_d m_d$.

Proposition 11.3. *Let M be an R-module. Then M is a finite free module if and only if it has a basis.*

Proof. Suppose that $\{m_1, \dots, m_d\}$ is a basis. As in example 6.7, we can define a homomorphism $\phi\colon R^d \to M$ by $\phi(u_1, \dots, u_d) = u_1 m_1 + \dots + u_d m_d$. By the definition of a basis, for each $m \in M$ there is precisely one element $u \in R^d$ such that $\phi(u) = m$. This means that ϕ is a bijection, and thus an isomorphism. Thus $M \simeq R^d$, so M is free.

Conversely, suppose that M is free, so we have an isomorphism $\phi\colon R^d \to M$ for some d. By the second half of Example 6.7, there is a list $m_1, \dots, m_d$ of elements of M such that $\phi(u_1, \dots, u_d) = u_1 m_1 + \dots + u_d m_d$ for all $(u_1, \dots, u_d) \in R^d$. As ϕ is an isomorphism, for each element $m \in M$ there is a unique element $u \in R^d$ with $\phi(u) = m$, and this means precisely that $\{m_1, \dots, m_d\}$ is a basis. □

Example 11.4. Put $M = \{(x, y, z) \in \mathbb{Z}^3 \mid x = y\}$. Then the vectors $m_1 := (1, 1, 0)$ and $m_2 := (0, 0, 1)$ clearly lie in M. Moreover, any vector in M can be written uniquely in the form $(x, x, z) = xm_1 + zm_2$. Thus $\{m_1, m_2\}$ is a basis for M as a $\mathbb{Z}$-module, so M is free.

Example 11.5. Put $M = \{(x, y, z) \in \mathbb{Z}^3 \mid x + y + z = 0 \pmod 3\}$, considered as a module over $\mathbb{Z}$. I claim that this is free as a $\mathbb{Z}$-module. To prove this, we need to find some elements of M which can form a basis. Firstly, if $x + y + z = 0$ then (x, y, z) will certainly lie in M. Some

obvious vectors satisfying $x+y+z=0$ are $m_1 := (1,-1,0)$ and $m_2 := (0,1,-1)$. Also, the vector $m_3 := (0,0,3)$ satisfies $x+y+z=3$ so $x+y+z=0 \pmod 3$ so $m_3 \in M$. We can thus define a map $\phi\colon \mathbb{Z}^3 \to M$ by $\phi(u_1,u_2,u_3) = u_1m_1 + u_2m_2 + u_3m_3$.

Now suppose we have an element $m=(x,y,z) \in M$. Note that $x+y+z=0 \pmod 3$, so $x+y+z=3t$ for some $t \in \mathbb{Z}$. We have

$$m - xm_1 = (x,y,z) - x(1,-1,0) = (0,x+y,z),$$

and so

$$m - xm_1 - (x+y)m_2 = (0,x+y,z) - (x+y)(0,1,-1) = (0,0,x+y+z) = (0,0,3t) = tm_3,$$

so $m = xm_1 + (x+y)m_2 + tm_3 = \phi(x,x+y,t)$. This shows that ϕ is surjective.

Now suppose that $(u,v,w) \in \ker(\phi)$, so $um_1+vm_2+wm_3=0$. By looking at the x coordinates we see that $u.1+v.0+w.0=0$ so $u=0$. Thus, the original equation becomes $vm_2+wm_3=0$. By looking at the y coordinates we see that $v.1+w.0=0$, so $v=0$, so the equation becomes $wm_3=0$. By looking at the z coordinates we see that $3w=0$ but w is just an integer so the only way $3w$ can be 0 is if $w=0$. This proves that $\ker(\phi)=\{(0,0,0)\}$, so ϕ is injective and thus is an isomorphism.

We next give a convenient test for showing that certain modules are *not* free.

Definition 11.6. For any R-module M, an element $m \in M$ is a *torsion element* if there is some nonzero element $a \in R$ such that $am=0$. We write $\mathrm{tors}(M)$ for the set of all torsion elements. We say that M is *torsion-free* if $\mathrm{tors}(M)=\{0\}$. Equivalently, a module M is torsion-free if whenever $a \in R$ and $x \in M$ are both nonzero, their product ax is also nonzero.

Lemma 11.7. *Every finite free module is torsion-free.*

Proof. Every finite free module is isomorphic to R^d for some d, so it is enough to show that R^d is torsion-free. Let a be a nonzero element of R, and let $u=(u_1,\dots,u_d)$ be a nonzero element of R^d. This means that $u_i \neq 0$ for some i. As R is an integral domain, it follows that $au_i \neq 0$, and thus that the vector $au=(au_1,\dots,au_d)$ is not the zero vector. □

Example 11.8. Suppose that $n>1$, and consider $\mathbb{Z}_n$ as a $\mathbb{Z}$-module. Then $\overline{1}$ is a nonzero element of $\mathbb{Z}_n$ and n is a nonzero element of $\mathbb{Z}$ but $n.\overline{1} = \overline{n} = \overline{0}$. Thus $\mathbb{Z}_n$ is not torsion-free as a $\mathbb{Z}$-module, so it cannot be a finite free module.

Example 11.9. Consider $C^\infty(\mathbb{R},\mathbb{R})$ as a module over $\mathbb{R}[D]$ in the usual way. Then the function $f(t)=e^t$ is a nonzero element of $C^\infty(\mathbb{R},\mathbb{R})$, and $D-1$ is a nonzero element of $\mathbb{R}[D]$, but $(D-1)f = f'-f = 0$. Thus $C^\infty(\mathbb{R},\mathbb{R})$ is not torsion-free, so it is not a finite free module over $C^\infty(\mathbb{R},\mathbb{R})$. For an even simpler proof, just consider the constant function $g(t)=1$, so $Dg=g'=0$, again showing that $C^\infty(\mathbb{R},\mathbb{R})$ is not torsion-free.

We now prove a key theorem, which makes the theory of modules over Euclidean domains much simpler than for other rings.

Theorem 11.10. *Any submodule of a finite free module over a Euclidean domain is free.*

Note that Examples 11.4 and 11.5 illustrate this.

Proof. As every finite free module is isomorphic to R^d for some d, it will be enough to show that every submodule of R^d is free. We do this by induction on d.

The case $d=0$ is easy. The module R^0 is just $\{0\}$, the only submodule of this is $\{0\}$ itself, and this is free because it is R^0.

The case $d=1$ is essentially given by Theorem 9.7. Let M be a submodule of $R^1=R$, or in other words an ideal in R. By Theorem 9.7 we have $M=Ra$ for some $a \in R$. If $a=0$ then $M=\{0\}=R^0$ so M is free. If $a \neq 0$ then I claim that $\{a\}$ is a basis for M. Indeed, as $M=Ra$, every element $m \in M$ can certainly be written as $m=ua$ for some $u \in R$. If we have $ua=m=va$ then $(u-v)a=0$ but $a \neq 0$ and R is an integral domain so $u-v=0$ so $u=v$. Thus m can be written *uniquely* in the form $m=ua$, so $\{a\}$ is a basis and so M is free.

Now suppose that $d > 1$ and that we have proved that every submodule of R^{d-1} is free. We need to show that every submodule $M \leq R^d$ is free. Let F be the set of vectors of the form $(x_1, x_2, \dots, x_{d-1}, 0)$ in R^d. Note that F is a copy of R^{d-1}, so every submodule of F is free. In particular, $M \cap F$ is a submodule of F so it is free, with basis $\{m_1, \dots, m_r\}$ say. Define a homomorphism $\pi\colon M \to R$ by $\pi(x_1, \dots, x_d) = x_d$. The image of π is a submodule of R so it has the form Ra for some $a \in R$.

Suppose that $a = 0$. Then $\text{image}(\pi) = Ra = \{0\}$, so $\pi(m) = 0$ for all $m \in M$, so every element of M has the form $(x_1, \dots, x_{d-1}, 0)$. This means that every element of M is an element of F, so M is a submodule of F, so M is free.

Suppose instead that $a \neq 0$. As $a \in Ra = \text{image}(\pi)$, we can choose $m_{r+1} \in M$ such that $\pi(m_{r+1}) = a$. I claim that $\{m_1, \dots, m_r, m_{r+1}\}$ is a basis for M. To see this, let m be an element of M. Then $\pi(m) \in \text{image}(\pi) = Ra$, so $\pi(m) = u_{r+1}a$ for some $u_{r+1} \in R$. Put $m' = m - u_{r+1}m_{r+1}$, so $m' \in M$ and

$$\pi(m') = \pi(m) - u_{r+1}\pi(m_{r+1}) = u_{r+1}a - u_{r+1}a = 0.$$

This means that the last coordinate of m' is 0, so $m' \in F$. As $m' \in M \cap F$ and $\{m_1, \dots, m_r\}$ is a basis for $M \cap F$ we see that $m' = u_1m_1 + \dots + u_rm_r$ for some $u_1, \dots, u_r \in R$, so

$$m = m' + u_{r+1}m_{r+1} = u_1m_1 + \dots + u_{r+1}m_{r+1}.$$

This shows that the elements $m_1, \dots, m_{r+1}$ generate M.

Now suppose that we have elements $v_1, \dots, v_{r+1} \in R$ satisfying $v_1m_1 + \dots + v_{r+1}m_{r+1} = 0$. I claim that $v_1 = \dots = v_{r+1} = 0$. Indeed, as $m_1, \dots, m_r \in F$ we have $\pi(m_1) = \dots = \pi(m_r) = 0$, so when we apply π to the previous equation we get $v_{r+1}\pi(m_{r+1}) = 0$, or equivalently $v_{r+1}a = 0$. As $a \neq 0$ and R is an integral domain this means that $v_{r+1} = 0$. Thus, our original equation becomes $v_1m_1 + \dots + v_rm_r = 0$. As $\{m_1, \dots, m_r\}$ is a basis for $M \cap F$, the only way we can have $v_1m_1 + \dots + v_rm_r = 0$ is if $v_1 = \dots = v_r = 0$, as claimed.

It now follows that $\{m_1, \dots, m_{r+1}\}$ is a basis for M, so M is free. This completes the inductive step, and thus the proof of the theorem. □

Corollary 11.11. *Let M be a finitely generated module over a Euclidean domain R, and let N be a submodule of M. Then N is also finitely generated.*

Proof. As M is finitely generated, there is a list $m_1, \dots, m_d$ of elements of M such that an arbitrary element $m \in M$ can be written in the form $u_1m_1 + \dots + u_dm_d$. Define $\phi\colon R^d \to M$ by $\phi(u_1, \dots, u_d) = u_1m_1 + \dots + u_dm_d$; this is clearly a surjective homomorphism. Put $L = \{u \in R^d \mid \phi(u) \in N\}$. I claim that this is a submodule of R^d. Indeed, if $u, v \in L$ then $\phi(u), \phi(v) \in N$ so $\phi(u + v) = \phi(u) + \phi(v) \in N$ so $u + v \in L$. Similarly, if $u \in L$ and $a \in R$ then $\phi(u) \in N$ so $\phi(au) = a\phi(u) \in N$ so $au \in L$, so L is a submodule as claimed. Submodules of R^d are finite free modules by the theorem, so we can choose a basis $\{p_1, \dots, p_r\}$ for L. Put $n_i = \phi(p_i)$; as $p_i \in L$ we have $n_i \in N$. I claim that the elements $n_1, \dots, n_r$ generate N. Indeed, suppose $n \in N$. Then $n \in M$ and the homomorphism $\phi\colon R^d \to M$ is surjective so we have $n = \phi(u)$ for some $u \in R^d$. As $\phi(u) = n \in N$ we see that $u \in L$, so u can be written in the form $u = v_1p_1 + \dots + v_rp_r$ for some $v_1, \dots, v_r \in R$. It follows that

$$n = \phi(u) = v_1\phi(p_1) + \dots + v_r\phi(p_r) = v_1n_1 + \dots + v_rn_r.$$

This shows that the elements $n_1, \dots, n_r$ generate N as claimed, so N is finitely generated. □

12. Row and column operations

We saw in the last section that a submodule of a finite free module over a Euclidean domain is free. We next give a systematic method for finding a basis for such a submodule.

Suppose we have vectors $u_1, \ldots, u_n$ in R^m, and we let N be the submodule generated by $u_1, \ldots, u_n$. It will be convenient to form the $n \times m$ matrix A whose columns are the vectors u_i.

Definition 12.1. An *elementary column operation* on a matrix A over a ring R is any of the following operations:

(a) Add a multiple of one column to another column
(b) Multiply a column by an invertible element of R
(c) Exchange two columns.

We say that a matrix A is in (unreduced) column echelon form if

(a) All nonzero columns occur to the left of all the zero columns.
(b) If the i'th and $(i+1)$'st columns are nonzero, then the top nonzero entry in the $(i+1)$'s column is below the top nonzero entry in the i'th column.

Example 12.2. The matrix

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 \\ 3 & 6 & 0 & 0 \\ 4 & 7 & 9 & 0 \\ 5 & 8 & 10 & 0 \end{pmatrix}$$

is in column echelon form, but the following matrices are not

$$\begin{pmatrix} 0 & 3 & 0 \\ 1 & 4 & 0 \\ 2 & 5 & 0 \end{pmatrix} \qquad \begin{pmatrix} 0 & 1 & 0 \\ 0 & 2 & 4 \\ 0 & 3 & 5 \end{pmatrix} \qquad \begin{pmatrix} 0 & 0 & 8 \\ 0 & 4 & 0 \\ 2 & 0 & 0 \end{pmatrix}.$$

Proposition 12.3. *Let A be an $n \times m$ matrix over a Euclidean domain R, and let N be the submodule of R^m generated by the columns of A. Then A can be reduced by elementary column operations to a matrix B in column echelon form, and the nonzero columns of B form a basis for N.*

Remark 12.4. The resulting matrix B is not unique; in general, a matrix A can be converted to many different matrices B in column echelon form.

The algorithm is a mixture of the Euclidean algorithm and the usual Gaussian algorithm for column-reducing a matrix over a field. Suppose that the top row of A is nonzero (if not, we simply ignore any rows of zeros at the top and work with the first nonzero row). Of all the nonzero elements in the top row, we find one whose valuation is as small as possible. After swapping the columns around, we can assume that this element occurs in the top left corner of the matrix, so it is the entry a_{11}. If we were working over a field, we would subtract a_{k1}/a_{11} times the first column from the k'th column (for $k = 2, \ldots, n$), and then the top row would have the form $(a_{11}, 0, \ldots, 0)$. As we are not working over a field, we do not necessarily have $a_{k1}/a_{11} \in R$, so we cannot perform these operations. As the next best thing, we divide a_{k1} by a_{11} by the division algorithm to obtain $a_{k1} = a_{11}q_k + r_k$ with either $r_k = 0$ or $\nu(r_k) < \nu(a_{11})$. We then subtract q_k times the first column from the k'the column for $k = 2, \ldots, n$. The top row is now $(a_{11}, r_2, \ldots, r_n)$. If $r_2 = \ldots = r_n = 0$ then we have a matrix of the form $\left(\begin{array}{c|c} a_{11} & 0 \\ \hline * & A' \end{array}\right)$. We can then column-reduce the smaller matrix A' by the same process, and eventually we get a column-reduced form for A. Generally, however, the elements $r_2, \ldots, r_n$ will not all be zero. Among those that are nonzero, we choose one whose valuation is as small as possible. Note that by construction this valuation is strictly less than that of a_{11} and valuations are always nonnegative integers, so this kind of step can only occur finitely many times. We swap the columns around to put the element of minimum valuation in the top left corner and repeat the whole process.

We end up with a matrix B in column echelon form. Let N' be the submodule of R^m generated by the columns of B. It is clear from the form of B that its columns are linearly independent, so they form a basis for N'. To prove Proposition 12.3, it will be enough to check that $N' = N$, or equivalently that when we perform an elementary column operation on a matrix, the module spanned by the columns does not change. We will explain this by example rather than giving a formal proof. Suppose that the matrix has three columns, say u_1, u_2 and u_3.

(a) A typical operation of the first type replaces the list (u_1, u_2, u_3) by $(u_1, u_2 + cu_1, u_3)$ for some $c \in R$. If a vector v can be written as $a_1u_1 + a_2u_2 + a_3u_3$, it can also be written as $(a_1 - ca_2)u_1 + a_2(u_2 + cu_1) + a_3u_3$, so $\text{span}\{u_1, u_2, u_3\} \subseteq \text{span}\{u_1, u_2 + cu_1, u_3\}$. Conversely, if v can be written as $b_1u_1 + b_2(u_2 + cu_1) + b_3u_3$, it can also be written as $(b_1 + cb_2)u_1 + b_2u_2 + b_3u_3$, so $\text{span}\{u_1, u_2 + cu_1, u_3\} \subseteq \text{span}\{u_1, u_2, u_3\}$.

(b) A typical operation of the second type replaces the list (u_1, u_2, u_3) by (u_1, cu_2, u_3) for some invertible element $c \in R$. If a vector v can be written as $a_1u_1 + a_2u_2 + a_3u_3$, it can also be written as $a_1u_1 + (a_2c^{-1})(cu_2) + a_3u_3$, so $\text{span}\{u_1, u_2, u_3\} \subseteq \text{span}\{u_1, cu_2, u_3\}$. Conversely, if v can be written as $b_1u_1 + b_2(cu_2) + b_3u_3$, it can also be written as $b_1u_1 + (b_2c)u_2 + b_3u_3$, so $\text{span}\{u_1, cu_2, u_3\} \subseteq \text{span}\{u_1, u_2, u_3\}$.

(c) A typical operation of the third type replaces the list (u_1, u_2, u_3) by (u_2, u_1, u_3), and this clearly does not change the span.

A formal proof would be essentially the same but would need more elaborate notation.

Example 12.5. Here we perform a column reduction over $\mathbb{Z}$ by strictly following the algorithm.

$$\begin{pmatrix} 5 & 8 & 11 & 3 \\ 16 & 25 & 34 & 9 \end{pmatrix} \xrightarrow{1} \begin{pmatrix} 3 & 5 & 8 & 11 \\ 9 & 16 & 25 & 34 \end{pmatrix}$$
$$\xrightarrow{2} \begin{pmatrix} 3 & 2 & 2 & 2 \\ 9 & 7 & 7 & 7 \end{pmatrix}$$
$$\xrightarrow{3} \begin{pmatrix} 2 & 3 & 2 & 2 \\ 7 & 9 & 7 & 7 \end{pmatrix}$$
$$\xrightarrow{4} \begin{pmatrix} 2 & 1 & 0 & 0 \\ 7 & 2 & 0 & 0 \end{pmatrix}$$
$$\xrightarrow{5} \begin{pmatrix} 1 & 2 & 0 & 0 \\ 2 & 7 & 0 & 0 \end{pmatrix}$$
$$\xrightarrow{6} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 3 & 0 & 0 \end{pmatrix}$$

In step 1 we note that the element of smallest valuation in the top row is the 3 in the 4'th column. We therefore move the 4'th column round to be the first column. In step 2 we divide the remaining entries in the top row by 3. We have $5 = 1 \times 3 + 2$ and $8 = 2 \times 3 + 2$ and $11 = 3 \times 3 + 2$ so we subtract 1 times the first column from the second column, 2 times the first column from the third column, and 3 times the first column from the fourth column. In step 3 we note that the element of smallest valuation on the top row is 2, so we swap the first two columns to put a 2 in the top left corner. Then in step 4, we subtract the first column from each of the other columns. The element of smallest valuation in the top row is now 1, so we swap the first two columns to put the 1 in the top left corner; this is step 5. Finally, we subtract twice the first column from the second column. Our matrix is now in the form $\left(\begin{array}{c|c} a_{11} & 0 \\ \hline * & A' \end{array}\right)$, and the matrix $A' = (3\ 0\ 0)$ is already in echelon form so we are finished.

If we use a little imagination rather than strictly following the algorithm, we can reduce the matrix more quickly:

$$\begin{pmatrix} 5 & 8 & 11 & 3 \\ 16 & 25 & 34 & 9 \end{pmatrix} \xrightarrow{1} \begin{pmatrix} -1 & -1 & -1 & 3 \\ -2 & -2 & -2 & 9 \end{pmatrix}$$
$$\xrightarrow{2} \begin{pmatrix} -1 & 0 & 0 & 0 \\ -2 & 0 & 0 & 3 \end{pmatrix}$$
$$\xrightarrow{3} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 3 & 0 & 0 \end{pmatrix}.$$

The conclusion is that if N is the submodule of $\mathbb{Z}^2$ generated by the vectors $(5,16)$, $(8,25)$, $(11,34)$ and $(3,9)$ then $\{(1,2),(0,3)\}$ is a basis for N.

Example 12.6. Here is a column reduction over $\mathbb{Q}[x]$.

$$\begin{pmatrix} x^2 & x-1 & x^3 \\ 0 & -1 & x \\ 0 & -1 & 0 \end{pmatrix} \xrightarrow{1} \begin{pmatrix} 1 & x-1 & 1 \\ x+1 & -1 & x^2+2x+1 \\ x+1 & -1 & x^2+x+1 \end{pmatrix}$$
$$\xrightarrow{2} \begin{pmatrix} 1 & 0 & 0 \\ x+1 & -x^2 & x^2+x \\ x+1 & -x^2 & x^2 \end{pmatrix}$$
$$\xrightarrow{3} \begin{pmatrix} 1 & 0 & 0 \\ x+1 & x^2 & x \\ x+1 & x^2 & 0 \end{pmatrix}$$
$$\xrightarrow{4} \begin{pmatrix} 1 & 0 & 0 \\ x+1 & x & 0 \\ x+1 & 0 & x^2 \end{pmatrix}$$

In step 1 we subtracted $x+1$ times the middle column from the first column, and x^2+x+1 times the middle column from the third column. In step 2 we subtracted $x-1$ times the first column from the middle column, and and subtracted the first column from the last column. In step 4 we multiplied the middle column by -1 and then subtracted it from the last column. In step 4 we subtracted x times the last column from the middle column, and then swapped the middle and last columns.

The conclusion is that the vectors $m_1 := (1, x+1, x+1)$, $m_2 := (0, x, 0)$ and $m_3 := (0, 0, x^2)$ form a basis for the submodule of $\mathbb{Q}[x]^3$ generated by the columns of the original matrix.

In fact, the columns of the original matrix also form a basis for this submodule (just because they are linearly independent over $\mathbb{Q}[x]$). However, the basis $\{m_1, m_2, m_3\}$ is easier to work with because it is in echelon form.

Example 12.7. Put $v_k = (k, k^2, k^3, k^4) \in \mathbb{Z}^4$, and let N be the subgroup of $\mathbb{Z}^4$ generated by $v_1, \ldots, v_5$. These vectors are the columns of the first matrix written below:

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 1 & 4 & 9 & 16 & 25 \\ 1 & 8 & 27 & 64 & 125 \\ 1 & 16 & 81 & 256 & 625 \end{pmatrix} \to \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 2 & 6 & 12 & 20 \\ 1 & 6 & 24 & 60 & 120 \\ 1 & 14 & 78 & 252 & 620 \end{pmatrix}$$
$$\to \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 2 & 0 & 0 & 0 \\ 1 & 6 & 6 & 24 & 60 \\ 1 & 14 & 36 & 168 & 480 \end{pmatrix}$$
$$\to \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 2 & 0 & 0 & 0 \\ 1 & 6 & 6 & 0 & 0 \\ 1 & 14 & 36 & 24 & 0 \end{pmatrix}$$

We deduce that the vectors $(1,1,1,1)$, $(0,2,6,14)$, $(0,0,6,36)$ and $(0,0,0,24)$ form a basis for N. Note that the "leading terms" of these vectors are 1!, 2!, 3! and 4!; this is part of a general pattern.

The above algorithm helps us to find a basis for a submodule N of R^n. However, we often want instead to investigate the structure of R^n/N, and so far we do not have a method for this. The simplest case is where N is generated by the elements $d_1e_1, \dots, d_re_r$ for some $r \leq n$ and some nonzero elements $d_1, \dots, d_r$ in R, where e_i is the i'th standard basis vector. It is then easy to see that

$$R^n/N \simeq R/d_1 \oplus \dots \oplus R/d_r \oplus R^{n-r}.$$

More generally, suppose we can find a (non-standard) basis $u_1, \dots, u_n$ for R^n and nonzero elements $d_1, \dots, d_r$ such that N is generated by $d_1u_1, \dots, d_ru_r$. We again find that $R^n/N \simeq R/d_1 \oplus \dots \oplus R/d_r \oplus R^{n-r}$. Our next algorithm will enable us to find bases of this type.

Definition 12.8. A matrix over a Euclidean domain is in *normal form* if it has the form $\left(\begin{array}{c|c} D & 0 \\ \hline 0 & 0 \end{array}\right)$, where

1. D is an $r \times r$ matrix for some r (called the *rank*)
2. The diagonal entries in D are nonzero, and the other entries are zero
3. If the diagonal entries are $d_1, \dots, d_r$, then $d_1|d_2|\dots|d_r$.

For example, the following matrix over $\mathbb{Z}$ is in normal form:

$$\left(\begin{array}{ccc|cc} 2 & 0 & 0 & 0 & 0 \\ 0 & 6 & 0 & 0 & 0 \\ 0 & 0 & 12 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 \end{array}\right)$$

We have already defined elementary column operations over a Euclidean domain, and elementary row operations are defined in the obvious analogous way.

Theorem 12.9. *Let A be an $n \times m$ matrix over a Euclidean domain R, and let N be the quotient of R^m by the span of the columns of A. Then A can be transformed by row and column operations to a matrix B in normal form. Moreover, if the nonzero diagonal entries in B are $d_1, \dots, d_r$ then $N \simeq R/d_1 \oplus \dots \oplus R/d_r \oplus R^{m-r}$.*

The proof will follow after some auxiliary definitions and preliminary results.

Remark 12.10. It could happen that some of the elements d_i are units. In this case we have $Rd_i = R$ and so $R/d_i = R/R = 0$ so we can drop the term R/d_i from the direct sum.

Definition 12.11. Let R be a Euclidean domain with valuation ν. For any nonzero matrix A over R, we let $\nu(A)$ be the smallest of the valuations of all the nonzero entries in A. For example, if $R = \mathbb{Z}$ we have $\nu\left(\begin{smallmatrix} 5 & -4 \\ 0 & -3 \end{smallmatrix}\right) = \nu(-3) = 3$.

Definition 12.12. We say that a nonzero $n \times m$ matrix A over R is *prenormal* if it has the form $\left(\begin{array}{c|c} d & 0 \\ \hline 0 & B \end{array}\right)$, where

1. d is a nonzero element of R
2. B is an $(n-1) \times (m-1)$ matrix over R
3. every entry in B is divisible by d.

Lemma 12.13. *Let A be a nonzero matrix over a Euclidean domain R. Then A can be transformed into prenormal form by elementary row and column operations.*

Proof. Let a_{ij} be the entry in the i'th column and j'th row of A. Put $n = \nu(A)$, which is a nonnegative integer. By definition we have $\nu(a_{ij}) = n$ for some i, j, and after reordering the rows and columns we may assume that $\nu(a_{11}) = n$. Put $d = a_{11}$.

Suppose that every entry in A is divisible by d. Then for each $i > 1$ we have $a_{1i} = q_id$ for some $q_i \in R$. We can subtract q_i times the top row from the i'th row for each i to get a new matrix

in which every row except the first starts with 0. It is easy to see that in this new matrix, every entry is still divisible by d. We can then subtract multiples of the first column from the other columns to get a matrix of the form $\left(\begin{array}{c|c} d & 0 \\ \hline 0 & B \end{array}\right)$ where every entry in B is divisible by d. Thus A can be made prenormal, as claimed.

Now suppose instead that some entry in A is not divisible by d. I claim that A can be transformed to a new matrix A' with $\nu(A') < n = \nu(A)$. Indeed, if some entry a_{1i} in the first column is not divisible by d then we have $a_{1i} = dq + r$ for some nonzero remainder r with $\nu(r) < \nu(d) = n$. If we subtract q times the first row from the i'th row we get a new matrix A' in which the i'th row starts with r, so $\nu(A') \leq \nu(r) < n$. A similar argument works if some entry a_{j1} in the first row is not divisible by d. This leaves the case where all entries in the first row or the first column are divisible by d, but some entry a_{ij} (with $i, j > 1$) is not divisible by d. After reordering some rows and columns we may assume that a_{22} is not divisible by d. We thus have $a_{22} = dq + r$ where $r \neq 0$ and $\nu(r) < \nu(d) = n$. We also have $a_{12} = ud$ and $a_{21} = vd$ for some u, v. The top left corner of our matrix looks like this:

$$\begin{pmatrix} a & vd \\ ud & qd + r \end{pmatrix}.$$

We now subtract v times the first row from the second row, then add the second row to the top row, then subtract $(q - uv + u)$ times the first column from the second column. The effect on the top left corner is as follows:

$$\begin{pmatrix} d & ud \\ vd & qd + r \end{pmatrix} \to \begin{pmatrix} d & ud \\ 0 & (q - uv)d + r \end{pmatrix} \to \begin{pmatrix} d & (q - uv + u)d + r \\ 0 & (q - uv)d + r \end{pmatrix} \to \begin{pmatrix} d & r \\ 0 & (q - uv)d + r \end{pmatrix}.$$

Thus, the matrix A' that we end up with has r as an entry, so $\nu(A') \leq \nu(r) < n$ as claimed.

We now repeat the whole process. We find that either A' can be made prenormal, or it can be transformed to another matrix A'' with $\nu(A'') < \nu(A')$, and so on. As valuations of matrices are nonnegative integers, we cannot keep on reducing them indefinitely, so eventually we must get to a matrix that can be made prenormal. □

Proposition 12.14. *Any matrix A over a Euclidean domain R can be transformed to normal form by row and column operations.*

Proof. Let A_0 be a matrix over a Euclidean domain, of shape $n \times m$ say. If $A_0 = 0$ then A_0 is already in normal form. Otherwise, by the Proposition, we can convert A_0 by row and column operations to the form $\left(\begin{array}{c|c} d_1 & 0 \\ \hline 0 & A_1 \end{array}\right)$, where every entry in A_1 is divisible by d_1. Clearly A_1 has shape $(n-1) \times (m-1)$. If A_1 is nonzero we can convert it to the form $\left(\begin{array}{c|c} d_2 & 0 \\ \hline 0 & A_2 \end{array}\right)$, where the entries in A_2 are divisible by d_2. As this is obtained from A_1 by row and column operations, we see that all the entries are still divisible by d_1, and in particular $d_1 | d_2$. This converts A_0 to the form

$$\left(\begin{array}{cc|c} d_1 & 0 & 0 \\ 0 & d_2 & 0 \\ \hline 0 & 0 & A_2 \end{array}\right).$$

The theorem follows by iterating this process in the evident way. □

Example 12.15.

$$\begin{pmatrix} 13 & 3 \\ 7 & 2 \end{pmatrix} \to \begin{pmatrix} 6 & 1 \\ 7 & 2 \end{pmatrix} \to \begin{pmatrix} 6 & 1 \\ -5 & 0 \end{pmatrix} \to \begin{pmatrix} 0 & 1 \\ 5 & 0 \end{pmatrix} \to \begin{pmatrix} 1 & 0 \\ 0 & 5 \end{pmatrix}.$$

This row reduction follows the method implicit in the above proof, except that we have not bothered to reorder the columns until the end. We start by finding the entry of smallest valuation, which is 2. We subtract the bottom row from the top row to make the entry above the 2 have smaller valuation than 2 does. Now the 1 in the top right hand corner has smaller valuation than everything else, and everything is divisible by 1. We can thus use row and column operations to

clear away the rest of the entries in the same row or column as the 1. We then swap the columns to put the matrix in normal form.

Example 12.16.

$$\begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix} \to \begin{pmatrix} 2 & 3 \\ 0 & 3 \end{pmatrix} \to \begin{pmatrix} 2 & 1 \\ 0 & 3 \end{pmatrix} \to \begin{pmatrix} 2 & 1 \\ -6 & 0 \end{pmatrix} \to \begin{pmatrix} 1 & 0 \\ 0 & 6 \end{pmatrix}.$$

Here we start with a diagonal matrix whose entry of smallest valuation is in the 2 in the top left corner. However, the other entries are not all divisible by 2, so the matrix is not prenormal. We first perform a row operation to ensure that there is a non-divisible entry on the top row. We then subtract 2 times the first column from the second column, leaving a remainder of 1 in the top right corner, which has valuation smaller than 2. We can now clear the entries below and to the left of the 1 and then reorder to get a normal matrix.

Example 12.17. Put $v_k = (k, k^2, k^3, k^4) \in \mathbb{Z}^4$, and let N be the subgroup of $\mathbb{Z}^4$ generated by $v_1, \dots, v_5$. We showed in Example 12.7 that the corresponding matrix can be column-reduced as follows:

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 1 & 4 & 9 & 16 & 25 \\ 1 & 8 & 27 & 64 & 125 \\ 1 & 16 & 81 & 256 & 625 \end{pmatrix} \to \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 2 & 0 & 0 & 0 \\ 1 & 6 & 6 & 0 & 0 \\ 1 & 14 & 36 & 24 & 0 \end{pmatrix}$$

We can now perform row operations to reduce this matrix further. This works easily in this particular example, because the entries underneath the 1 are divisible by 1, the entries under the 2 are divisible by 2, and the entry under the 6 is divisible by 6. Explicitly, we subtract the first row from each of the other rows; then we subtract 3 times the second row from the third row, and 7 times the second row from the last row; then we subtract 6 times the third row from the last row. This gives the matrix

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 6 & 0 & 0 \\ 0 & 0 & 0 & 24 & 0 \end{pmatrix}.$$

The conclusion will be that $\mathbb{Z}^4/N \simeq \mathbb{Z}_2 \oplus \mathbb{Z}_6 \oplus \mathbb{Z}_{24}$. (We have omitted the $\mathbb{Z}_1$ because $\mathbb{Z}_1 = 0$.)

Definition 12.18. Let E be an $n \times n$ matrix over a ring R. We say that E is *elementary* if either

1. It is obtained from the identity matrix by exchanging two rows (or equivalently, exchanging two columns); or
2. The diagonal entries are all equal to 1, there is a single nonzero entry off the diagonal, and all other entries are zero; or
3. One of the diagonal entries is an invertible element of R, the remaining diagonal entries are equal to 1, and all other entries are zero.

Note that elementary matrices are always invertible, and that their inverses are also elementary matrices.

Just as in the case of fields, if A' is obtained from A by a single elementary row operation then $A' = EA$ for some elementary matrix E. For example, let A be an $n \times 3$-matrix. Then

1. Exchanging the first and third rows is the same as multiplying on the left by $\left(\begin{smallmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{smallmatrix}\right)$.
2. Adding the third row to the first is the same as multiplying on the left by $\left(\begin{smallmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{smallmatrix}\right)$.
3. Multiplying the second row by -1 is the same as multiplying the matrix on the left by $\left(\begin{smallmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{smallmatrix}\right)$.

Similarly, if A' is obtained from A by a single column operation then $A' = AE$ for some elementary matrix E. Thus, if A' is obtained from A by a sequence of row and column operations then $A' = PAQ$ for some matrices P and Q which are products of elementary matrices and thus are invertible. Using this, the following corollary follows immediately from Proposition 12.14

Corollary 12.19. *If A is an $n \times m$ matrix over a Euclidean domain, then there exist invertible square matrices P and Q (of size m and n respectively) such that PAQ is in normal form.* □

Lemma 12.20. *Let F be a free module over R, and let $u_1, \dots, u_m$ be a basis for F. Suppose that $r \leq m$ and $d_1, \dots, d_r \in R$. Let M be the submodule of F generated by $d_1u_1, \dots, d_ru_r$. Then $F/M \simeq R/d_1 \oplus \dots \oplus R/d_r \oplus R^{m-r}$.*

Proof. Write $Q = R/d_1 \oplus \dots \oplus R/d_r \oplus R^{m-r}$ for brevity. Any element $x \in F$ can be written uniquely in the form $x_1u_1 + \dots + x_mu_m$ with $x_1, \dots, x_m \in R$. It therefore makes sense to define a map $\alpha\colon F \to Q$ by

$$\alpha(x_1u_1 + \dots + x_mu_m) = (x_1 + Rd_1, \dots, x_r + Rd_r, x_{r+1}, \dots, x_m).$$

This is easily seen to be a homomorphism. Suppose we have an element $y = (y_1, \dots, y_m) \in Q$, so $y_i \in R/d_i$ for $i \leq r$ and $y_i \in R$ for $i > r$. For $i \leq r$ we can choose $x_i \in R$ such that $y_i = x_i + Rd_i$. We can then define $u = (x_1, \dots, x_r, y_{r+1}, \dots, y_m) \in R^m$ and we find that $\alpha(u) = y$. Thus α is surjective, and the First Isomorphism Theorem tells us that $Q \simeq F/\ker(\alpha)$.

We now need to understand $\ker(\alpha)$. If $x = \sum_i x_iu_i \in \ker(\alpha)$ then $(x_1 + Rd_1, \dots, x_r + Rd_r, x_{r+1}, \dots, x_m) = (0, \dots, 0)$ which means that $x_i = 0$ for $i > r$ and $x_i \in Rd_i$ for $i \leq r$. This means that for $i \leq r$ we have elements $y_i \in R$ such that $x_i = y_id_i$ and so

$$x = x_1u_1 + \dots + x_ru_r = y_1(d_1u_1) + \dots + y_r(d_ru_r),$$

so $x \in \operatorname{span}\{d_1u_1, \dots, d_ru_r\} = M$. This shows that $\ker(\alpha) \subseteq M$, and the reverse inclusion is easy so $\ker(\alpha) = M$. We thus have $Q \simeq F/M$ as claimed. □

Proof of Theorem 12.9. Let A be an $n \times m$ matrix over a Euclidean domain R, let let M be the submodule of R^m spanned by the columns of A, and put $N = R^m/M$. Choose P and Q as in Corollary 12.19, so the matrix $C := PAQ$ is in normal form, with diagonal entries $d_1, \dots, d_r$ say. Put $B = AQ$. This is obtained from A by elementary column operations, so the columns of B span the same submodule as the columns of A; in other words, they span N.

Next, note that $PB = C$ so $B = P^{-1}C$. Let $u_1, \dots, u_m$ be the columns of P^{-1}; as P^{-1} is invertible, these form a basis for R^m. I claim that the columns of $P^{-1}C$ are $d_1u_1, \dots, d_ru_r, 0, \dots, 0$ (with $n - r$ zeros). This is clear in the following special case, where $n = 4$, $m = 3$ and $r = 2$:

$$P^{-1} = \begin{pmatrix} u_{11} & u_{21} & u_{31} \\ u_{12} & u_{22} & u_{32} \\ u_{13} & u_{23} & u_{33} \end{pmatrix}$$

$$C = \begin{pmatrix} d_1 & 0 & 0 & 0 \\ 0 & d_2 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

$$P^{-1}C = \begin{pmatrix} d_1u_{11} & d_2u_{21} & 0 & 0 \\ d_1u_{12} & d_2u_{22} & 0 & 0 \\ d_1u_{13} & d_2u_{23} & 0 & 0 \end{pmatrix}$$

The general case is the same except for more elaborate notation.

Thus M is spanned by $d_1u_1, \dots, d_ru_r$, and it follows that $N = R^m/M$ is isomorphic to $R/d_1 \oplus \dots \oplus R/d_r \oplus R^{m-r}$. □

Corollary 12.21. *Let M be a finitely generated module over a Euclidean domain R. Then*

$$M \simeq R/d_1 \oplus \dots \oplus R/d_r \oplus R^s$$

for some $r, s \geq 0$ and some nonunits $d_1, \dots, d_r \in R$ with $d_1|d_2|\dots|d_r$.

Proof. As M is finitely generated, we can find a finite list $v_1, \dots, v_m$ of elements that generate M. We can then define a homomorphism $\alpha\colon R^m \to M$ by

$$\alpha(a_1, \dots, a_m) = a_1v_1 + \dots + a_mv_m.$$

This is surjective because the elements v_i span M. Thus, the First Isomorphism Theorem for modules tells us that $M \simeq R^m/\ker(\alpha)$. We also know that $\ker(\alpha)$ is a submodule of R^m and

thus is a finite free module, with basis $u_1, \dots, u_n$ say. Let A be the $n \times m$ matrix with columns $u_1, \dots, u_n$, so M is isomorphic to the quotient of R^m by the span of the columns of A. The claim now follows □

Corollary 12.22. *Let M be a finitely generated Abelian group. Then*

$$M \simeq \mathbb{Z}_{d_1} \oplus \dots \oplus \mathbb{Z}_{d_r} \oplus \mathbb{Z}^s$$

for some $r, s \geq 0$ and some natural numbers $d_1, \dots, d_r$ with $1 < d_1 | d_2 | \dots | d_r$. □

13. Primary decomposition

Corollary 12.21 gives us in some sense a complete list of all finitely generated R-modules. However, it turns out not to be the most convenient kind of list to work with. Moreover, it still leaves us with a question about uniqueness. For example, both $\mathbb{Z}_4 \oplus \mathbb{Z}_{12}$ and $\mathbb{Z}_2 \oplus \mathbb{Z}_{24}$ appear in the list of finitely generated Abelian groups, and so far we have no way of knowing whether they might be isomorphic. In this section we will modify the classification given in Corollary 12.21 to get a new classification which is more useful in practice and which allows us to prove a uniqueness theorem.

Let R be a Euclidean domain, and choose a complete set of irreducibles $\mathcal{P}$.

Let M be an R-module. Recall that in Definition 11.6 we defined $\mathrm{tors}(M)$ to be the set of torsion elements in M, in other words the set of elements $m \in M$ such that there is a nonzero element $a \in R$ with $am = 0$.

Proposition 13.1. $\mathrm{tors}(M)$ *is a submodule of M.*

Proof. Suppose that $m, n \in \mathrm{tors}(M)$. Then there are nonzero elements $a, b \in R$ such that $am = 0$ and $bn = 0$. It follows that $ab \neq 0$ and $ab(m+n) = b(am) + a(bn) = b.0 + a.0 = 0$, so $m + n \in \mathrm{tors}(M)$. Similarly, for any $c \in R$ we have $a(cm) = c(am) = 0$, so $cm \in \mathrm{tors}(M)$, so $\mathrm{tors}(M)$ is a submodule as claimed. □

Definition 13.2. We say that M is a *torsion module* if $\mathrm{tors}(M) = M$. We say that M is a *finite torsion module* if it is finitely generated as well as being a torsion module.

Definition 13.3. A *basic* R-module is an R-module of the form R/p^k for some $p \in \mathcal{P}$ and $k > 0$.

We will show that any finite torsion module is isomorphic to a direct sum of basic modules.

Example 13.4. For any $m \in \mathbb{Z}_{12}$ we have $12m = 0$, so m is a torsion element. This shows that $\mathbb{Z}_{12}$ is a torsion module over $\mathbb{Z}$. More generally, let M be any finite Abelian group, considered as a $\mathbb{Z}$-module. If d is the order of M then $dm = 0$ for all $m \in M$, which shows that M is a torsion module. We saw in Example 5.25 that M is also finitely generated, so it is a finite torsion module.

Example 13.5. Let W_2 be the set of functions of the form $f(t) = a + bt + ct^2$, considered as an $\mathbb{R}[D]$-module in the usual way. For a function f of this form we have $f'(t) = b + 2ct$ so $f''(t) = 2c$ so $f'''(t) = 0$. This means that $D^3 f = 0$, so f is a torsion element of W. It follows that W_2 is a torsion module. We saw in Example 5.26 that W_2 is a cyclic module over $\mathbb{R}[D]$ and thus is finitely generated, so it is a finite torsion module over $\mathbb{R}[D]$. Similarly, the space W_d of polynomials of degree at most d is a finite torsion module over $\mathbb{R}[D]$.

Example 13.6. Consider the Abelian group $M = \mathbb{Q}/\mathbb{Z}$ as a $\mathbb{Z}$-module. Any element $m \in M$ has the form $a/b + \mathbb{Z}$ for some $a, b \in \mathbb{Z}$ with $b \neq 0$. it follows that $bm = a + \mathbb{Z}$, but $a \in \mathbb{Z}$ so $a + \mathbb{Z} = \mathbb{Z}$, which is the zero element of the group M. Thus $bm = 0$, proving that m is a torsion element. It follows that M is a torsion module. It is not finitely generated, however.

Example 13.7. Let K be a field, and let λ and μ be elements of K. Consider K^2 as a module over $K[x]$ using the endomorphism $x.(u, v) = \phi(u, v) = (\lambda u, \mu v)$. Then

$$(x - \lambda).(u, v) = (\lambda u, \mu v) - (\lambda u, \lambda v) = (0, (\mu - \lambda)v),$$

and

$$(x - \mu).(u, v) = (\lambda u, \mu v) - (\mu u, \mu v) = ((\lambda - \mu)u, 0),$$

so

$$(x-\mu)(x-\lambda).(u,v) = (x-\mu).(0,(\mu-\lambda)v) = (0,0).$$

In other words, the element $p(x) = (x-\mu)(x-\lambda) \in K[x]$ satisfies $p(x).(u,v) = (0,0)$ for all $(u,v) \in K^2$, which proves that K^2 is a torsion module.

Example 13.8. More generally, let V be any module over $K[x]$ that is finite-dimensional (of dimension d say) when considered as a vector space over K. I claim that V is a finite torsion module over $K[x]$. Indeed, suppose that $v \in V$. We then have vectors $v, xv, x^2v, \dots, x^dv$ in V. There are $d+1$ vectors in this list, but V only has dimension d, so the vectors in our list must be linearly dependent. Thus, there are scalars $a_0, \dots, a_d \in K$ (not all zero) such that $a_0v + a_1xv + \dots + a_dx^dv = 0$. This means that the polynomial $f(x) = a_0 + a_1x + \dots + a_dx^d$ is nonzero and satisfies $f.v = 0$, so v is a torsion element. This shows that V is a torsion module. Moreover, we can choose a basis $v_1, \dots, v_d$ for V as a vector space, and it follows that these elements generate V as a module over $K[x]$, so V is finitely generated.

Next recall that for any two rings R_0, R_1 we can form the product ring $R_0 \times R_1$. The elements of $R_0 \times R_1$ are pairs (a_0, a_1) with $a_0 \in R_0$ and $a_1 \in R_1$. Addition and multiplication are defined in the obvious way:

$$(a_0,a_1) + (b_0,b_1) = (a_0+b_0, a_1+b_1)$$
$$(a_0,a_1)(b_0,b_1) = (a_0b0, a_1b_1).$$

The additive identity is the element $(0,0)$, and the multiplicative identity is the element $(1,1)$.

The following result is called the *Chinese remainder theorem.*

Proposition 13.9. *If a and b are coprime then $R/ab \simeq R/a \times R/b$ as rings (or as R-modules).*

Proof. Define $\alpha\colon R \to R/a \times R/b$ by $\alpha(t) = (t+aR, t+bR)$. Note that $\alpha(s+t) = \alpha(s) + \alpha(t)$ and $\alpha(st) = \alpha(s)\alpha(t)$ and $\alpha(1) = 1$, so α is a homomorphism of rings.

As a and b are coprime, we have $xa + yb = 1$ for some $x, y \in R$.

Suppose that $\alpha(t) = (0,0)$. Then $t + aR$ is the zero coset $0 + aR$, so t is divisible by a, say $t = au$ for some u. Similarly $t = bv$ for some v. This means that

$$t = 1.t = (xa+yb)t = xat + ybt = xa(bv) + yb(au) = (xv+yu)ab,$$

so t is divisible by ab. Conversely, if t is divisible by ab then it is divisible by both a and b, so $\alpha(t) = (0,0)$. Thus $\ker(\alpha) = Rab$.

Now suppose we have some element $(u+aR, v+bR) \in R/a \times R/b$. Consider the element $t = ybu + xav \in R$. Note that $t = (1-xa)u + xav = u + xa(v-u) = u \pmod a$ and $t = ybu + (1-yb)v = v + yb(u-v) = v \pmod b$, so $\alpha(t) = (t+Ra, t+Rb) = (u+Ra, v+Rb)$. This shows that α is surjective. Thus, Theorem 8.10 says that $R/ab \simeq R/a \oplus R/b$. □

Corollary 13.10. *Any finite torsion R-module is isomorphic to a direct sum of basic R-modules.*

Proof. Let M be a finite torsion R-module. By Corollary 12.21 we have $M \simeq R/d_1 \oplus \dots \oplus R/d_r \oplus R^s$ say. If $s > 0$ then M contains a copy of R so it cannot be a torsion module, contrary to assumption. Thus we must have $s = 0$ and so $M = R/d_1 \oplus \dots \oplus R/d_r$. It will thus be enough to show that R/d is a direct sum of basic modules for all $d \neq 0$. For this we factor d as $up_1^{n_1} \dots p_r^{m_r}$ as in Theorem 10.11. As $Rd = Rdu^{-1}$ we may replace d by du^{-1} and thus assume that $d = p_1^{n_1} \dots p_r^{n_r}$. Put $q_i = p_i^{n_i}$ and $b_i = q_iq_{i+1} \dots q_r$, so $b_1 = d$. Note that $d = q_1b_2$ and q_1 and b_2 are coprime. Thus Proposition 13.9 tells us that $R/d \simeq R/q_1 \oplus R/b_2$. Similarly, $b_2 = q_2b_3$ and q_2 and b_3 are coprime so $R/b_2 \simeq R/q_2 \oplus R/b_3$ so $R/d = R/q_1 \oplus R/q_2 \oplus R/b_3$. Continuing if the obvious way we find that $R/d = R/q_1 \oplus \dots \oplus R/q_r$, and the modules R/q_i are basic, as required. □

Example 13.11. In the case $R = \mathbb{Z}$, we deduce that any finite Abelian group is a direct sum of groups of the form $\mathbb{Z}_{p^k} = \mathbb{Z}/p^k$ where p is prime and $k > 0$. Suppose that $M = B_1 \oplus \dots \oplus B_t$ where $B_i \simeq \mathbb{Z}/p_i^{n_i}$. (Note that the primes p_i need not all be different.) We then have

$$|M| = |B_1| \dots |B_t| = p_1^{n_1} \dots p_t^{n_t}.$$

Consider for example the case where $|M| = 81 = 3^4$. Clearly all the primes p_i must be equal to 3, and $3^4 = |M| = 3^{n_1} \dots 3^{n_t} = 3^{n_1+\dots n_t}$ so $n_1 + \dots + n_t = 4$. Given this, it is not hard to see that the possibilities are as follows:

$$\begin{aligned} M &\simeq \mathbb{Z}_{81} \\ M &\simeq \mathbb{Z}_{27} \oplus \mathbb{Z}_3 \\ M &\simeq \mathbb{Z}_9 \oplus \mathbb{Z}_9 \\ M &\simeq \mathbb{Z}_9 \oplus \mathbb{Z}_3 \oplus \mathbb{Z}_3 \\ M &\simeq \mathbb{Z}_3 \oplus \mathbb{Z}_3 \oplus \mathbb{Z}_3 \oplus \mathbb{Z}_3. \end{aligned}$$

Now consider instead the case where $|M| = 36 = 2^2 3^2$. Then all the primes p_i must be either 2 or 3. The orders $|B_i| = p_i^{n_i}$ must divide 36 so we must have $n_i = 1$ or 2. Using this it is not hard to see that the possibilities are as follows:

$$\begin{aligned} M &\simeq \mathbb{Z}_4 \oplus \mathbb{Z}_9 \\ M &\simeq \mathbb{Z}_4 \oplus \mathbb{Z}_3 \oplus \mathbb{Z}_3 \\ M &\simeq \mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_9 \\ M &\simeq \mathbb{Z}_2 \oplus \mathbb{Z}_2 \oplus \mathbb{Z}_3 \oplus \mathbb{Z}_3. \end{aligned}$$

It remains to discuss the question of uniqueness. Could it happen, for example, that the groups $A := \mathbb{Z}_9 \oplus \mathbb{Z}_9$ and $B := \mathbb{Z}_9 \times \mathbb{Z}_3 \times \mathbb{Z}_3$ are isomorphic? The answer is no, because A contains 9 elements satisfying $3a = 0$, whereas B contains 27 elements satisfying $3b = 0$. An elaboration of this argument will show that any finite torsion module can be written in an essentially unique way as a direct sum of basic modules.

Definition 13.12. Suppose that $p \in \mathcal{P}$ and $k \in \mathbb{N}$. For any finite torsion module M, we define

$$F_p^k(M) = \{x \in p^{k-1}M \mid px = 0\}.$$

This is easily seen to be a submodule of M. As M is finitely generated, we see from Corollary 11.11 that $F_p^k(M)$ is finitely generated. As $px = 0$ for all $x \in F_p^k(M)$, we can regard $F_p^k(M)$ as a module over R/p. As R/p is a field, every finitely generated module over it has a well-defined dimension, so we can define

$$f_p^k(M) = \dim_{R/p}(F_p^k(M)).$$

We also define

$$g_p^k(M) = f_p^k(M) - f_p^{k+1}(M).$$

Remark 13.13. It is easy to see that a pair $(x, y) \in M \oplus N$ lies in $F_p^k(M \oplus N)$ if and only if $x \in F_p^k(M)$ and $y \in F_p^k(N)$. It follows that $F_p^k(M \oplus N) = F_p^k(M) \oplus F_p^k(N)$ and thus that $f_p^k(M \oplus N) = f_p^k(M) + f_p^k(N)$ and $g_p^k(M \oplus N) = g_p^k(M) + g_p^k(N)$.

Remark 13.14. Suppose that $M \simeq M'$. I claim that $F_p^k(M) \simeq F_p^k(M')$, and thus that $f_p^k(M) = f_p^k(M')$ and $g_p^k(M) = g_p^k(M')$. Indeed, let $\phi\colon M \to M'$ be an isomorphism. Then if $x \in F_p^k(M)$ then $x = p^{k-1}y$ for some y and $px = 0$, so $\phi(x) = p^{k-1}\phi(y)$ and $p\phi(x) = 0$, so $\phi(x) \in F_p^k(M')$. Thus, ϕ gives a homomorphism from $F_p^k(M)$ to $F_p^k(M')$. Similarly, the homomorphism $\phi^{-1}\colon M' \to M$ restricts to give a homomorphism from $F_p^k(M')$ to $F_p^k(M)$. It is easy to see that these two maps are inverse to each other, so $F_p^k(M) \simeq F_p^k(M')$ as claimed.

Proposition 13.15. *We have*

$$g_p^k(R/q^j) = \begin{cases} 1 & \text{if } p = q \text{ and } k = j \\ 0 & \text{otherwise.} \end{cases}$$

Proof. We first prove that

$$f_p^k(R/q^j) = \begin{cases} 0 & \text{if } p \neq q \\ 0 & \text{if } p = q \text{ and } k > j \\ 1 & \text{if } p = q \text{ and } k \leq j. \end{cases}$$

First suppose that $p \neq q$. Then p^k and q^j are coprime, so $ap^k + bq^j = 1$ for some $a, b \in R$. If $x \in F_p^k(R/q^j)$ then $x = p^{k-1}y$ for some y and $px = 0$ so $p^k y = 0$. On the other hand, it is clear from the definition of R/q^j that $q^j z = 0$ for all $z \in R/q^j$, so $q^j y = 0$. We thus have $y = 1.y = ap^k y + bq^j y = 0$, and thus $x = p^{k-1}y = 0$. Thus $F_p^k(R/q^j) = 0$ and so $f_p^k(R/q^j) = 0$, as required.

Now suppose that $q = p$ and $j < k$. If $x \in R/p^j$ then $p^j x = 0$ and $k - 1 \geq j$ so $p^{k-1}x = 0$. Thus $p^{k-1}.(R/p^j) = 0$ and therefore $F_p^k(R/p^j) = 0$ so $f_p^k(R/p^j) = 0$.

Now suppose instead that $q = p$ and $k \leq j$. Put $e = p^{j-1} + p^j R \in R/p^j$, so clearly $pe = 0$. We also have $e = p^{k-1}e'$, where $e' = p^{j-k} + p^j R \in R/p^j$, so $e \in p^{k-1}(R/p^j)$, so $e \in F_p^k(R/p^j)$. If $\overline{a}$ is another element of $F_p^k(R/p^j)$ then $p\overline{a} = 0$ so pa is a multiple of p^j so a is a multiple of p^{j-1} so $\overline{a}$ is a multiple of e. As $e \neq 0$ and e spans $F_p^k(R/p^j)$ we conclude that $F_p^k(R/p^j)$ has dimension one, so $f_p^k(R/p^j) = 1$, as required.

It is now easy to deduce our description of $g_p^k(R/q^j)$. If $q \neq p$ then $f_p^k(R/q^j) = 0$ for all k and it follows easily that $g_p^k(R/q^j) = 0$. Suppose instead that $p = q$. If $k > j$ then $k + 1 > j$ as well so $f_p^k(R/p^j) = f_p^{k+1}(R/p^j) = 0$ so $g_p^k(R/p^j) = 0$ as claimed. If $k < j$ then both k and $k + 1$ are less than or equal to j, so $f_p^k(R/p^j) = f_p^{k+1}(R/p^j) = 1$ so $g_p^k(R/p^j) = 0$ as claimed. If $k = j$ then $f_p^k(R/p^j) = 1$ and $f_p^{k+1}(R/p^j) = 0$ so $g_p^k(R/p^j) = 1$ as claimed. □

Corollary 13.16. *Let M be an finite torsion module. Then M can be expressed in a unique way as a direct sum of basic modules. The number of copies of R/p^k in the direct sum is $g_p^k(M)$.*

Proof. We know from Corollary 13.10 that M can be written as $B_1 \oplus \ldots \oplus B_t$, where each B_i is a basic module. Let n_p^k be the number of copies of R/p^k in this list. We know that $g_p^k(M) = g_p^k(B_1) + \ldots + g_p^k(B_t)$. In this sum we get a 1 for every B_i that is a copy of R/p^k and a 0 for all other B_i's. Thus implies that $g_p^k(M) = n_p^k$.

Now suppose we have another splitting, say $M = C_1 \oplus \ldots \oplus C_s$ where each C_j is basic. Let m_p^k be the number of copies of R/p^k in this list. The same argument as before shows that $g_p^k(M) = m_p^k$, so $m_p^k = n_p^k$. Thus the lists $B_1, \ldots, B_t$ and $C_1, \ldots, C_s$ contain the same number of copies of R/p^k for all p and k, so the two lists must be the same up to reordering. Thus, M can be written in an essentially unique way as a direct sum of basic modules. □

14. Canonical forms for square matrices

Given any $n \times n$ matrix A over a field K we have a module M_A over $K[x]$. We see from Example 13.8 that M_A is a finite torsion module, and Corollary 13.16 gives a classification of such modules. In this section, we will see what this tells us about square matrices.

For simplicity, we will restrict attention to the case $K = \mathbb{C}$, where the irreducibles are easy to understand. As we saw in Example 10.6, the set

$$\mathcal{P} = \{x - \lambda \mid \lambda \in \mathbb{C}\}$$

is a complete set of irreducibles in $\mathbb{C}[x]$. The basic $\mathbb{C}[x]$ modules are thus the modules

$$B(\lambda, k) := B^k_{x-\lambda} = \mathbb{C}[x]/(x - \lambda)^k.$$

Thus Corollary 13.16 says that any finite torsion module over $\mathbb{C}[x]$ can be written essentially uniquely as a direct sum of $B(\lambda, k)$'s, in an essentially unique way.

The next proposition explains the most basic case of this. Recall that M_λ is the module whose elements are the complex numbers, with the multiplication rule $f.z = f(\lambda)z$.

Proposition 14.1. $B(\lambda, 1) \simeq M_\lambda$.

Proof. Define a map $\alpha\colon \mathbb{C}[x] \to M_\lambda$ by $\alpha(f) = f(\lambda)$. This is a $\mathbb{C}[x]$-module map because

$$\alpha(g.f) = (gf)(\lambda) = g(\lambda)f(\lambda) = g.f(\lambda) = g.\alpha(f).$$

If a is a constant polynomial then $\alpha(a) = a$, and this shows that α is surjective. We also have $\alpha(f) = 0$ iff $f(\lambda) = 0$ iff $f(x)$ is divisible by $x-\lambda$, so $\ker(\alpha)$ is the principal ideal $\mathbb{C}[x].(x-\lambda)$. Thus, the first isomorphism theorem gives us an isomorphism $\overline{\alpha}\colon B(\lambda, 1) = \mathbb{C}[x]/(x - \lambda) \to M_\lambda$. □

Now suppose that A is a diagonalizable $n \times n$ matrix over $\mathbb{C}$, with eigenvalues $\lambda_1, \dots, \lambda_n$ say. Then there exists a matrix P (whose columns are eigenvectors of A) such that $D := P^{-1}AP$ is a diagonal matrix, with entries $\lambda_1, \dots, \lambda_n$ on the diagonal. Recall from Example 5.16 that the direct sum of matrices is defined by

$$A \oplus B = \left(\begin{array}{c|c} A & 0 \\ \hline 0 & B \end{array}\right).$$

If we regard λ_i as a 1×1 matrix and use this notation, we find that $D = \lambda_1 \oplus \dots \oplus \lambda_n$.

We saw in Corollary 6.19 that $M_A \simeq M_{\lambda_1} \oplus \dots \oplus M_{\lambda_n}$, and we can now rewrite this as

$$M_A \simeq B(\lambda_1, 1) \oplus \dots \oplus B(\lambda_n, 1).$$

To extend this picture to nondiagonalizable matrices, we need the following definition.

Definition 14.2. Given $\lambda \in \mathbb{C}$ and $k > 0$ we let $J(\lambda, k)$ be the $k \times k$ matrix such that

1. Every entry on the diagonal is λ
2. Every entry just below the diagonal is 1
3. Every other entry is 0.

This is called a *Jordan block* of size k and eigenvalue λ. For example, we have

$$J(\lambda, 4) = \begin{pmatrix} \lambda & 0 & 0 & 0 \\ 1 & \lambda & 0 & 0 \\ 0 & 1 & \lambda & 0 \\ 0 & 0 & 1 & \lambda \end{pmatrix}.$$

Proposition 14.3. *$B(\lambda, k)$ is isomorphic to $M_{J(\lambda,k)}$.*

Proof. Put $y = x - \lambda \in \mathbb{C}[x]$, so that $B(\lambda, k) = \mathbb{C}[x]/y^k$. Put $A = J(\lambda, k) - \lambda I$, so for $v \in M_{J(\lambda,k)} = \mathbb{C}^k$ we have $y.v = J(\lambda, k)v - \lambda v = Av$. From the definition of $J(\lambda, k)$ we see that A has 1's just below the diagonal and 0's everywhere else. In the case $k = 4$ we have

$$y.v = Av = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{pmatrix} = \begin{pmatrix} 0 \\ v_1 \\ v_2 \\ v_3 \end{pmatrix}.$$

It is not hard to see that the general case follows the same pattern, so we have

$$y.(v_1, \dots, v_k) = (0, v_1, \dots, v_{k-1}).$$

It follows that $y^2.(v_1, \dots, v_k) = (0, 0, v_1, \dots, v_{k-2})$ and so on, so $y^{k-1}.(v_1, \dots, v_k) = (0, \dots, 0, v_1)$ and $y^k v = 0$.

Let $\{e_1, \dots, e_k\}$ be the usual basis for $M_{J(\lambda,k)} = \mathbb{C}^k$ over $\mathbb{C}$, so $y.e_i = e_{i+1}$ for $i < k$ and $y.e_k = 0$. We define $\alpha : \mathbb{C}[x] \to M_{J(\lambda,k)}$ by $\alpha(f) = f.e_1 = f(J(\lambda, k))e_1$. This is easily seen to be a map of $\mathbb{C}[x]$-modules.

I next claim that α is surjective. Indeed, for any $v = (v_1, \dots, v_k) \in M_{J(\lambda,k)}$ we can put $f = v_1 + v_2 y + \dots + v_k y^{k-1} \in \mathbb{C}[x]$ and we find that

$$\alpha(f) = f.e_1 = v_1 e_1 + v_2 y.e_1 + \dots + v_k y^{k-1}.e_k = v_1 e_1 + \dots + v_k e_k = v,$$

which proves surjectivity.

We next claim that $\ker(\alpha) = \mathbb{C}[x].y^k$. Indeed, suppose we have some polynomial $f(x)$ with $\alpha(f) = f.e_1 = 0$. By putting $x = y + \lambda$ and expanding everything out, we can write $f(x)$ in the form $a_0 + a_1 y + \dots + a_d y^d$. (For example, if $f(x) = x^2 + x + 1$ then $f(x) = (y+\lambda)^2 + (y+\lambda) + 1 = (1 + \lambda + \lambda^2) + (1 + 2\lambda)y + y^2$.) We then have

$$f.e_1 = a_0 e_1 + \dots + a_{k-1} y^{k-1}.e_1 = a_0 e_1 + \dots + a_{k-1} e_k = (a_0, a_1, \dots, a_{k-1}).$$

As $f.e_1 = 0$ we must have $a_0 = \dots = a_{k-1} = 0$ so $f(x) = a_k y^k + \dots + a_d y^d$, so $f(x)$ is divisible by y^k as required.

The first isomorphism theorem now tells us that

$$M_{J(\lambda,k)} = \operatorname{image}(\alpha) \simeq \mathbb{C}[x]/\ker(\alpha) = \mathbb{C}[x]/y^k = B(\lambda, k)$$

as claimed. □

Theorem 14.4. *Any square matrix A over $\mathbb{C}$ is conjugate to a matrix A' (called the* Jordan normal form *or* JNF *of A) that is a a direct sum of Jordan blocks.*

Proof. We know from Corollary 13.16 that M_A is isomorphic to a direct sum of modules of the form $B(\lambda, k)$, say

$$M_A \simeq B(\lambda_1, k_1) \oplus \dots \oplus B(\lambda_t, k_t).$$

Put $A' = J(\lambda_1, k_1) \oplus \dots \oplus J(\lambda_t, k_t)$, so

$$M_{A'} \simeq M_{J(\lambda_1,k_1)} \oplus \dots \oplus M_{J(\lambda_t,k_t)} \simeq B(\lambda_1, k_1) \oplus \dots \oplus B(\lambda_t, k_t) \simeq M_A.$$

As $M_A \simeq M_{A'}$, Proposition 6.18 tells us that A is conjugate to A', as claimed. □

The row-reduction algorithm described previously can be used write M_A as a direct sum of cyclic modules, or in other words modules of the form $\mathbb{C}[x]/f(x)$. If we factor $f(x)$ as $(x - \lambda_1)^{k_1} \dots (x - \lambda_r)^{k_r}$ (with all the λ's different) we see from Proposition 13.9 that

$$\mathbb{C}[x]/f(x) \simeq \mathbb{C}[x]/(x - \lambda_1)^{k_1} \oplus \dots \mathbb{C}[x]/(x - \lambda_r)^{k_r} = B(\lambda_1, k_1) \oplus \dots \oplus B(\lambda_r, k_r).$$

We can use this to give an explicit expression for M_A as a direct sum of $B(\lambda, k)$'s and thus to determine the JNF of A.

A different, and usually easier, approach is to use Corollary 13.16, which tells us that the number of copies of $B(\lambda, k)$ in the decomposition of M_A is just $g^k_{x-\lambda}(M_A)$, as defined in Definition 13.12. To calculate these numbers $g^k_{x-\lambda}(M_A)$, we need to recall the definition of the characteristic polynomial and minimal polynomial of a matrix.

Definition 14.5. The *characteristic polynomial* $\operatorname{char}(A)$ of a square matrix A is the polynomial $\det(tI - A)$.

Note that this is easy to calculate directly from A; for example, if

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}$$

then the characteristic polynomial is

$$\begin{vmatrix} t-1 & -2 & -3 \\ -4 & t-5 & -6 \\ -7 & -8 & t-9 \end{vmatrix} = (t-1)\begin{vmatrix} t-5 & -6 \\ -8 & t-9 \end{vmatrix} - (-2)\begin{vmatrix} -4 & -6 \\ -7 & t-9 \end{vmatrix} + (-3)\begin{vmatrix} -4 & t-5 \\ -7 & -8 \end{vmatrix}$$
$$= (t-1)((t-5)(t-9)-48) + 2(-4(t-9)-42) - 3(32+7(t-5))$$
$$= t^3 - 15t^2 - 18t.$$

It is also not hard to see that for any $n \times n$ matrix A we have $\mathrm{char}(A) = t^n +$ lower terms ; in other words, $\mathrm{char}(A)$ is a monic polynomial of degree n.

We now define the minimal polynomial of a square matrix A. First, we put $I = \{f \in \mathbb{C}[x] \mid f(A) = 0\}$. Clearly if $f(A) = g(A) = 0$ and h is arbitrary then $(f+g)(A) = f(A)+g(A) = 0$ and $(hf)(A) = h(A)f(A) = 0$, so I is an ideal. By Theorem 9.7, we see that $I = \mathbb{C}[x]g$ for some polynomial g. I next claim that I is never the zero ideal. Indeed, the set of all $n \times n$-matrices over $\mathbb{C}$ is a vector space over $\mathbb{C}$ of dimension n^2, so any list of n^2+1 such matrices must be linearly dependent. In particular, the list $I, A, \dots, A^{n^2}$ is linearly dependent, so there is some list $a_0, \dots, a_{n^2}$ (not all zero) such that $a_0 I + a_1 A + \dots + a_{n^2} A^{n^2} = 0$. If we put $f(x) = a_0 + a_1 x + \dots + a_{n_2} x^{n^2}$ we find that $f \neq 0$ and $f \in I$, so $I \neq 0$. As $I = \mathbb{C}[x]g$, we deduce that $g \neq 0$. After multiplying g by a nonzero constant, we may assume that g is a monic polynomial. This justifies the following definition:

Definition 14.6. The *minimal polynomial* $\min(A)$ of a square matrix A is the unique monic polynomial that generates the ideal $I_A = \{f \in \mathbb{C}[x] \mid f(A) = 0\}$.

Theorem 14.7. *Let A be an $n \times n$ matrix, with $\mathrm{char}(A) = \prod_{i=1}^{r}(t-\lambda_i)^{r_i}$, where all the λ_i's are distinct and $r_i > 0$. Then*

(a) *We have $\min(A) = \prod_{i=1}^{r}(t-\lambda_i)^{s_i}$, where $0 < s_i \leq r_i$. In other words, the roots of $\min(A)$ are precisely the same as the roots of $\mathrm{char}(A)$, and the multiplicities of the roots in $\min(A)$ are at most as large as the multiplicities in $\mathrm{char}(A)$.*

(b) *The JNF of A contains only blocks of the form $J(\lambda_i, k)$, where λ_i is a root of the characteristic polynomial, as before. The number of such blocks is $\dim(\ker(A-\lambda_i I)) = n - \mathrm{rank}(A-\lambda_i I)$. The maximum value of k that occurs with λ_i is precisely s_i, and the sum of all the k's that occur with λ_i is r_i.*

Remark 14.8. In particular, part (a) says that $\min(A)(t)$ divides $\mathrm{char}(A)(t)$, so $\mathrm{char}(A)(t) \in I_A$, so if we substitute the matrix A into its characteristic polynomial (in other words, evaluate $\mathrm{char}(A)(A)$) we get the zero matrix. This is the *Cayley-Hamilton Theorem*.

The proof will follow after some examples and preliminary results.

Example 14.9. Consider the following matrix over $\mathbb{C}$:

$$A = \begin{pmatrix} 2 & 3 & 4 \\ 0 & 2 & 3 \\ 0 & 0 & 2 \end{pmatrix}.$$

The characteristic polynomial is

$$\mathrm{char}(A) = \det \begin{pmatrix} t-2 & -3 & -4 \\ 0 & t-2 & -3 \\ 0 & 0 & t-2 \end{pmatrix} = (t-2)^3.$$

Here we have used the fact that if a matrix has zeros everywhere below the diagonal, then the determinant is just the product of the entries on the diagonal. It follows from the theorem that the minimal polynomial must be $(t-2)$ or $(t-2)^2$ or $(t-2)^3$. However, we find that

$$A - 2I = \begin{pmatrix} 0 & 3 & 4 \\ 0 & 0 & 3 \\ 0 & 0 & 0 \end{pmatrix} \qquad (A-2I)^2 = \begin{pmatrix} 0 & 0 & 9 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

If the minimal polynomial is $f(t)$, we must have $f(A) = 0$. Thus, the above shows that $\min(A) \neq t-2$ and $\min(A) \neq (t-2)^2$, so $\min(A)$ must be $(t-2)^3$.

We also see that the JNF of A can only contain blocks of the form $J(2,k)$, so it must have the form $J(2,k_1) \oplus \ldots \oplus J(2,k_r)$, and we may as well order the factors so that $0 < k_1 \leq \ldots \leq k_r$. By comparing characteristic polynomials we see that $(t-2)^3 = \operatorname{char}(A) = (t-2)^{k_1+\ldots+k_r}$. By comparing minimal polynomials (and noting that $\max(k_1, \ldots, k_r) = k_r$) we see that $(t-2)^3 = \min(A) = (t-2)^{k_r}$. Thus $3 = k_1 + \ldots + k_r = k_r$. As all the k_i's are supposed to be positive, this can only work if $r = 1$ and $k_1 = 3$. Thus the JNF of A is just the single Jordan block $J(2,3)$, and thus $M_A \simeq B(2,3) = \mathbb{C}[x]/(x-2)^3$.

For an alternative approach, we can observe from our formula for $A - 2I$ that $\operatorname{rank}(A-2I) = 2$ and so $\dim(\ker(A-2I)) = 3-2 = 1$. Part (b) of Theorem 14.7 therefore tells us that there is only one block in the JNF, so A is conjugate to $J(2,k)$ for some k. As A is a 3×3 matrix, we must have $k = 3$.

Example 14.10. Consider the following matrix over $\mathbb{C}$:

$$A = \begin{pmatrix} 2 & i & i & 2 \\ i & 0 & 0 & i \\ i & 0 & 0 & i \\ 2 & i & i & 2 \end{pmatrix}.$$

The characteristic polynomial is the determinant of the first matrix shown below:

$$\begin{pmatrix} t-2 & -i & -i & -2 \\ -i & t & 0 & -i \\ -i & 0 & t & -i \\ -2 & -i & -i & t-2 \end{pmatrix} \to \begin{pmatrix} t & 0 & 0 & -t \\ 0 & t & -t & 0 \\ -i & 0 & t & -i \\ -2 & -i & -i & t-2 \end{pmatrix} \to \begin{pmatrix} t & 0 & 0 & 0 \\ 0 & t & 0 & 0 \\ -i & 0 & t & -2i \\ -2 & -i & -2i & t-4 \end{pmatrix}.$$

It is perfectly possible to evaluate the determinant directly, but it is more efficient to perform some row and column operations first, as shown above. In the first step we have subtracted the fourth row from the first row and the third row from the second row. In the second step we have added the first column to the fourth column and the second column to the third column. None of these operations change the determinant. The determinant of our final matrix is t^2 times the determinant of the 2×2 block in the bottom right corner, which is $t^2 - 4t + 4$. Thus $\det(tI - A) = t^2(t^2 - 4t + 4) = t^2(t-2)^2$. The factor $(t-2)^2$ comes from some Jordan blocks of the form $J(2,k)$, each of which contributes a factor $(t-2)^k$. The only possibilities are $J(2,2)$ and $J(2,1) \oplus J(2,1)$. Similarly, the factor t^2 comes from $J(0,2)$ or $J(0,1) \oplus J(0,1)$. This gives four possibilities for the JNF of A; we list these below, together with their minimal polynomials.

$$\begin{array}{rl} J(0,1) \oplus J(0,1) \oplus J(2,1) \oplus J(2,1) & t(t-2) \\ J(0,2) \oplus J(2,1) \oplus J(2,1) & t^2(t-2) \\ J(0,1) \oplus J(0,1) \oplus J(2,2) & t(t-2)^2 \\ J(0,2) \oplus J(2,2) & t^2(t-2)^2. \end{array}$$

The minimal polynomial of A must be one of the polynomials in the list above. By direct calculation, we find that

$$A(A-2I) = 2\begin{pmatrix} 1 & i & i & 1 \\ i & -1 & -1 & i \\ i & -1 & -1 & i \\ 1 & i & i & 1 \end{pmatrix}$$

$$A^2(A-2I) = 4\begin{pmatrix} 1 & i & i & 1 \\ i & -1 & -1 & i \\ i & -1 & -1 & i \\ 1 & i & i & 1 \end{pmatrix}$$

$$A(A-2I)^2 = A^2(A-2I)^2 = 0.$$

It follows that the minimal polynomial of A is $t(t-2)^2$. By comparing this with the list, we deduce that the JNF of A must be $J(0,1) \oplus J(0,1) \oplus J(2,2)$.

Example 14.11. Consider the following matrix over $\mathbb{C}$:

$$A = \begin{pmatrix} 0 & 0 & 1 & 1 \\ -1 & 1 & 1 & 0 \\ -1 & 0 & 2 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Here we will just calculate the characteristic polynomial directly by expanding along the top row, although more efficient methods are certainly possible. We have

$$\begin{vmatrix} t & 0 & -1 & -1 \\ 1 & t-1 & -1 & 0 \\ 1 & 0 & t-2 & -1 \\ 0 & 0 & 0 & t-1 \end{vmatrix} = t \begin{vmatrix} t-1 & -1 & 0 \\ 0 & t-2 & -1 \\ 0 & 0 & t-1 \end{vmatrix} - 0 \begin{vmatrix} 1 & -1 & 0 \\ 1 & t-2 & -1 \\ 0 & 0 & t-1 \end{vmatrix} +$$

$$(-1) \begin{vmatrix} 1 & t-1 & 0 \\ 1 & 0 & -1 \\ 0 & 0 & t-1 \end{vmatrix} - (-1) \begin{vmatrix} 1 & t-1 & -1 \\ 1 & 0 & t-2 \\ 0 & 0 & 0 \end{vmatrix}$$

$$\begin{vmatrix} t-1 & -1 & 0 \\ 0 & t-2 & -1 \\ 0 & 0 & t-1 \end{vmatrix} = (t-1)^2(t-2)$$

$$\begin{vmatrix} 1 & t-1 & 0 \\ 1 & 0 & -1 \\ 0 & 0 & t-1 \end{vmatrix} = -(t-1)^2$$

$$\begin{vmatrix} 1 & t-1 & -1 \\ 1 & 0 & t-2 \\ 0 & 0 & 0 \end{vmatrix} = 0$$

so

$$\begin{aligned} \operatorname{char}(A) &= t(t-1)^2(t-2) - 0 + (-1)(-(t-1)^2) - 0 \\ &= (t-1)^2(t(t-2)+1) \\ &= (t-1)^2(t^2-2t+1) = (t-1)^4. \end{aligned}$$

It follows that the minimal polynomial is $(t-1)^k$ for some k with $1 \leq k \leq 4$. We have

$$(A-I) = \begin{pmatrix} -1 & 0 & 1 & 1 \\ -1 & 0 & 1 & 0 \\ -1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

$$(A-I)^2 = 0.$$

It follows that the minimal polynomial must be $(t-1)^2$. We also see that $\operatorname{rank}(A-I) = 2$, so $\dim(\ker(A-I)) = 4-2 = 2$, so there are 2 blocks in the JNF. The JNF is thus $J(1,k_1) \oplus J(1,k_2)$ for some k_1, k_2 with $k_1 \leq k_2$. By comparing characteristic polynomials we find that $k_1 + k_2 = 2$, and by comparing minimal polynomials we find that $k_2 = \max(k_1, k_2) = 2$. It follows that $k_1 = k_2 = 2$, so the JNF is $J(1,2) \oplus J(1,2)$.

Proposition 14.12. *If A is conjugate to B, then* $\operatorname{char}(A) = \operatorname{char}(B)$ *and* $\min(A) = \min(B)$ *and* $\dim(\ker(A-\lambda I)) = \dim(\ker(B-\lambda I))$ *for all* $\lambda \in \mathbb{C}$.

Proof. As A and B are conjugate, we have $A = PBP^{-1}$ for some invertible matrix P. It follows that $P(tI-B)P^{-1} = tPP^{-1} - PBP^{-1} = tI - A$ so

$$\det(tI-A) = \det(P(tI-B)P^{-1}) = \det(P)\det(tI-B)\det(P)^{-1} = \det(tI-B),$$

so $\operatorname{char}(A) = \operatorname{char}(B)$ as claimed.

Next, we have $f(A) = 0$ iff $f.m = 0$ for all $m \in M_A$, and $f(B) = 0$ iff $f.m = 0$ for all $m \in M_B$. As A is conjugate to B the modules M_A and M_B are isomorphic, so $f(A) = 0$ iff $f(B) = 0$. (More directly, one can check that $f(A) = Pf(B)P^{-1}$, and again it follows that $f(A) = 0$ iff $f(B) = 0$.)

Thus, the ideals I_A and I_B are the same, so they have the same monic generator, in other words $\min(A) = \min(B)$.

Finally, note that $\ker(A - \lambda I) = \{m \in M_A \mid (x - \lambda)m = 0\}$ and $\ker(B - \lambda I) = \{m \in M_B \mid (x - \lambda)m = 0\}$. As $M_A \simeq M_B$ we see that these two vector spaces are isomorphic and thus have the same dimension, as claimed. □

Proposition 14.13. *Let A and B be square matrices of sizes n and m. Then* $\operatorname{char}(A \oplus B) = \operatorname{char}(A)\operatorname{char}(B)$, *and* $\min(A \oplus B)$ *is the least common multiple of* $\min(A)$ *and* $\min(B)$. *Moreover, for any $\lambda \in \mathbb{C}$ we have*

$$\dim\ker(A \oplus B - \lambda I_{n+m}) = \dim\ker(A - \lambda I_n) + \dim\ker(B - \lambda I_m).$$

Proof. We first claim that $\det(A \oplus B) = \det(A)\det(B)$. We have

$$A \oplus B = \left(\begin{array}{c|c} A & 0 \\ \hline 0 & B \end{array}\right) = \left(\begin{array}{c|c} A & 0 \\ \hline 0 & I_m \end{array}\right)\left(\begin{array}{c|c} I_n & 0 \\ \hline 0 & B \end{array}\right) = (A \oplus I_m)(I_n \oplus B),$$

so $\det(A \oplus B) = \det(A \oplus I_m)\det(I_n \oplus B)$. By expanding along the top row we find that $\det(I_n \oplus B) = \det(I_{n-1} \oplus B)$, and it follows inductively that $\det(I_n \oplus B) = \det(B)$ for all n. Similarly, by expanding along the bottom row we see that $\det(A \oplus I_m) = \det(A)$ for all m, so the equation $\det(A \oplus B) = \det(A \oplus I_m)\det(I_n \oplus B)$ gives $\det(A \oplus B) = \det(A)\det(B)$ as claimed.

Next, observe that $tI_{n+m} - A \oplus B = (tI_n - A) \oplus (tI_m - B)$, so $\det(tI_{n+m} - A \oplus B) = \det(tI_n - A)\det(tI_m - B) = \operatorname{char}(A)\operatorname{char}(B)$ as claimed.

Next, note that $f(A \oplus B) = f(A) \oplus f(B)$. Thus

$$\begin{aligned} f \text{ is divisible by } \min(A \oplus B) &\Leftrightarrow f(A \oplus B) = 0 \\ &\Leftrightarrow f(A) = 0 \text{ and } f(B) = 0 \\ &\Leftrightarrow f \text{ is divisible by both } \min(A) \text{ and } \min(B). \end{aligned}$$

This means that $\min(A \oplus B)$ is the least common multiple of $\min(A)$ and $\min(B)$.

Finally, we can regard $\mathbb{C}^{n+m}$ as $\mathbb{C}^n \oplus \mathbb{C}^m$ and we then have $(A \oplus B).(u, v) = (Au, Bv)$. It follows that $(A \oplus B - \lambda I).(u, v) = (Au - \lambda u, Bv - \lambda v)$, and thus that $\ker(A \oplus B - \lambda I)$ is the set of pairs (u, v) for which $(A - \lambda I)u = 0$ and $(B - \lambda I)v = 0$. In other words we have $\ker(A \oplus B - \lambda I) = \ker(A - \lambda I) \oplus \ker(B - \lambda I)$ and so $\dim\ker(A \oplus B - \lambda I) = \dim\ker(A - \lambda I) + \dim\ker(B - \lambda I)$. □

Proposition 14.14. *We have* $\operatorname{char}(J(\lambda, k)) = \min(J(\lambda, k)) = (t - \lambda)^k$. *Moreover, we have*

$$\dim\ker(J(\lambda, k) - \mu I) = \begin{cases} 1 & \text{if } \lambda = \mu \\ 0 & \text{if } \lambda \neq \mu. \end{cases}$$

Proof. I first claim that $\det(J(\lambda, k)) = \lambda^k$. To see this, we need to recall the usual row-expansion method for evaluating determinants. Suppose we have a $k \times k$ matrix A with entries $a_1, \ldots, a_k$ on the top row, and that A_i is obtained from i by deleting the top row and the i'th column; then

$$\det(A) = a_1 \det(A_1) - a_2 \det(A_2) + \ldots \pm a_k \det(A_k).$$

If we take $A = J(\lambda, k)$ then $a_1 = \lambda$, $A_1 = J(\lambda, k-1)$ and $a_2 = \ldots = a_k = 0$, so $\det(J(\lambda, k)) = \lambda \det(J(\lambda, k-1))$. Moreover, $J(\lambda, 1)$ is the 1×1 matrix (λ) so $\det(J(\lambda, 1)) = \lambda$. It follows inductively that $\det(J(\lambda, k)) = \lambda^k$ for all $k > 0$, as claimed. Here we illustrate the case $k = 5$:

$$J(\lambda, 5) = \left(\begin{array}{c|cccc} \lambda & 0 & 0 & 0 & 0 \\ \hline 1 & \lambda & 0 & 0 & 0 \\ 0 & 1 & \lambda & 0 & 0 \\ 0 & 0 & 1 & \lambda & 0 \\ 0 & 0 & 0 & 1 & \lambda \end{array}\right) = \left(\begin{array}{c|c} \lambda & 0 \\ \hline * & J(\lambda, 4) \end{array}\right).$$

Next, note that $J(\lambda, k) - tI = J(\lambda - t, k)$, so $\det(J(\lambda, k) - tI) = (\lambda - t)^k$. For a $k \times k$ matrix we have $\det(-A) = (-1)^k \det(A)$, so $\det(tI - J(\lambda, k)) = (-1)^k(\lambda - t)^k = (t - \lambda)^k$ as claimed.

We next need to understand which polynomials $f(x)$ have $f(J(\lambda, k)) = 0$. This happens iff $f.m = 0$ for all m in the module $M_{J(\lambda,k)}$, which is isomorphic to $B(\lambda, k) = \mathbb{C}[x]/(x - \lambda)^k$. It is clear that $f.B(\lambda, k) = \{0\}$ iff f is divisible by $(x - \lambda)^k$, so $\min(J(\lambda, k)) = (x - \lambda)^k$ as well.

Finally, note that $J(\lambda,k)-\mu I = J(\lambda-\mu,k)$. If $\mu \neq \lambda$ we deduce that $\det(J(\lambda,k)-\mu I) = (\lambda-\mu)^k \neq 0$, so $J(\lambda,k)-\mu I$ is invertible and $\ker(J(\lambda,k)-\mu I) = \{0\}$. On the other hand, if $\mu = \lambda$ then $J(\lambda,k)-\mu I = J(0,k)$, and it is not hard to see that $J(0,k).(x_1,\dots,x_k) = (0,x_1,\dots,x_{k-1})$. Thus $J(0,k).\underline{x} = 0$ iff $x_1 = \dots = x_{k-1} = 0$, so $\underline{x} = (0,\dots,0,x_k)$ for some $x_k \in \mathbb{C}$. It follows that $\ker(J(\lambda,k)-\lambda I)$ is the span of the standard basis vector e_k, and so $\dim\ker(J(\lambda,k)-\lambda I) = 1$. □

Proof of Theorem 14.7. Let A' be the JNF of A, so A' is conjugate to A and is a block sum of matrices of the form $J(\lambda,k)$ for various λ's and k's. We know from Proposition 14.12 that $\mathrm{char}(A) = \mathrm{char}(A')$ and $\min(A) = \min(A')$ and $\dim\ker(A-\lambda I) = \dim\ker(A'-\lambda I)$ for all λ.

First suppose that all the Jordan blocks in A' have the same eigenvalue λ. Then we can write

$$A' = J(\lambda,k_1) \oplus \dots \oplus J(\lambda,k_d)$$

for some sequence $k_1,\dots,k_d$ of positive integers. We then have

$$\mathrm{char}(A') = \prod_i \mathrm{char}(J(\lambda,k_i)) = \prod_i (t-\lambda)^{k_i} = (t-\lambda)^k,$$

where $r = \sum_i k_i$. We also see that $\min(A')$ is the least common multiple of the polynomials $(t-\lambda)^{k_i}$, which is $(t-\lambda)^s$, where $s = \max(k_1,\dots,k_d)$. As each k_i is positive this means that $0 < s \leq r$. Finally, we see that

$$\begin{aligned}\dim\ker(A-\lambda I) &= \dim\ker(A'-\lambda I)\\ &= \sum_{i=1}^{d} \dim\ker(J(\lambda,k_i)-\lambda I)\\ &= \sum_{i=1}^{d} 1 = d.\end{aligned}$$

A similar argument shows that $\dim\ker(A-\mu I) = 0$ for $\mu \neq \lambda$.

More generally, there will be a number of different eigenvalues, say $\lambda_1,\dots,\lambda_d$. Let A'_i be the block sum of all the terms of the form $J(\lambda_i,k)$ for some k. The above argument shows that $\mathrm{char}(A'_i) = (t-\lambda_i)^{r_i}$ and $\min(A'_i) = (t-\lambda_i)^{s_i}$ for some integers r_i, s_i with $0 < s_i \leq r_i$. We also have $A' = A'_1 \oplus \dots \oplus A'_d$, so

$$\mathrm{char}(A) = \mathrm{char}(A') = \prod_i (t-\lambda_i)^{r_i}.$$

Similarly, we find that $\min(A)$ is the least common multiple of the polynomials $(t-\lambda_1)^{s_1},\dots,(t-\lambda_d)^{s_d}$. As these polynomials are all powers of inequivalent irreducibles, their lcm is just their product, so

$$\min(A) = \prod_i (t-\lambda_i)^{s_i}.$$

We also see that $\dim\ker(A-\lambda_i) = \sum_j \dim\ker(A_j-\lambda_i)$. The terms for $j \neq i$ are zero, and the term for $j = i$ is just the number of Jordan blocks in A_i.

The theorem now follows immediately. □

We conclude by studying the structure of cyclic modules over $\mathbb{C}[x]$.

Let $f(x)$ be a monic polynomial of degree n over $\mathbb{C}$. We then have a cyclic module $\mathbb{C}[x]/f(x)$ over $\mathbb{C}[x]$. For any polynomial $g(x) \in \mathbb{C}[x]$ we have an element $\overline{g(x)} = g(x)+\mathbb{C}[x]f(x) \in \mathbb{C}[x]/f(x)$, and $\overline{g(x)} = \overline{h(x)}$ iff $g(x)-h(x)$ is divisible by $f(x)$.

Proposition 14.15. *If f is as above then the elements $\overline{1},\overline{x},\dots,\overline{x}^{n-1}$ form a basis for $\mathbb{C}[x]/f(x)$ over $\mathbb{C}$. In particular, we have $\dim_{\mathbb{C}}(\mathbb{C}[x]/f(x)) = n$.*

Proof. Every element of $\mathbb{C}[x]/f(x)$ can be written as $\overline{g(x)}$ for some $g(x) \in \mathbb{C}[x]$. We can divide g by f to get $g(x) = f(x)q(x)+r(x)$ for some polynomial $r(x)$ of degree at most $n-1$. It follows that $g(x)-r(x)$ is divisible by $f(x)$, so $\overline{g(x)} = \overline{r(x)}$. As $\deg(r) < n$ we have $r(x) =$

$a_0 + a_1x + \ldots + a_{n-1}x^{n-1}$ for some $a_0, \ldots, a_{n-1} \in \mathbb{C}$. It follows that $\overline{r(x)} = \sum_{i=0}^{n-1} a_i \overline{x}^i$, so the elements $\overline{1}, \overline{x}, \ldots, \overline{x}^{n-1}$ span $\mathbb{C}[x]/f(x)$ over $\mathbb{C}$.

Now suppose we have a linear relation among these elements, say $b_0\overline{1} + \ldots + b_{n-1}\overline{x}^{n-1} = \overline{0}$. This means that the polynomial $g(x) := b_0 + b_1x + \ldots + b_{n-1}x^{n-1}$ satisfies $\overline{g(x)} = \overline{0}$, so g is divisible by f. As the degree of g is less than the degree of f,this can only happen if $g = 0$, which means that $b_0 = \ldots = b_{n-1} = 0$. Thus, the elements $\overline{1}, \ldots, \overline{x}^{n-1}$ are linearly independent over $\mathbb{C}$. □

Theorem 14.16. *Let A be an $n \times n$ matrix over $\mathbb{C}$, with JNF $J(\lambda_1, k_1) \oplus \ldots \oplus J(\lambda_r, k_r)$. Then the following statements are all equivalent (so if any one of them is true, then all of them are true):*

(a) *M_A is a cyclic module over $\mathbb{C}[x]$.*
(b) *The numbers λ_i are all different.*
(c) $\min(A) = \operatorname{char}(A)$.
(d) $M_A \simeq \mathbb{C}[x]/\operatorname{char}(A)(x)$.
(e) *There is a vector $v \in M_A$ such that $\{v, Av, \ldots, A^{n-1}v\}$ is a basis for M_A over $\mathbb{C}$.*

Proof. (a)⇒(b): If M_A is cyclic then $M_A \simeq \mathbb{C}[x]/f(x)$ for some polynomial $f(x)$. This must be nonzero, otherwise M_A would be the same as $\mathbb{C}[x]$ and thus would have infinite dimension over $\mathbb{C}$, which is impossible because $\dim_{\mathbb{C}}(M_A) = n$. We can factor $f(x)$ as $c(x-\mu_1)^{k_1} \ldots (x-\mu_s)^{k_s}$ for some nonzero constant c and numbers $\mu_1, \ldots, \mu_s$. By collecting terms we may assume that the μ_i's are all different. Using the Chinese Remainder Theorem we find that

$$M_A \simeq \mathbb{C}[x]/f(x) = \bigoplus_i \mathbb{C}[x]/(x-\mu_i)^{k_i} = \bigoplus_i B(\mu_i, k_i),$$

so the JNF of A is $J(\mu_1, k_1) \oplus \ldots \oplus J(\mu_s, k_s)$. Thus the μ's are the same as the λ's (up to possible reordering) and so the λ_i's are all different.

(b)⇒(a): Conversely, suppose that the λ_i's are all different. Then the polynomials $(x-\lambda_i)^{k_i}$ are all coprime to each other, so if we put $f(x) = \prod_i (x-\lambda_i)^{k_i}$ the Chinese Remainder Theorem tells us that

$$\mathbb{C}[x]/f(x) \simeq \bigoplus_i \mathbb{C}[x]/(x-\lambda_i)^{k_i} = \bigoplus_i B(\lambda_i, k_i) \simeq M_A,$$

so M_A is cyclic.

(b)⇔(c): Note that A is conjugate to $J(\lambda_1, k_1) \oplus \ldots \oplus J(\lambda_r, k_r)$. We know from Propositions 14.12, 14.13 and 14.14 that $\operatorname{char}(A)$ is the product of the polynomials $(x-\lambda_i)^{k_i}$, and that $\min(A)$ is their least common multiple. The product is the same as the least common multiple iff the factors $(x-\lambda_i)^{k_i}$ are all coprime to each other, which is true iff the numbers λ_i are all different.

(a)⇔(d): If (d) holds (ie $M_A \simeq \mathbb{C}[x]/\operatorname{char}(A)(x)$) then M_A has the form $\mathbb{C}[x]/f(x)$ and is certainly cyclic.

Conversely, suppose that (a) holds, so $M_A \simeq \mathbb{C}[x]/f(x)$ for some f. Just as in the proof that (a)⇒(b) we see that f is a unit multiple of $\prod_i (x-\lambda_i)^{k_i}$, which we know is the same as $\operatorname{char}(A)(x)$. It follows that $M_A \simeq \mathbb{C}[x]/\operatorname{char}(A)(x)$, so (d) holds.

(e)⇒(a): Suppose that $v \in M_A$ and that $\{v, Av, \ldots, A^{n-1}v\}$ is a basis for M_A over $\mathbb{C}$. Then for any $w \in M_A$ there exist $a_0, \ldots, a_{n-1} \in \mathbb{C}$ such that $w = a_0v + a_1Av + \ldots + a_{n-1}A^{n-1}v$. Thus, if we put $g(x) = \sum_i a_i x^i \in \mathbb{C}[x]$ we find that $w = g.v$. It follows that v generates M_A as a $\mathbb{C}[x]$-module, so M_A is cyclic.

(a)⇒(e): Suppose that M_A is cyclic, so we can choose an isomorphism $\alpha\colon \mathbb{C}[x]/f(x) \to M_A$. It follows that $\deg(f) = \dim(\mathbb{C}[x]/f(x)) = \dim(M_A) = n$. It follows from Proposition 14.15 that $\{\overline{1}, \ldots, \overline{x}^{n-1}\}$ is a basis for $\mathbb{C}[x]/f(x)$, and thus that $\{\alpha(\overline{1}), \ldots, \alpha(\overline{x}^{n-1})\}$ is a basis for M_A. If we put $v = \alpha(\overline{1})$ we find that $\alpha(\overline{x}^k) = \alpha(x^k.\overline{1}) = x^k.\alpha(\overline{1} = A^kv$. This means that our basis is just $\{v, Av, \ldots, A^{n-1}v\}$, as required. □

Example 14.17. Consider the following matrices:

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

We find that $\mathrm{char}(A) = \mathrm{char}(B) = (t-1)^4$. Moreover $(A-I)^3 \neq 0$, so $\min(A) = (t-)^4 = \mathrm{char}(A)$, so M_A is cyclic. On the other hand, $(B-I)^2 = 0$ (and $B - I \neq 0$) so $\min(B) = (t-1)^2 \neq \mathrm{char}(B)$, so M_B is not cyclic.

Example 14.18. Consider the following matrix:

$$A = \begin{pmatrix} 0 & 0 & 0 & -3 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ 3 & 0 & 0 & 0 \end{pmatrix}.$$

We find that

$$\mathrm{char}(A) = 9 + 10x^2 + x^4 = (x^2+1)(x^2+9) = (x+i)(x-i)(x+3i)(x-3i).$$

As the four roots are distinct, we see that the minimal polynomial is the same as the characteristic polynomial, so M_A is cyclic. If we put $v = (1,1,1,1) \in \mathbb{C}^4$ we find that

$$\begin{aligned} v &= (1,1,1,1) \\ Av &= (-3,-1,1,3) \\ A^2v &= (-9,-1,-1,-9) \\ A^3v &= (27,1,-1,-27). \end{aligned}$$

By standard methods we can check that these vectors are linearly independent and thus form a basis of $\mathbb{C}^4$. This gives another proof that M_A is cyclic.