A3: Rings and Modules, 2020–2021

Tom Sanders

We begin with the course overview as described on `https://courses.maths.ox.ac.uk/node/44027`.

Course Overview:

The first abstract algebraic objects which are normally studied are groups, which arise naturally from the study of symmetries. The focus of this course is on rings, which generalise the kind of algebraic structure possessed by the integers: a ring has two operations, addition and multiplication, which interact in the usual way. The course begins by studying the fundamental concepts of rings (already met briefly in core Algebra): what are maps between them, when are two rings isomorphic *etc.* much as was done for groups. As an application, we get a general procedure for building fields, generalising the way one constructs the complex numbers from the reals. We then begin to study the question of factorization in rings, and find a class of rings, known as Unique Factorization Domains, where any element can be written uniquely as a product of prime elements generalising the case of the integers. Finally, we study modules, which roughly means we study linear algebra over certain rings rather than fields. This turns out to have powerful applications to ordinary linear algebra and to abelian groups.

Learning Outcomes:

Students should become familiar with rings and fields, and understand the structure theory of modules over a Euclidean domain along with its implications. The material underpins many later courses in algebra and number theory, and thus should give students a good background for studying these more advanced topics.

Course Synopsis:

Recap on rings (not necessarily commutative) and examples: $\mathbb{Z}$, fields, polynomial rings (in more than one variable), matrix rings. Zero-divisors, integral domains. Units. The

Last updated: 17th February, 2021.

characteristic of a ring. Discussion of fields of fractions and their characterization (proofs non-examinable) [2]

Homomorphisms of rings. Quotient rings, ideals and the first isomorphism theorem and consequences, e.g. Chinese remainder theorem. Relation between ideals in R and R/I. Prime ideals and maximal ideals, relation to fields and integral domains. Examples of ideals. Application of quotients to constructing fields by adjunction of elements; examples to include $\mathbb{C} = \mathbb{R}[X]/\langle X^2 + 1\rangle$ and some finite fields. Degree of a field extension, the tower law. [4]

Euclidean Domains. Examples. Principal Ideal Domains. EDs are PIDs. Unique factorisation for PIDs. Gauss's Lemma and Eisenstein's Criterion for irreducibility. [3]

Modules: Definition and examples: vector spaces, abelian groups, vector spaces with an endomorphism. Submodules and quotient modules and direct sums. The first isomorphism theorem. [2]

Row and column operations on matrices over a ring. Equivalence of matrices. Smith Normal form of matrices over a Euclidean Domain. [1.5]

Free modules and presentations of finitely generated modules. Structure of finitely generated modules of a Euclidean domain. [2]

Application to rational canonical form and Jordan normal form for matrices, and structure of finitely generated Abelian groups. [1.5]

References

There is an alternative approach to the course given in Earl's notes [Ear19] which is an excellent source for further examples.

> Forest green text denotes material which is unlectured background from previous courses.
> Blue text denotes material which is unlectured and more advanced.

1 Rings

A set R equipped with two binary operations $+$ and $\times$ is a **ring** if

- R equipped with $+$ is a commutative group called the **additive group**;
- $\times$ is an associative binary operation on R with an identity[1];
- $\times$ is distributive[2] over $+$.

We call $+$ the **addition** of the ring. Identities of binary operations are unique when they exist, and so we can unambiguously write 0 for the identity of addition – it is called the **zero** of the ring. Inverses[3] for elements w.r.t. associative binary operations are unique when they exist and so we can unambiguously write $-x$ for the additive inverse of $x \in R$; the map $R \to R; x \mapsto -x$ is called **negation** and$-(-x) = x$ for all $x \in R$, and $-0 = 0$ since an identity for a binary operation is always an inverse for itself. We write $x - y$ for $x + (-y)$.

We call $\times$ the **multiplication** of the ring and write xy in place of $x \times y$. Again, we can unambiguously write 1 for the identity of multiplication. Not all elements of R need have a multiplicative inverse; those that do are called **units** and we write[4] $U(R)$ for the set of units. Again, if $x \in U(R)$ we can unambiguously write x^{-1} for the multiplicative inverse of x, $(x^{-1})^{-1} = x$ for all $x \in U(R)$, and $1 \in U(R)$ with $1^{-1} = 1$.

Occasionally we shall have multiple rings and it will be instructive to clarify which particular ring we are referring to. We shall do this with subscripts writing, for example, $+_R$, $\times_R$, 0_R and 1_R in reference to the addition, multiplication, zero, and multiplicative identity of a ring R.

We say R is a **commutative** ring if the multiplication is commutative. ⚠The modern notion of commutative ring can be traced back to Emmy Noether [Noe21, §1] (translated into English in [Ber14]), though her definition does *not* assume the multiplication has an identity.

Proposition 1.1 (Group of units)**.** *Suppose that R is a ring. Then multiplication on R restricts to a well-defined binary operation on $U(R)$ giving it the structure of a group with identity 1, and if $x \in U(R)$ then $x^{-1} \in U(R)$ and it is the inverse of x with respect to this group operation on $U(R)$. Furthermore, if R is commutative then so is the group $U(R)$.*

Proof. First, suppose that $x, y \in U(R)$. Then $(xy)(y^{-1}x^{-1}) = x((yy^{-1})x^{-1}) = xx^{-1} = 1$ and similarly $(y^{-1}x^{-1})(xy) = 1$ so that $xy \in U(R)$. Hence multiplication on R restricts to a well-defined binary operation on $U(R)$. Since multiplication is associative on R, it is *a fortiori*

[1] e is an **identity** for a binary operation $*$ on a set X if $x * e = x = e * x$ for all $x \in X$.

[2] Meaning $x \times (y + z) = (x \times y) + (x \times z)$ *and* $(x + y) \times z = (x \times z) + (y \times z)$ for all $x, y, z \in R$.

[3] y is an **inverse** for x w.r.t. a binary operation $*$ on X if $x * y = y * x$ is an (and so *the*) identity for $*$.

[4] ⚠Some authors (*e.g.* [Lan02, p84] and [Lam07, xiv]) write R^* for $U(R)$.

associative when restricted to $U(R)$. Since $1 \in U(R)$ is an identity for multiplication on R it is *a fortiori* an identity for multiplication restricted to $U(R)$. Finally, if $x \in U(R)$ then $xx^{-1} = 1 = x^{-1}x$ and so $x^{-1} \in U(R)$ (with inverse x) and so every $x \in U(R)$ is invertible w.r.t. multiplication on R restricted to $U(R)$, and its inverse is the same as its inverse in R. Finally, if R is commutative then multiplication is commutative on R and *a fortiori* it is commutative when restricted to $U(R)$. The result is proved. □

Remark 1.2. If R is a finite commutative ring then $U(R)$ is a finite commutative group, but exactly which finite commutative groups occur as the group of units of a ring is an open problem called Fuchs' problem [Fuc58, Problem 72, p299].

To say that multiplication is distributive over addition is exactly to say that the left and right multiplication maps[5] are homomorphisms of the additive group. Group homomorphisms preserve identities and hence inverses; put another way:

Lemma 1.3. *Suppose that R is a ring.*

(i) (Zero annihilates) $x0 = 0x = 0$ *for all* $x \in R$*;*

(ii) (Negation distributes) $x(-y) = (-x)y = -(xy)$ *for all* $x, y \in R$*.*

Remark 1.4. Suppose that R is a ring and $z \in U(R)$. Then $(-z)(-z^{-1}) = (-(-z))(z^{-1}) = zz^{-1} = 1$ and similarly $(-z^{-1})(-z) = 1$, whence $-z \in U(R)$ and $(-z)^{-1} = -z^{-1}$. In particular, since $1 \in U(R)$ we have $-1 \in U(R)$.

We write[4] R^* for the set of non-zero elements of a ring R.

Remark 1.5. It is almost always the case that $U(R) \subset R^*$. Indeed, in view of Lemma 1.3 (i), 0 cannot have a multiplicative inverse unless $0 = 1$ (and of course if $0 = 1$ then 0 *does* have a multiplicative inverse – it is 0) so that $0 \in U(R)$ if and only if $0 = 1$.

If $0 = 1$ in a ring R we call it **trivial**, and if $0 \neq 1$ then we call it **non-trivial**.[6]

Proposition 1.6 (Trivial rings)**.** *Suppose that R is a ring. Then R is trivial if and only if R has one element.*

Proof. First, if R contains only one element, then since $0, 1 \in R$ we must have $0 = 1$. On the other hand, if $0 = 1$ then for any $x \in R$ we have $x = 1x = 0x = 0$ by Lemma 1.3, and so $R = \{0\}$. □

[5] The left (resp. right) multiplication maps are the maps $R \to R; y \mapsto xy$ (resp. $R \to R; y \mapsto yx$) for $x \in R$.

[6] Some authors (*e.g.* [Lam07]) use the terms **zero** and **non-zero** in place of trivial and non-trivial.

A ring R is said to be an **integral domain** if R is non-trivial, commutative, and R^* is closed under multiplication *i.e.* $xy \in R^*$ whenever $x, y \in R^*$.

In a ring R we call $x \in R$ a **left** (resp. **right**) **zero divisor** if there is $y \in R^*$ such that $xy = 0$ (resp. $yx = 0$).

Remark 1.7. If $x \in R$ is *not* a left (resp. right) zero divisor then left (resp. right) multiplication by x, which is a group homomorphism of the additive group, has trivial kernel and so is injective.

Remark 1.8. An integral domain has no non-zero zero divisors.

Remark 1.9. Units are never zero-divisors: if $x \in U(R)$ and $xy = 0$, then $0 = x^{-1}0 = x^{-1}(xy) = (x^{-1}x)y = 1y = y$ so $y \notin R^*$.

We say that $\mathbb{F}$ is a **field** if it is a commutative ring with $U(\mathbb{F}) = \mathbb{F}^*$.

Remark 1.10. Since $0 \notin U(\mathbb{F})$, $\mathbb{F}$ is non-trivial.

2 Homomorphisms, isomorphisms, and subrings

A **ring homomorphism**[7] is a map $\phi : R \to S$ between two rings such that

$$\phi(x + y) = \phi(x) + \phi(y) \text{ and } \phi(xy) = \phi(x)\phi(y) \text{ for all } x, y \in R, \text{ and } \phi(1) = 1.$$

There are some basic properties of homomorphisms we shall need.

Lemma 2.1. *Suppose that $\phi : R \to S$ and $\psi : S \to T$ are ring homomorphisms. Then $\psi \circ \phi$ is a ring homomorphism $R \to T$.*

Proof. This is immediate from the definition. □

Lemma 2.2. *Suppose that $\phi : R \to S$ is a ring homomorphism. Then $\phi(0) = 0$, $\phi(-x) = -\phi(x)$ for all $x \in R$, and if $x \in U(R)$ then $\phi(x) \in U(S)$ and $\phi(x^{-1}) = \phi(x)^{-1}$.*

Proof. First, $\phi(0) = 0$ and $\phi(-x) = -\phi(x)$ for all $x \in R$, since ϕ is a group homomorphism of the additive group and homomorphisms preserve identities and inverses. If $x \in U(R)$ then there is some $y \in U(R)$ such that $xy = yx = 1$ and hence $\phi(x)\phi(y) = \phi(y)\phi(x) = \phi(1) = 1$ so that $\phi(x) \in U(S)$. Thus ϕ restricts to a group homomorphism $U(R) \to U(S)$, and again group homomorphisms preserve inverses so the result is proved. □

Remark 2.3. ⚠If R is a non-trivial ring[8] then the map $\phi : \{0\} \to R; 0 \mapsto 0$ has $\phi(1) = \phi(0) = 0 \neq 1$ since $\{0\}$ is trivial and R is non-trivial, and so it is *not* a ring homomorphism. In particular, this example shows that we may not dispense with the requirement that $\phi(1) = 1$ in the definition of ring homomorphism.

[7] With an eye to generalisation, one might argue that the most natural definition of ring homomorphism would include the conclusions of Lemma 2.2 – if they did not follow we would add them in as assumptions.

[8] ⚠We have not yet shown that such a thing exists, but it is perhaps not surprising that it does.

Proposition 2.4. *Suppose that $\phi : \mathbb{F} \to R$ is a ring homomorphism, $\mathbb{F}$ is a field and R is non-trivial. Then ϕ is injective.*

Proof. If $\phi(x) = \phi(y)$ and $x \neq y$ then $x - y \in \mathbb{F}^*$ and so there is u such that $(x - y)u = 1$ whence $0 = 0\phi(u) = (\phi(x) - \phi(y))\phi(u) = \phi((x - y)u) = \phi(1) = 1$, which contradicts the non-triviality of R. We conclude that ϕ is injective as claimed. □

A ring S is a **subring** of a ring R if the map $j : S \to R; s \mapsto s$ is a well-defined ring homomorphism; S is a **proper subring** if S is a subring of R and $S \neq R$.

Remark 2.5. Subrings inherit some properties of the containing ring, *e.g.* being commutative and being non-trivial, and hence being an integral domain; but not others *e.g.* being a field.

A ring S is a **(proper) subfield** of a ring R if S is a field and a (proper) subring of R.

Remark 2.6. ⚠A ring that is not a field may have have a subfield.

Lemma 2.7 (Subring test)**.** *Suppose that R is a ring and S is a subset of R such that*

$$1 \in S, -x \in S \text{ for all } x \in S, \text{ and } x + y, xy \in S \text{ for all } x, y \in S. \tag{2.1}$$

Then the addition and multiplication on R restrict to well-defined operations on S giving it the structure of a subring of R.

Proof. First S is non-empty and closed under addition and negation so by the subgroup test addition on R restricts to a well-defined binary operation on S giving it the structure of a commutative group. Since S is closed under multiplication it also restricts to a well-defined binary operation on S, and is *a fortiori* associative since it is associative on R. Finally, $1 \in S$ and since this is an identity for R it is *a fortiori* an identity for S. Multiplication and addition restricted are *a fortiori* distributive when restricted to S, and so we conclude that S is a ring. The inclusion map is then well-defined and a ring homomorphism and the result is proved. □

Given a subset satisfying the hypotheses of the above lemma, we make the common abuse of calling it a subring on the understanding that we are referring to the induced operations.

Remark 2.8. Note that if R is a subring of S and S is a subring of T then R is a subring of T in view of Lemma 2.1.

Proposition 2.9. *Suppose that $\phi : R \to S$ is a homomorphism. Then* $\operatorname{Im} \phi$ *is a subring of S.*

Proof. This is immediate from the subring test and Lemma 2.2. □

Example 2.10 (Centre of a ring)**.** Given a ring R the **centre** (or **center**) is the set $Z(R) := \{z \in R : zr = rz \text{ for all } r \in R\}$; in words it is the set of elements of R that commute with all other elements of R. In particular, R is commutative if and only if $Z(R) = R$.

The centre is a subring of R by the subring test: $1 \in Z(R)$ since $1r = r = r1$ for all $r \in R$; if $x \in Z(R)$ then $(-x)r = -(xr) = -(rx) = r(-x)$ by Lemma 1.3; and if $x, y \in Z(R)$ then $(x+y)r = xr + yr = rx + ry = r(x+y)$ for all $r \in R$ by distributivity, so $x + y \in Z(R)$; and $(xy)r = x(yr) = x(ry) = (xr)y = (rx)y = r(xy)$ for all $r \in R$ by associativity of multiplication, so $xy \in Z(R)$.

Homomorphisms between rings can be particularly useful for endowing the image with additional structure.

Proposition 2.11. *Suppose that R is a ring, $\mathbb{F}$ is a field and $\phi : \mathbb{F} \to R$ is a ring homomorphism. Then the additive group of R equipped with scalar multiplication defined by $\mathbb{F} \times R \to R; (\lambda, v) \mapsto \lambda.v := \phi(\lambda)v$ is an $\mathbb{F}$-vector space. Furthermore, if the image of $\mathbb{F}$ is in the centre of R then the ring multiplication on R considered as an $\mathbb{F}$-vector space is bilinear.*[9]

Proof. First, the additive group of R is a commutative group by definition. Secondly, $(\lambda\mu).v = \lambda.(\mu.v)$ for all $\lambda, \mu \in \mathbb{F}$ and $v \in R$ since multiplication in R is associative and $\phi(xy) = \phi(x)\phi(y)$. Thirdly $1.v = v$ for all $v \in R$ since 1 is a multiplicative identity and $\phi(1) = 1$. Finally, $(\lambda + \mu).v = \lambda.v + \mu.v$ for all $\lambda, \mu \in \mathbb{F}$ and $v \in R$ since right multiplication in R and ϕ are both group homomorphisms; and $\lambda.(v + w) = \lambda.v + \lambda.w$ for all $\lambda \in \mathbb{F}$ and $v, w \in R$ since left multiplication is a group homomorphism.

Suppose that $\lambda, \mu \in \mathbb{F}$ and $u, v, w \in R$. Then the ring multiplication is linear in its first argument since right multiplication is a group homomorphism and multiplication in R is associative:

$$(\lambda.v + \mu.w)u = (\phi(\lambda)v + \phi(\mu)w)u = \lambda.(vu) + \mu.(wu);$$

it is linear in its second argument since left multiplication is a group homomorphism, and multiplication in R is associative *and* the image of $\mathbb{F}$ is in the centre of R:

$$\begin{aligned} u(\lambda.v + \mu.w) = u(\phi(\lambda)v + \phi(\mu)w) &= u(\phi(\lambda)v) + u(\phi(\mu)w) \\ &= \phi(\lambda)(uv) + \phi(\mu)(uw) = \lambda.(uv) + \mu.(uw). \end{aligned}$$

The result is proved. □

A **ring isomorphism** is a map $\phi : R \to S$ that is a bijective ring homomorphism.[10]

[9] In the literature this is sometimes expressed by saying that R is a **unital associative $\mathbb{F}$-algebra**.

[10] If we did not have Lemma 2.12 then we would also insist here that the inverse map be a ring homomorphism. There is a comparison here with the situation in which topological spaces (or more concretely subsets of the reals) replace rings and continuous maps replace homomorphisms. The map $f : [0, 1) \cup \{2\} \to [0, 1]$ with $f(x) := x$ if $x \neq 2$ and $f(2) := 1$, is a continuous bijection but does not have a continuous inverse.

Lemma 2.12. *Suppose that $\phi : R \to S$ is a ring isomorphism. Then ϕ^{-1} is a ring homomorphism, and hence a ring isomorphism.*

Proof. First, $\phi(1) = 1$ and ϕ is a bijection so $\phi^{-1}(1) = 1$. Secondly, ϕ is a bijective group homomorphism of the additive group and so ϕ^{-1} is also a group homomorphism of the additive group of S. Finally, if $x, y \in S$ then by surjectivity there are elements $u, v \in R$ such that $\phi(u) = x$ and $\phi(v) = y$, and

$$\phi^{-1}(xy) = \phi^{-1}(\phi(u)\phi(v)) = \phi^{-1}(\phi(uv)) = uv = \phi^{-1}(x)\phi^{-1}(y).$$

We conclude that ϕ^{-1} is a homomorphism and the result is proved. □

We say that two rings R and S are **(ring) isomorphic** and write $R \cong S$ if there is a ring isomorphism $R \to S$.

Proposition 2.13. *$\cong$ is an equivalence relation.*

Proof. The identity map on a ring is an isomorphism so $\cong$ is reflexive. $\cong$ is symmetric in view of Lemma 2.12. Finally, $\cong$ is transitive since the composition of bijections is a bijection, and composition of ring homomorphisms is a ring homomorphism – this is Lemma 2.1. □

3 The natural numbers

We write $\mathbb{N}_0$ for the natural numbers including 0, and $\mathbb{N}^*$ for the naturals *without* 0. These come equipped with a map[11] $\mathbb{N}_0 \to \mathbb{N}_0; x \mapsto x + 1$ which is an injection with image $\mathbb{N}^*$; and enjoy the inductive axiom that if $X \subset \mathbb{N}_0$ has $0 \in X$ and $X + 1 \subset X$ then $X = \mathbb{N}_0$. These are essentially Peano's axioms for the natural numbers and we shall not concern ourselves with the question of whether such an object exists[12].

The first axiom here is a way of capturing the fact that the natural numbers are infinite[13] without reference to the usual definition of finite[14] which would be circular. The second axiom – induction – captures the fact that $\mathbb{N}_0$ is minimal subject to the requirement that it is infinite.

[11] Called the **successor function**.

[12] We fall back on the 1886 quote "Die ganzen Zahlen hat der liebe Gott gemacht, alles andere ist Menschenwerk" attributed to Kronecker by Weber [Web92, p19], and translated as "God made the integers, all else is the work of man" by Gray [Gra08, p153].

[13] A set X is said to be **Dedekind finite** if any injection $f : X \to X$ is surjective. In particular the successor function bears witness to the fact that the naturals are *not* Dedekind finite.

[14] Recall that a set I is **finite** if there is some $n \in \mathbb{N}_0$ and a bijection $\phi : \{0, \dots, n-1\} \to I$ with the convention that if $n = 0$ (so that there is no natural number predecessor $n - 1$) then the domain is the empty set.

These properties of the naturals are basic to recursive definitions and can immediately be used to produce the usual binary operations of addition[15] and multiplication[16]. Addition gives rise to a total order[17] on $\mathbb{N}_0$ in which $x \geqslant y$ if and only if there is $z \in \mathbb{N}_0$ such that $x = y + z$, and this also allows us to give an equivalent[18] formulation of the induction axiom as the well-ordering principle, that if $X \subset \mathbb{N}_0$ is non-empty then it has a minimal element.

Iterated sums and products in rings

Given a ring R, the binary operations of addition and multiplication can be applied recursively: We define

$$R^n \to R; (x_0, \dots, x_{n-1}) \mapsto \sum_{i=0}^{n-1} x_i \text{ and } R^n \to R; (x_0, \dots, x_{n-1}) \mapsto \prod_{i=0}^{n-1} x_i \tag{3.1}$$

to be the constant values 0_R and 1_R respectively when $n = 0$, and then recursively by

$$\sum_{i=0}^{n} x_i := \left(\sum_{i=0}^{n-1} x_i\right) + x_n \text{ and } \prod_{i=0}^{n} x_i := \left(\prod_{i=0}^{n-1} x_i\right) x_n.$$

We extend this notation so that for $x_m, \dots, x_n \in R$ we write

$$\sum_{i=m}^{n} x_i := \sum_{i=0}^{n-m} x_{m+i} \text{ and } \prod_{i=m}^{n} x_i := \prod_{i=0}^{n-m} x_{m+i},$$

with the convention that if $n < m$ the sum is 0_R and the product is 1_R. Finally we also write

$$x_m + \cdots + x_n := \sum_{i=m}^{n} x_i \text{ and } x_m \cdots x_n := \prod_{i=m}^{n} x_i,$$

which is compatible with existing notation when $n = 2$.

Iterated sums and products of the identities

Induction and the fact that 0_R and 1_R are identities for their respective operations gives

$$\sum_{i=0}^{n-1} 0_R = 0_R \text{ and } \prod_{i=0}^{n-1} 1_R = 1_R.$$

[15] $x + 0 := x$ and $x + (y + 1) := (x + y) + 1$ for all $x, y \in \mathbb{N}_0$.

[16] $x \times 0 := 0$ and $x \times (y + 1) := x \times y + x$ for all $x, y \in \mathbb{N}_0$.

[17] $\geqslant$ is a **total order** on X if $\geqslant$ is reflexive, meaning $x \geqslant x$ for all x; $\geqslant$ is transitive, meaning $x \geqslant y$ and $y \geqslant z$ implies $x \geqslant z$; $\geqslant$ is anti-symmetric, meaning $x \geqslant y$ and $y \geqslant x$ if and only if $x = y$; and trichotomous meaning $x \geqslant y$ or $y \geqslant x$.

[18] There is a slight subtlety with this equivalence which we have avoided by insisting that the successor function has image the whole of $\mathbb{N}^*$ rather than just a proper subset. One can prove the former from the latter using induction but not using well-ordering. See [Öhm19] for a discussion.

Associativity of iterated sums and products

Induction with ring associativity (and the recursive definition of addition) shows that for $x \in R^{n+m}$ we have

$$\sum_{i=0}^{n+m-1} x_i = \left(\sum_{i=0}^{n-1} x_i\right) + \left(\sum_{i=0}^{m-1} x_{n+i}\right) \text{ and } \prod_{i=0}^{n+m-1} x_i = \left(\prod_{i=0}^{n-1} x_i\right)\left(\prod_{i=0}^{m-1} x_{n+i}\right).$$

Distributivity of iterated sums and products

A further induction with ring distributivity (and the recursive definition of multiplication) shows that for $x \in R^n, y \in R^m$ we have

$$\sum_{k=0}^{nm-1} z_k = \left(\sum_{i=0}^{m-1} x_i\right)\left(\sum_{j=0}^{n-1} y_j\right) = \sum_{k=0}^{mn-1} w_k$$

where $z_{ni+j} := x_iy_j$ and $w_{mj+i} := x_iy_j$ for $0 \leqslant i \leqslant m-1, 0 \leqslant j \leqslant n-1$. As it happens these two equalities are really the same because addition is commutative *c.f.* Exercise I.1.

Commutativity of iterated sums and (in commutative rings) products

Commutativity of ring addition coupled with the fact that the permutations of $\{0, \ldots, n-1\}$ are generated by transpositions of consecutive elements shows that for $x \in R^n$ we have

$$\sum_{i=0}^{n-1} x_{\pi(i)} = \sum_{i=0}^{n-1} x_i \text{ for all bijections } \pi : \{0, \ldots, n-1\} \to \{0, \ldots, n-1\}. \tag{3.2}$$

This fact permits a definition of unordered sum: suppose that I is a finite set and $x \in R^I$. We write

$$\sum_{i\in I} x_i := \sum_{i=0}^{n-1} x_{\sigma(i)} \text{ where } \sigma : \{0, \ldots, n-1\} \to I \text{ is } \textit{any} \text{ bijection.} \tag{3.3}$$

Since I is finite[14] there is an $n \in \mathbb{N}_0$ such that such a bijection exists, and furthermore different choices of bijection give rise to the same sum in view of (3.2).

In a commutative ring we have an analogue of (3.2) for products, and hence for $x \in R^I$ we can define an analogue of (3.3) for products which we write as $\prod_{i\in I} x_i$.

Homomorphisms of iterated sums and products

Finally, given a ring homomorphism $\phi : R \to S$ and $x \in R^n$, by induction we get

$$\phi\left(\sum_{i=0}^{n-1} x_i\right) = \sum_{i=0}^{n-1} \phi(x_i) \text{ and } \phi\left(\prod_{i=0}^{n-1} x_i\right) = \prod_{i=0}^{n-1} \phi(x_i).$$

Remark 3.1. One should not worry too much about the above. These definitions have to be made, but the rough idea is that iterated operations ‘work in the way we expect’.

The integers

By a series of inductions we can show that the natural numbers equipped with their usual addition and multiplication satisfy all the axioms of a ring except for the existence of additive inverses.

There is a standard construction for ‘adding in’ the negative numbers, which in some sense goes back at least to Liu Hui (劉徽) [SCL99, p404], that extends the usual addition and multiplication on the natural numbers, and is minimal in the sense that *every* integer can be written as a difference of two natural numbers.

Example 3.2 ($\mathbb{Z}$)**.** The integers, denoted $\mathbb{Z}$, have the structure of an integral domain in which the zero is 0, the multiplicative identity is 1, and $U(\mathbb{Z}) = \{-1, 1\}$.

In *any* ring we can write sums and products of differences as differences by the following lemma.

Lemma 3.3. *Suppose that R is a ring. Then*

$$(a-d)+(b-c)=(a+b)-(c+d) \text{ and } (a-d)(b-c)=(ab+dc)-(ac+db)$$

for all $a, b, c, d \in R$.

Proof. The first of these follows by commutativity and associativity of addition and distributivity of negation over addition. The second by distributivity of multiplication, commutativity and associativity of addition, and distributivity of negation over addition and multiplication. □

Remark 3.4. These identities can actually be used to *define* the addition and multiplication on $\mathbb{Z}$ in terms of that on $\mathbb{N}_0$.

Remark 3.5. As a special case of the definition of the iterated product, for $x \in R$ we put $x^{n+1} := x^n x$ for $n \in \mathbb{N}_0$ and $x^0 := 1_R$. Induction shows that

$$x^0 = 1_R, x^{n+m} = x^n x^m \text{ and } x^{nm} = (x^n)^m \text{ for all } n, m \in \mathbb{N}_0 \text{ and } x \in R$$

Moreover, if $x \in U(R)$ then this extends to the integers by $x^{n-m} := x^n (x^{-1})^m$ for $n, m \in \mathbb{N}_0$. This extension is well-defined (in particular the two possible meanings of x^{-1} coincide) and has

$$x^0 = 1_R, (x^n)^{-1} = (x^{-1})^n, x^{n+m} = x^n x^m \text{ and } x^{nm} = (x^n)^m \text{ for all } n, m \in \mathbb{Z} \text{ and } x \in U(R)$$

Similarly, for addition we have $0.x := 0_R$ and $(n+1).x := (n.x) + x$, and $(n-m).x := (n.x) - (m.x)$ which is also well-defined and has

$$0.x = 0_R, -(n.x) = (-n).x, (n+m).x = (n.x) + (m.x) \text{ and } (nm).x = n.(m.x) \tag{3.4}$$

for all $n, m \in \mathbb{Z}$ and $x \in R$.

Remark 3.6. ⚠$n.(x+y) = n.x + n.y$ for all $n \in \mathbb{Z}$ and $x, y \in R$, *but* $(xy)^n = x^n y^n$ for all $n \in \mathbb{Z}$ if and only if x and y commute. ⚠Of course x and y must be units in this last case.

The integers have an important relationship[19] to rings in general captured by the following proposition.

Proposition 3.7. *Suppose that R is a ring. Then the map $\mathbb{Z} \to R; n \mapsto n.1_R$ is a ring homomorphism into the centre of R, and this map is the only ring homomorphism $\mathbb{Z} \to R$.*

Proof. First, $1.1_R = 0.1_R + 1_R = 0_R + 1_R = 1_R$. Secondly, $(n+m).1_R = n.1_R + m.1_R$ for all $n, m \in \mathbb{Z}$ by (3.4). Finally, induction using distributivity of multiplication in R gives $(nm).1_R = (n.1_R)(m.1_R)$ for all $n, m \in \mathbb{N}_0$ and then Lemma 3.3 for *both* $\mathbb{Z}$ and R extends this to the integers. Hence the given map is a homomorphism. For $n \in \mathbb{N}_0$, $n.1_R$ is in the centre of R by induction and since the centre is a ring we conclude that the image of the homomorphism is in the centre of R.

In the other direction, suppose that $\psi : \mathbb{Z} \to R$ is a homomorphism. Then $\psi(n) = n.1_R$ for all $n \in \mathbb{N}_0$ by induction since ψ is a homomorphism of the additive group and $\psi(1) = 1_R$. But then $\psi(n-m) = \psi(n) - \psi(m) = n.1_R - m.1_R = (n-m).1_R$ for all $n, m \in \mathbb{N}_0$ by Lemma 2.2. The result is proved. □

The **characteristic** of a ring R is 0 if $n.1_R = 0$ implies $n = 0$, and otherwise it is the smallest $n \in \mathbb{N}^*$ such that $n.1_R = 0_R$.

Example 3.8. The characteristic of $\mathbb{Z}$ is 0.

Remark 3.9. Note that if S is a subring of R then the characteristic of S is the same as that of R.

Example 3.10. $\mathbb{Z}$ has no proper subring: If S is a subring of $\mathbb{Z}$, then the inclusion $j : S \to \mathbb{Z}$ is a ring homomorphism. By Proposition 3.7 there a ring homomorphism $\phi : \mathbb{Z} \to S$, and then $j \circ \phi$ is a ring homomorphism $\mathbb{Z} \to \mathbb{Z}$ so by the uniqueness of Proposition 3.7 it is the identity. It follows that j is surjective and hence $S = \mathbb{Z}$.

4 Examples

Perhaps unsurprisingly there are more examples of rings than just the integers.

[19] In the language of category theory the integers are an **initial object** in the category of rings, and one might feel lured into describing them as one ring (up to isomorphism) ruling (by copies of the 'line' of integers) all others, though the relevance of ambient light conditions and the extent to which the integers actually find, bring or bind any other ring is less clear.

Example 4.1 ($\mathbb{Z}_N$)**.** Given $N \in \mathbb{N}^*$ we write $x \equiv y \pmod{N}$ if $N \mid x - y$. This is an equivalence relation on $\mathbb{Z}$ and it is compatible with the addition and multiplication there, meaning if $x \equiv y \pmod{N}$ and $x' \equiv y' \pmod{N}$ then

$$x + x' \equiv y + y' \pmod{N} \text{ and } xx' \equiv yy' \pmod{N}.$$

We write $\mathbb{Z}_N$ for the set of congruence classes of integers under this relation, and the compatibility above means that the multiplication and addition on $\mathbb{Z}$ induce a ring structure on $\mathbb{Z}_N$. This construction is an example of a quotient ring which we shall meet in more generality in §5.

$\mathbb{Z}_N$ is commutative, has the congruence class of 0 as its zero, the congruence class of 1 as its multiplicative identity, and the congruence class of $-x$ as the additive inverse of the congruence class of x.

The characteristic of $\mathbb{Z}_N$ is N.

If N is composite then $N = ab$ for $1 < a, b < N$ and so $ab \equiv 0 \pmod{N}$ while $a, b \not\equiv 0 \pmod{N}$. It follows that $\mathbb{Z}_N$ is *not* an integral domain.

Integers x and y are said to be **coprime** if their only common factors are units, meaning if $a \mid x$ and $a \mid y$ then a is a unit *i.e.* $a \in \{-1, 1\}$.

Theorem 4.2 (Bezout's Lemma)**.** *Suppose that $x, y \in \mathbb{Z}$ are coprime. Then there are $\alpha, \beta \in \mathbb{Z}$ such that $\alpha x + \beta y = 1$.*

This has been covered in Prelims Mathematics I but will also follow from Example 5.9. A large part of the course will concern rings where we have an analogue of Bezout's Lemma – this is roughly what a PID is. This will be defined formally in Remark 5.10.

Example 4.3 (The field $\mathbb{F}_p$, and the group of units of $\mathbb{Z}_N$)**.** The units of $\mathbb{Z}_N$ are the (congruence classes of) integers coprime to N: If r and N are coprime then Bezout's Lemma exactly tells us that there is $\alpha \in \mathbb{Z}$ such that $\alpha r \equiv 1 \pmod{N}$, and so by commutativity of $\mathbb{Z}_N$, the congruence class of r is a unit. Conversely, if a is a non-unit common factor of r and N then $r \times (N/a) \equiv 0 \pmod{N}$ and so the congruence class of r is a zero-divisor (since $(N/a) \not\equiv 0 \pmod{N}$) and so not a unit by Remark 1.9.

Suppose that p is prime. Then every $1 \leqslant r < p$ is coprime to p, and so $U(\mathbb{Z}_p)$ contains every non-zero congruence class. Since $p > 1$ we also have that $\mathbb{Z}_p$ is non-trivial so $U(\mathbb{Z}_p) = \mathbb{Z}_p^*$. Finally, $\mathbb{Z}_p$ is commutative and so it is a field; we write $\mathbb{F}_p$ for $\mathbb{Z}_p$ to emphasise this property.[20]

The rationals – another important example for us – are a field which can be constructed from the integers in a way which generalises as follows to any integral domain.

[20] ⚠$\mathbb{Z}_p$ is sometimes (*e.g.* [Lam07]) used to denote a different ring (which we shall not consider) called the **p-adic integers**.

Theorem 4.4 (Field of fractions and its characterisation)**.** *Suppose that R is an integral domain. Then there is a field $\mathbb{F}$ such that R is a subring of $\mathbb{F}$, and no proper subfield of $\mathbb{F}$ contains R. Moreover, if $\mathbb{K}$ is a field containing R and no proper subfield of $\mathbb{K}$ contains R then there is an isomorphism $\phi : \mathbb{K} \to \mathbb{F}$ which is the identity on R.*

Remark 4.5. The proof is not hard and can be found in many places *e.g.* [Hun80, Theorem 4.3] and [Lan02, Chapter II, §4], and it is not dissimilar to the construction of the integers from the naturals by 'adding in' the negative numbers. It is omitted from the syllabus because all it really does is formalise the content of Remark 4.6 below.

Remark 4.6. Suppose that $\mathbb{F}$ is a field of fractions for R, and consider the set $F(R) := \{ab^{-1} : a \in R, b \in R^*\}$ as a subset of $\mathbb{F}$. This contains $1 = 1/1^{-1}$ and is closed under additive inverses as well as addition and multiplication since

$$ac^{-1} + bd^{-1} = (ad + bc)(cd)^{-1} \text{ and } (ac^{-1})(bd^{-1}) = (ab)(cd)^{-1}.$$

It follows from the subring test that $F(R)$ is a subring and it contains R. Now, if $ab^{-1} \neq 0$ then $a \in R^*$ so $ba^{-1} \in F(R)$, and hence $F(R)$ is closed under multiplicative inverses and so a field, whence $F(R) = \mathbb{F}$. This motivates the name field of fractions: all the elements of $\mathbb{F}$ can be written as a 'fraction' ab^{-1}.

Remark 4.7. Note that the field of fractions is not just unique up to isomorphism as a field, but – and this is ensured by the part of the statement that says ϕ is the identity on R – R 'sits inside' its field of fractions in a unique way.

Example 4.8 ($\mathbb{Q}$)**.** The rationals, denoted $\mathbb{Q}$, are the field of fractions of the integers. The order on the integers extends to a total order (also denoted $\geqslant$) on the rationals such that $x + z \geqslant y + z$ whenever $x \geqslant y$ and $xy \geqslant 0$ whenever $x, y \geqslant 0$. Of course the well-ordering of the naturals w.r.t. $\geqslant$, which in $\mathbb{Z}$ manifests as the fact that every non-empty subset of $\mathbb{Z}$ that is bounded below has a minimum element, does *not* extend to $\mathbb{Q}$.

The rationals are an example of a field with a subring – $\mathbb{Z}$ – that is not a field, bearing out the last part of Remark 2.5.

⚠$U(\mathbb{Q}) \cap \mathbb{Z} \not\subset U(\mathbb{Z})$ so being a unit in $\mathbb{Q}$ and an element of $\mathbb{Z}$ does *not* guarantee the status of unit in $\mathbb{Z}$.

⚠Let $\psi : \mathbb{Z} \to \mathbb{Q}$ be the usual inclusion and suppose $f, g : \mathbb{Q} \to R$ are ring homomorphisms such that[21] $f \circ \psi = g \circ \psi$. For $m \in \mathbb{Z}^*$ we have $g(1/m) = g(1/m)f(m)f(1/m) = g(1/m)g(m)f(1/m) = f(1/m)$ and hence $g(n/m) = g(n)g(1/m) = f(n)f(1/m) = f(n/m)$ for all $n \in \mathbb{Z}$. It follows that $f \equiv g$ despite the fact that ψ is *not* surjective.

[21] ⚠As it happens this hypothesis is automatically satisfied for $\mathbb{Z}$ because there is a unique homomorphism from $\mathbb{Z} \to R$ (Proposition 3.7) and the composition of homomorphisms is a homomorphism. It is included because the purpose of the example is to illustrate that 'right cancellability' of ring homomorphisms does not require them to be surjective, unlike functions more generally.

Example 4.9 ($\mathbb{R}$)**.** The reals, like the rationals, are a field with a total order $\geqslant$ such that $x+z \geqslant y+z$ whenever $x \geqslant y$ and $xy \geqslant 0$ whenever $x, y \geqslant 0$. However, they also have the additional property that any non-empty subset of $\mathbb{R}_{\geqslant 0}$ that is bounded below has a greatest lower bound (though this lower bound may not be in the set, unlike $\mathbb{Z}$).

Example 4.10. The ring $\mathbb{Q}[\sqrt{2}] := \{a + b\sqrt{2} : a, b \in \mathbb{Q}\}$ is a subfield of $\mathbb{R}$. Indeed, it is a commutative subring of $\mathbb{R}$ by the subring test (since $\sqrt{2}^2 \in \mathbb{Q}$). Now suppose $a + b\sqrt{2} \in \mathbb{Q}[\sqrt{2}]^*$ so that $a^2 - 2b^2 \neq 0$ (since $\sqrt{2}$ is irrational). Then

$$(a+b\sqrt{2})\left(\frac{a}{a^2-2b^2} - \frac{b}{a^2-2b^2}\sqrt{2}\right) = 1$$

so that $a + b\sqrt{2} \in U(\mathbb{Q}[\sqrt{2}])$ as claimed.

Example 4.11 ($\mathbb{C}$)**.** The complex numbers are a ring whose additive group is a vector space over $\mathbb{R}$ with basis 1 and i, and with bilinear multiplication determined by $i^2 = -1$. In particular, for $z, w \in \mathbb{C}$ there are unique $x, y, u, v \in \mathbb{R}$ such that $z = x + iy$ and $w = u + iv$, and we can compute that

$$-z = (-x) + i(-y), z + w = (x+u) + i(y+v), \text{ and } zw = (xu - yv) + i(xv + yu).$$

⚠Note that unlike the rationals and reals there is *no* total order $\geqslant$ on $\mathbb{C}$ such that $x + z \geqslant y + z$ whenever $x \geqslant y$, and $xy \geqslant 0$ whenever $x, y \geqslant 0$. Indeed, suppose that there were. By trichotomy either $1 \geqslant 0$ or $0 \geqslant 1$; in the latter case $(-1) > 1 + (-1) = 0$. Hence either $1 = 1^2 \geqslant 0$ or $1 = (-1)^2 \geqslant 0$; we conclude that $1 > 0$ (since $1 \neq 0$ by non-triviality of $\mathbb{C}$). Again, by trichotomy either $i \geqslant 0$ or $0 \geqslant i$. In the former case $-1 = i^2 \geqslant 0$ so $0 \geqslant 1$, a contradiction; in the latter $-i \geqslant i + (-i) = 0$ and so $-1 = (-i)^2 \geqslant 0$ and a contradiction again.

Example 4.12. The ring $\mathbb{Z}[i] := \{a + bi : a, b \in \mathbb{Z}\}$ is called the ring of **Gaussian integers**. It is a subring of $\mathbb{C}$ by the subring test since $i^2 \in \mathbb{Z}$ and so an integral domain, and $U(\mathbb{Z}[i]) = \{1, -1, i, -i\}$. To see this last fact suppose that $(a + ib)(c + id) = 1$ for $a, b, c, d \in \mathbb{Z}$, then taking absolute values we have $(a^2 + b^2)(c^2 + d^2) = 1$. We conclude $a^2 + b^2 = 1$, and hence $(a, b) \in \{(1, 0), (-1, 0), (0, 1), (0, -1)\}$ as required.

Direct products of rings

For this we require some more notation. Suppose that S_i is a set for each $i \in I$. We write $\prod_{i \in I} S_i$ for the set of $x = (x_i)_{i \in I}$ with $x_i \in S_i$ for all $i \in I$. If $S_i = S$ for all $i \in I$ then we write S^I or $\prod_{i \in I} S$ for $\prod_{i \in I} S_i$, and for $n \in \mathbb{N}_0$ we write S^n for $S^{\{0,\dots,n-1\}}$ and sometimes $(s_0, \dots, s_{n-1})$ in place of $(s_i)_{i \in \{0,\dots,n-1\}}$. In particular $S^0 = S^{\varnothing}$ contains one element which we denote $()$.

Proposition 4.13. *Suppose that $(R_i)_{i\in I}$ is a family of rings. Then $\prod_{i\in I} R_i$ may be equipped with the structure of a ring, called the* **direct product** *and denoted $\prod_{i\in I} R_i$, with*

$$(-x)_i = -x_i, (x+y)_i = x_i + y_i, \text{ and } (xy)_i = x_iy_i \text{ for all } i \in I, x, y \in \prod_{i\in I} R_i,$$

and $0_{\prod_{i\in I} R_i} = (0_{R_i})_{i\in I}$ and $1_{\prod_{i\in I} R_i} = (1_{R_i})_{i\in I}$, and

$$U\left(\prod_{i\in I} R_i\right) = \prod_{i\in I} U(R_i) \text{ and if } x \in U\left(\prod_{i\in I} R_i\right) \text{ then } (x^{-1})_i = x_i^{-1} \text{ for all } i \in I.$$

Moreover, $\prod_{i\in I} R_i$ is commutative if and only if R_i is commutative for all $i \in I$, and the projection maps $\prod_{i\in I} R_i \to R_j; x \mapsto x_j$ are ring homomorphisms.

Proof. Since $x = y$ if and only if $x_i = y_i$ for all $i \in I$, all the axioms of a ring, along with the description of the group of units, and the commutativity of multiplication if it holds, are inherited from the corresponding axioms for the rings R_i *e.g* addition is associative for $\prod_{i\in I} R_i$ because addition is associative in R_i for all $i \in I$. The fact that the projection maps are ring homomorphisms is a quick check. □

Remark 4.14. The empty product of rings has the structure of the trivial ring.

Example 4.15. Given a ring R, R^2 is *never* an integral domain: indeed, if R is trivial then R^2 is trivial and so not an integral domain; if R is non-trivial then $(0,1), (1,0) \in (R^2)^*$, but $(0,1)(1,0) = (0,0) = 0_{R^2}$ and so R^2 is not an integral domain.

Polynomial rings

Given a commutative ring R, the **polynomial ring over R with indeterminate X** is a commutative ring, denoted $R[X]$ and whose elements are called **polynomials**, with a distinguished element X such that every element $p = p(X) \in R[X]$ can be written in the form

$$p(X) = a_0 + a_1X + \cdots + a_nX^n \text{ for some } n \in \mathbb{N}_0 \text{ and } a_0, \dots, a_n \in R,$$

and furthermore,

$$\sum_{i=0}^{n} a_iX^i = 0_{R[X]} \text{ if and only if } a_0, \dots, a_n = 0_R. \tag{4.1}$$

If $p(X) = a_0 + a_1X + \cdots + a_nX^n \in R[X]^*$ then there is a largest $d \leqslant n$ such that $a_d \neq 0_R$. We call this d the **degree** of p and denote it $\deg p$; we call $a_0, a_1, \dots, a_d$ the **coefficients** of p; and we call a_d the **lead coefficient** and a_0 the **constant coefficient**.

Remark 4.16. It is reasonably straightforward to show that such a ring exists and is unique though it is not especially illuminating to do so. The important point is that the ring structure already tells us how to add and multiply polynomials as we shall discuss again in Remark 4.20.

Remark 4.17. The 'freeness' of the polynomial ring comes from the 'only if' part of (4.1); when we are using it we shall often say 'by equating coefficients'.

⚠While the maps $\mathbb{F}_p \to \mathbb{F}_p; \lambda \mapsto \lambda^p$ and $\mathbb{F}_p \to \mathbb{F}_p; \lambda \mapsto \lambda$ are the same by Fermat's Little Theorem, the polynomials X^p and X in $\mathbb{F}_p[X]$ are distinct.

Example 4.18. Suppose that $\mathbb{F}$ is a field. Then the inclusion map $\mathbb{F} \to \mathbb{F}[X]$ gives $\mathbb{F}[X]$ the structure of an $\mathbb{F}$-vector space by Proposition 2.11, and $1, X, X^2, \dots$ is linearly independent by the 'only if' part of (4.1), and spanning by definition of $\mathbb{F}[X]$. We conclude that $1, X, X^2, \dots$ is a basis for $\mathbb{F}[X]$.

When we meet modules later, we shall see how this generalises to $R[X]$.

Remark 4.19. We define $R[X_1, \dots, X_n] := R[X_1, \dots, X_{n-1}][X_n]$ and call $R[X_1, \dots, X_n]$ the **polynomial ring in the indeterminants** $X_1, \dots, X_n$.

Remark 4.20. There is some basic algebra of polynomial expressions that is useful. Suppose that R is a subring of S and $\lambda \in S$ commutes with all elements of R. Then we have

$$-\left(\sum_{i=0}^{n} a_i\lambda^i\right) = \sum_{i=0}^{n}(-a_i)\lambda^i, \text{ and[22] } \left(\sum_{i=0}^{n} a_i\lambda^i\right) + \left(\sum_{i=0}^{m} b_i\lambda^i\right) = \sum_{i=0}^{\max\{n,m\}} (a_i + b_i)\lambda^i, \qquad (4.2)$$

and

$$\left(\sum_{i=0}^{n} a_i\lambda^i\right)\left(\sum_{i=0}^{m} b_i\lambda^i\right) = \sum_{i=0}^{n+m}\left(\sum_{j=0}^{i} a_j b_{i-j}\right)\lambda^i. \qquad (4.3)$$

Note, in particular if $S = R[X]$ and $\lambda = X$, then in particular we have the rules for adding polynomials.

The important consequence of the 'freeness' mentioned in Remark 4.17 is the following.

Proposition 4.21. *Suppose that $\phi : R \to S$ is a ring homomorphism from a commutative ring R and $\lambda \in S$ commutes with all elements in the image of ϕ. Then the map*

$$R[X] \to S; p(X) = a_0 + a_1X + \cdots + a_dX^d \mapsto p(\lambda) := \phi(a_0) + \phi(a_1)\lambda + \cdots + \phi(a_d)\lambda^d$$

is a well-defined ring homomorphism.

Proof. The map is well-defined by the 'freeness' of the polynomial ring, that is (4.1), and the additive algebra of the polynomials, that is (4.2); denote it by $\tilde{\phi}$. We have $\tilde{\phi}(1) = 1$ since $\phi(1) = 1$, and $\tilde{\phi}$ is a homomorphism of the additive groups by the additive algebra of polynomials since ϕ is a homomorphism of the additive groups, and $\tilde{\phi}(pq) = \tilde{\phi}(p)\tilde{\phi}(q)$ by the multiplicative algebra of polynomials, that is (4.3), and the fact homomorphisms respect iterated sums (and also products), and the fact that λ commutes with all elements of the image of ϕ. □

[22] We take the convention convention that if $m < n \leqslant \max\{n, m\}$ then $b_i = 0_R$ for $m < i \leqslant n$, and if $n < m \leqslant \max\{n, m\}$ then $a_i = 0_R$ for $n < i \leqslant m$.

Remark 4.22. This homomorphism is called the **evaluation homomorphism** and its image is a ring by Proposition 2.9; we denote it $R[\lambda]$.

⚠Note that the particular homomorphism $\phi : R \to S$ does *not* appear in the notation $R[\lambda]$. Often the homomorphism is the inclusion map. Indeed, the inclusion map $R \to R[X]$ is a homomorphism (this is exactly the statement that R is a subring of $R[X]$) and $X \in R[X]$ commutes with all elements of (the image under inclusion of) R. With this map the two possible meanings of $R[X]$ (the one here and the one defined when we defined polynomial rings) coincide. Similarly the inclusion maps $\mathbb{Q} \to \mathbb{Q}[\sqrt{2}]$ and $\mathbb{Z} \to \mathbb{Z}[i]$ for the rings already defined in Examples 4.10 and 4.12 give notation that is compatible with that already set in those examples.

Integral domains produce polynomial rings where the degree function behaves nicely:

Proposition 4.23. *Suppose that R is a non-trivial commutative ring. Then TFAE:*

(i) R is an integral domain;

(ii) $R[X]$ is an integral domain;

(iii) for every $p, q \in R[X]^*$ we have $pq \in R[X]^*$ and $\deg pq = \deg p + \deg q$.

Proof. Certainly (ii) implies (i) since R is a subring of $R[X]$, and (iii) implies (ii) since R is non-trivial and $R[X]$ is commutative, and so the fact $R[X]$ is an integral domain follows by forgetting the second part of (iii).

To see (i) implies (iii) suppose that $p, q \in R[X]^*$ have degree n and m, and lead coefficients a_n and b_m respectively. Then from the multiplicative algebra of polynomials (4.3) we see that $\deg pq \leqslant n + m$, and the coefficient of X^{n+m} is $a_n b_m \in R^*$ since R is an integral domain and $a_n, b_m \in R^*$ by definition of degree. We conclude that $pq \in R[X]^*$ and $\deg pq = n + m$ as required. □

The **constant polynomials** are the degree 0 polynomials along with the zero polynomial.

Example 4.24. Given an integral domain R the group of units of $R[X]$ is the set of non-zero constant polynomials whose (only) non-zero coefficient is a unit of R. Indeed, suppose that $p \in U(R[X])$. Then there is some $q \in U(R[X])$ such that $pq = 1$, and so by Proposition 4.23 (iii) we have $0 = \deg p + \deg q$ and so $\deg p = 0$ and $\deg q = 0$ and hence $p(X) = a_0$ and $q(X) = b_0$ for some $a_0, b_0 \in R$. Since $a_0 b_0 = 1$ and R is commutative we conclude that $a_0 \in U(R)$ as required. The converse is immediate.

Matrix rings

Given a field $\mathbb{F}$ we write $M_{n,m}(\mathbb{F})$ for the set of $n \times m$ matrices with values in $\mathbb{F}$ and $M_n(\mathbb{F}) := M_{n,n}(\mathbb{F})$.

Proposition 4.25. *Suppose that $\mathbb{F}$ is a field and $n \in \mathbb{N}^*$. Then $M_n(\mathbb{F})$ is a ring with addition and multiplication satisfying*

$$A + B = (A_{i,j} + B_{i,j})_{i,j=1}^n \text{ and } AB = \left(\sum_{k=1}^n A_{i,k}B_{k,j}\right)_{i,j=1}^n \text{ for } A, B \in M_n(\mathbb{F}),$$

zero $(0_{\mathbb{F}})_{i,j=1}^n$, multiplicative identity I where $I_{i,i} = 1_{\mathbb{F}}$ for $1 \leqslant i \leqslant n$ and $I_{i,j} = 0_{\mathbb{F}}$ for $i \neq j$, and $-A = (-A_{i,j})_{i,j=1}^n$ for $A \in M_n(\mathbb{F})$.

Remark 4.26. It is reasonably straightforward to show that such a ring exists though it is not especially illuminating to do so. We shall revisit this in Proposition 13.9.

Example 4.27. The map $\mathbb{F} \to M_n(\mathbb{F}); \lambda \mapsto \lambda I$ is a homomorphism mapping $\mathbb{F}$ into the centre of $M_n(\mathbb{F})$. Proposition 2.11 shows that this gives $M_n(\mathbb{F})$ the structure of an $\mathbb{F}$-vector space such that multiplication on $M_n(\mathbb{F})$ is bilinear, and the n^2 matrices $E^{(i,j)}$ for $1 \leqslant i, j \leqslant n$ (which have $E^{(i,j)}_{i,j} = 1_{\mathbb{F}}$ and $E^{(i,j)}_{k,l} = 0_{\mathbb{F}}$ when $(k,l) \neq (i,j)$) form a basis.

Example 4.28. Given a field $\mathbb{F}$ and $A \in M_n(\mathbb{F})$, by $\mathbb{F}[A]$ we mean the ring defined in Remark 4.22 with the homomorphism $\mathbb{F} \to M_n(\mathbb{F})$ from Example 4.27; and if $\mathbb{K}$ is a subfield of $\mathbb{F}$, then by $\mathbb{K}[A]$ we mean the ring defined in Remark 4.22 with the homomorphism from Example 4.27 composed with the inclusion homomorphism $\mathbb{K} \to \mathbb{F}$.

Remark 4.29. For $n \geqslant 2$ the ring $M_n(\mathbb{F})$ is not commutative. This is easy to check for $n = 2$ where[23]

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

For $n > 2$ we get an example by embedding the matrices above: we place each one in the top left of an $n \times n$ matrix and then fill the 3rd, 4th, $\dots$, and nth row and column with 0s.

Remark 4.30. The group $U(M_n(\mathbb{F}))$ is often denoted $\mathrm{GL}_n(\mathbb{F})$, the set of matrices with non-zero determinant.

[23] ⚠As it happens we *can* have $2 = 0$ in a field *e.g.* $\mathbb{F}_2$, but we *cannot* have $2 = 1$ since then $1 = 0$ contradicting non-triviality of the field. This latter impossibility is what we need for this example.

5 Ideals and quotient rings

The rings $\mathbb{Z}_N$ – discussed in Example 4.1 – are examples of a very general construction. To describe this we first need a suitable generalisation of 'multiple of N': An **ideal**[24] in a ring R is a subgroup of the additive group of R closed under left and right multiplication by elements of R. The notation $I \lhd R$ is used in places (*e.g.* [Coh00, p12]) to mean I is an ideal of R.

Remark 5.1. ⚠Note the difference between ideals and subrings: an ideal is closed under multiplication by any element of the containing ring, while a subring is only closed under multiplication by elements of itself. On the other hand a subring contains 1, while an ideal doe not in general contain 1. The ring R itself is the only set that is both an ideal and a subring.

Example 5.2. Given a ring R, the sets $\{0_R\}$ and R are ideals.

Example 5.3. Given a commutative ring R and $x \in R$ the set $xR := \{xr : r \in R\}$ is an ideal. ⚠The requirement that R be commutative cannot be dropped: suppose that $\mathbb{F}$ is a field and consider the ring $M_2(\mathbb{F})$ and the matrices

$$A := \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \text{ and } P := \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

The set $AM_2(\mathbb{F})$ is a set of matrices all of which have rank at most 1, and if $AM_2(\mathbb{F})$ were an ideal then it would contain $I = A + PAP$ which has rank 2.

Proposition 5.4. *Suppose that R is a ring and $(I_i)_{i\in S}$ is collection of ideals of R (with S non-empty). Then $\bigcap_{i\in S} I_i$ is an ideal.*

Proof. The requirement that S be non-empty ensures that the intersection is well-defined. Since I_i is an (additive) subgroup of R for all $i \in S$, we have $0_R \in I_i$ and hence $0_R \in \bigcap_{i\in S} I_i$. Now, suppose $x, y \in \bigcap_{i\in S} I_i$. Then $x, y \in I_i$ for all $i \in S$, and hence $x + (-y) \in I_i$ for all $i \in S$, and $x + (-y) \in \bigcap_{i\in S} I_i$; we conclude that $\bigcap_{i\in S} I_i$ is a subgroup by the subgroup test. Finally, if $x \in \bigcap_{i\in S} I_i$ and $r \in R$ then $x \in I_i$ for all $i \in S$, and hence $xr, rx \in I_i$ for all $i \in I$ so $xr, rx \in \bigcap_{i\in S} I_i$. The result is proved. □

Given a ring R and elements $x_1, \dots, x_n \in R$ we define

$$\langle x_1, \dots, x_n \rangle := \bigcap \{I : I \text{ is an ideal in } R \text{ and } x_1, \dots, x_n \in I\}, \tag{5.1}$$

which is an ideal by the preceding proposition and Example 5.2, which ensures that there is at least some ideal containing $x_1, \dots, x_n$; we call $\langle x_1, \dots, x_n \rangle$ the **ideal generated by** $x_1, \dots, x_n$. An ideal generated by one element is called a **principal ideal**.

[24] When R is not commutative these are often called **two-sided ideals**.

Remark 5.5. ⚠The ideal generated by an element depends on the ambient ring: for example if $N \in \mathbb{N}^*$ then $\langle N \rangle = N\mathbb{Z}$ as an ideal in $\mathbb{Z}$, while $\langle N \rangle = \mathbb{Q}$ as an ideal in $\mathbb{Q}$.

Remark 5.6. If $x \in R$ is a unit then $\langle x \rangle = R$. Conversely, if R is commutative and $\langle x \rangle = R$ then there is $y \in R$ such that $xy = 1$ and hence $x \in U(R)$. ⚠Commutativity is essential for the converse: with the notation of Example 5.3, $\langle A \rangle = M_2(\mathbb{F})$, but A is *not* a unit.

Given a ring R and $A_1, \dots, A_n \subset R$ we write

$$A_1 + \cdots + A_n := \{a_1 + \cdots + a_n : a_1 \in A_1, \dots, a_n \in A_n\}.$$

Proposition 5.7. *Suppose that R is a ring and $I_1, \dots, I_n$ are ideals in R. Then $I_1 + \cdots + I_n$ is an ideal in R.*

Proof. Since $0_R \in I_i$ for all $1 \leqslant i \leqslant n$ we have $0_R \in I_1 + \cdots + I_n$. Suppose that $r \in R$ and $z, w \in I_1 + \cdots + I_n$ so that there are elements $z_i, w_i \in I_i$ for all $1 \leqslant i \leqslant n$ with $z = z_1 + \cdots + z_n$ and $w = w_1 + \cdots + w_n$. Then $rz = \sum_{i=1}^n (rz_i) \in I_1 + \cdots + I_n$ by distributivity, and similarly for zr. By commutativity and associativity of addition we have $z - w = z + (-1)w = \sum_{i=1}^n (z_i + (-1)w_i) \in I_1 + \cdots + I_n$. Hence by the subgroup test $I_1 + \cdots + I_n$ is a group, and so the result is proved. □

Remark 5.8. Suppose that R is a commutative ring and $x_1, \dots, x_n \in R$. Then

$$\langle x_1, \dots, x_n \rangle = x_1 R + \cdots + x_n R.$$

Indeed, certainly any ideal containing $x_1, \dots, x_n$ must contain the set on the right, and by Proposition 5.7 and Example 5.3 the right hand side is an ideal.

Example 5.9 (Ideals in $\mathbb{Z}$)**.** For each $N \in \mathbb{N}_0$, $\langle N \rangle = N\mathbb{Z}$ is an ideal in $\mathbb{Z}$ – this is the set of integer multiples of N.

In fact all ideals in $\mathbb{Z}$ have this form: suppose that I is a non-zero ideal in $\mathbb{Z}$ then I contains a positive element (since ideals are closed under multiplication by -1); let $N \in I$ be the minimal positive element of I. Of course $I \supset N\mathbb{Z}$; if $I \backslash N\mathbb{Z} \neq \varnothing$ then it contains a positive element and so a minimal positive element, say M. By minimality of N we have $M > N$ and of course $M - N \in I$. By minimality of M and positivity of $M - N$ we have $M - N \in N\mathbb{Z}$ whence $M \in N\mathbb{Z}$, a contradiction. It follows that $I = N\mathbb{Z}$ and the result is proved.

Remark 5.10. An integral domain in which every ideal is principal is called a **principal ideal domain** or **PID**. Example 5.9 shows that $\mathbb{Z}$ is a PID. PIDs are the central object of study of this course – roughly they capture the properties of $\mathbb{Z}$ which we are trying to generalise.

Example 5.11. The ideal $\langle 2, X\rangle$ in $\mathbb{Z}[X]$ is the set of polynomials with even constant coefficient. Certainly the polynomials with even constant coefficient form an ideal in $\mathbb{Z}[X]$ containing 2 and X, and conversely every such polynomial is in $\langle 2, X\rangle$ since it can be written in the form $2q + Xp(X)$ for some $p \in \mathbb{Z}[X]$ and constant polynomial $q \in \mathbb{Z}[X]$.

Now the ideal $\langle 2, X\rangle$ is not principal, so $\mathbb{Z}[X]$ is *not* a PID. To see this, suppose that $p \in \mathbb{Z}[X]$ is such that $\langle 2, X\rangle = \langle p\rangle$. Then there would be polynomials $q, r \in \mathbb{Z}[X]^*$ such that $X = p(X)q(X)$ and $2 = p(X)r(X)$. Since $\mathbb{Z}$ is an integral domain, Proposition 4.23 (iii) and the second of these tells us $\deg p = 0$, and then the first that $\deg q = 1$. Write $p(X) = a$ and $q(X) = bX + c$ where $a, b, c \in \mathbb{Z}$. Then $0 = (ab-1)X + ac$, and so $ab = 1$. We conclude that a is a unit. But then p is a unit in $\mathbb{Z}[X]$ and so $\langle p\rangle = \mathbb{Z}[X]$. This contradicts the fact that all polynomials in $\langle 2, X\rangle$ have even coefficients.

Kernels

Given a ring homomorphism $\phi : R \to S$, the **kernel of** ϕ is its kernel as a homomorphism of additive groups, that is the set of $r \in R$ such that $\phi(r) = 0_S$.

Remark 5.12. In particular, a ring homomorphism $\phi : R \to S$ is injective if and only if $\ker \phi = \{0_R\}$.

Proposition 5.13. *Suppose that $\phi : R \to S$ is a ring homomorphism. Then $\ker \phi$ is an ideal.*

Proof. Since ϕ is a group homomorphism the kernel is an additive subgroup of R. Now suppose $x \in \ker \phi$ and $r \in R$. Then $\phi(xr) = \phi(x)\phi(r) = 0\phi(r) = 0$ by Lemma 1.3, and similarly $\phi(rx) = 0$. It follows that $xr, rx \in \ker \phi$ so that $\ker \phi$ is an ideal. □

Remark 5.14. Given a commutative ring R we say that λ is a **root** of p if $p(\lambda) = 0$. $p(\lambda)$ is defined in Proposition 4.21 applied with the identity homomorphism $R \to R$.

Example 5.15. Suppose that R is a commutative ring and $\lambda \in R$. Then $I := \{p \in R[X] : p(\lambda) = 0\}$ is an ideal (as it is the kernel of the evaluation homomorphism $R[X] \to R; p \mapsto p(\lambda)$) and $I = \langle X - \lambda\rangle$. To see this, first note that I contains $X - \lambda$, and so $\langle X - \lambda\rangle$; and secondly, if $p \in I$ then

$$\begin{aligned} p(X) = p(X) - p(\lambda) &= \sum_{n=0}^{d} a_n(X^n - \lambda^n) \\ &= (X - \lambda)\sum_{n=1}^{d} a_n(X^{n-1} + \cdots + \lambda^{n-1}) \in \langle X - \lambda\rangle. \end{aligned}$$

This result is sometimes called the Factor Theorem.

The Factor Theorem already gives us a very useful result about integral domains.

Proposition 5.16. *Suppose that R is an integral domain and $p \in R[X]^*$ has degree d. Then p has at most d roots in R.*

Proof. We proceed by induction on d. If $d = 0$ then p is a non-zero constant and so has no roots. Now, suppose that $d > 0$ and λ is a root of p. Then there is a polynomial q such that $p(X) = (X - \lambda)q(X)$, and since R is an integral domain Proposition 4.23 (iii) applies so that $\deg q = d - 1$ and so by induction q has at most $d - 1$ roots. Since R is an integral domain, if $\lambda' \in R$ is a root of p then either $\lambda' - \lambda = 0$ or $q(\lambda') = 0$ so that λ' is a root of q. We conclude that p has at most $1 + (d - 1) = d$ roots as claimed. □

Remark 5.17. Note that if R is a non-trivial commutative ring that is not an integral domain then there are elements $a, b \in R^*$ with $ab = 0$. Then polynomial $aX \in R[X]$ then has degree 1 but at least two roots: 0 and b.

Quotient rings and the isomorphism theorems

Ideals can be used to produce equivalence relations that are compatible with the underlying ring operations in the same way as Example 4.1. This extends what happens for commutative groups.

With this we can construct quotient rings.

Theorem 5.18 (Quotient Rings). *Suppose that R is a ring and I is an ideal. Then the commutative group R/I may be endowed with a multiplication such that the quotient map q is a surjective ring homomorphism with kernel I. If R is commutative then so is R/I.*

Proof. The key is to show that $q(xy) = q(x'y')$ whenever $x+I = x'+I$ and $y+I = y'+I$. By distributivity of multiplication and negation we have that $xy - x'y' = (x - x')y + x'(y - y')$. But then $x - x' \in I$ and $y - y' \in I$ and so $xy - x'y' \in Iy + x'I \subset I$ since I is closed under multiplication by *any* element of R (in this case y on the right and x' on the left). We conclude that $q(xy) = q(x'y')$ as required, and so we may define $\widehat{\times}$ on R/I: first, for $u, v \in R/I$ let $x, y \in R$ be such that $q(x) = u$ and $q(y) = v$. Then put $u \widehat{\times} v := q(xy)$, which is well-defined.

For $u, v, w \in R/I$, let $x, y, z \in R$ be such that $u = q(x)$, $v = q(y)$ and $w = q(z)$. Then $(u \widehat{\times} v) \widehat{\times} w = q((xy)z) = q(x(yz)) = u \widehat{\times} (v \widehat{\times} w)$ so that $\widehat{\times}$ is associative. $q(1)q(x) = q(x) = q(x)q(1)$ so $q(1)$ is an identity for $\widehat{\times}$ since q is surjective. Finally, for $q(x) \in R/I$, we have $q(x) \widehat{\times} (q(y) + q(z)) = q(x(y + z)) = q(xy + xz) = q(xy) + q(xz) = q(x) \widehat{\times} q(y) + q(x) \widehat{\times} q(z)$ and since q is surjective it follows that left multiplication by $q(x)$ is a homomorphism. So is right multiplication by a similar argument, and hence (again since q is surjective) it follows that $\widehat{\times}$ distributes over addition.

Finally, we have seen that $q(1)$ is the identity; q is a homomorphism of the additive by definition of the quotient group; and q is multiplicative by definition. Thus q is a homomorphism. Moreover, $\widehat{\times}$ is commutative if the multiplication on R is commutative. The result is proved. □

Remark 5.19. Given an ideal I of a ring R, we have

$$(x+I) +_{R/I} (y+I) = (x+y)+I, -(x+I) = (-x)+I, \text{ and } 0_{R/I} = I;$$

and

$$(x+I) \times_{R/I} (y+I) = (xy)+I \text{ and } 1_{R/I} = 1+I,$$

and if $x \in U(R)$ then $x+I \in U(R/I)$ and $(x+I)^{-1} = x^{-1}+I$. ⚠Not every unit in R/I is the image of a unit as can be seen by considering $R = \mathbb{Z}$ and $I = \langle N \rangle$ for $N = 5$ or $N > 6$.

Remark 5.20. If $R = \mathbb{Z}$ and $I = \langle N \rangle = N\mathbb{Z}$ then R/I is the ring $\mathbb{Z}_N$. In the light of this we generalise the notation for modular arithmetic: if R is a ring and I is an ideal in R then we write $x \pmod{I}$ in place of $x+I$ or $q(x)$ (where q is as in Theorem 5.18). The intuition here is that quotient ring R/I is the ring R with the elements of I 'set to zero'.

Example 5.21. Suppose that $\mathbb{F}$ is a field and that I is an ideal in $\mathbb{F}$. Then the map $q : \mathbb{F} \to \mathbb{F}/I$ is a ring homomorphism with kernel I and so by Proposition 2.4 either $\mathbb{F}/I$ is trivial *i.e.* $I = \mathbb{F}$; or this homomorphism is injective and so $I = \{0\}$. It follows that for fields the only two ideals are the whole field and the zero ideal *c.f.* Example 5.2.

⚠The converse is not true, for example the ring of Exercise I.9, called the **quaternions**, is a non-commutative ring (and so in particular not a field) with only two ideals.

Theorem 5.22 (First Isomorphism Theorem)**.** *Suppose that $\phi : R \to S$ is a ring homomorphism. Then* $\operatorname{Im}\phi$ *is a subring of S;* $\ker\phi$ *is an ideal in R; and the map*

$$\psi : R/\ker\phi \to S; x + \ker\phi \mapsto \phi(x)$$

is a well-defined injective ring homomorphism.

Proof. The first conclusion is Proposition 2.9, but it is perhaps clearer to say it is by the subring test and Lemma 2.2; the second is Proposition 5.13, but it is also perhaps easier to just say by Lemma 1.3. With this Theorem 5.18 tells us $R/\ker\phi$ is a ring.

Now, $x + \ker\phi = y + \ker\phi$ if and only if $x - y \in \ker\phi$ which is true if and only if $0 = \phi(x-y) = \phi(x) - \phi(y)$, which in turn is true if and only if $\phi(x) = \phi(y)$. It follows that ψ is a well-defined injection. ψ is a homomorphism of the additive group since

$$\begin{aligned}\psi((x+\ker\phi)+(y+\ker\phi)) &= \psi((x+y)+\ker\phi)\\ &= \phi(x+y) = \phi(x)+\phi(y) = \psi(x+\ker\phi)+\psi(y+\ker\phi);\end{aligned}$$

and moreover

$$\begin{aligned}\psi((x+\ker\phi)(y+\ker\phi)) &= \psi((xy)+\ker\phi)\\ &= \phi(xy)=\phi(x)\phi(y)=\psi(x+\ker\phi)\psi(y+\ker\phi),\end{aligned}$$

and $\psi(1+\ker\phi)=\phi(1)=1$. The result is proved. □

We turn to some applications.

Example 5.23. The First Isomorphism Theorem applied to the ring homomorphism $R\to R; x\mapsto x$ gives the isomorphism $R/\{0\}\cong R$.

Example 5.24. Suppose that R is a commutative ring and $\lambda\in R$. Then $R[X]/\langle X-\lambda\rangle$ is isomorphic to R by applying the First Isomorphism Theorem to the evaluation homomorphism $R[X]\to R; p\mapsto p(\lambda)$.

Proposition 5.25. *Suppose that R is an integral domain of characteristic $p\neq 0$. Then p is prime and the additive group of R is a vector space over $\mathbb{F}_p$ in such a way that multiplication on R is bilinear.*

Proof. Let $\psi:\mathbb{Z}\to R$ be the homomorphism of Proposition 3.7, and suppose that R has characteristic p. If $p=ab$ for $a,b\geqslant 1$ then $0_R=\psi(p)=\psi(a)\psi(b)$, and since R is an integral domain we conclude that $\psi(a)=0$ or $\psi(b)=0$; say the former. Then by definition $a\geqslant p$ and so $a=p$ and $b=1$. We conclude that p is prime.

The kernel of ψ contains p and is an ideal in $\mathbb{Z}$. Since $\mathbb{Z}$ is a PID it has the form $\langle N\rangle$ for some $N\in\mathbb{N}_0$, but then $N\mid p$, whence $N=1$ or $N=p$. If $N=1$ then $1_R=\psi(1)=\psi(0)=0_R$ contradicting the non-triviality of R. We conclude that $N=p$ and the ring $\mathbb{Z}/\langle p\rangle$ is the ring $\mathbb{F}_p$ which is a field (Example 4.3). By the First Isomorphism Theorem there is then an injective ring homomorphism $\mathbb{F}_p\to R$ and so R has the structure of a vector space over $\mathbb{F}_p$ as described by Proposition 2.11.

□

Remark 5.26. For a finite field $\mathbb{F}$ the homomorphism $\psi:\mathbb{Z}\to\mathbb{F}$ (from Proposition 3.7) cannot be injective and so the kernel contains a non-zero, and hence positive element so the characteristic is non-zero and hence by the above prime. It follows from this that every finite field has order p^n for some prime p and $n\in\mathbb{N}^*$. In Exercise II.9 it is shown that there is a field of order p^n for every prime p and $n\in\mathbb{N}^*$.

Remark 5.27. Integral domains of characteristic 0 need not be vector spaces – *e.g.* $\mathbb{Z}$. If $\mathbb{Z}$ were a vector space over a field $\mathbb{F}$ with scalar multiplication denoted $\lambda.z$ for $\lambda\in\mathbb{F}$, $z\in\mathbb{Z}$, then we have two cases: $\mathbb{F}$ has characteristic 2, so $0=(1_{\mathbb{F}}+1_{\mathbb{F}}).1=1+1$, a contradiction; or $\mathbb{F}$ has characteristic greater than 2, and there is $\lambda\in\mathbb{F}$ such that $(1_{\mathbb{F}}+1_{\mathbb{F}})\lambda=1_{\mathbb{F}}$, and $1=(1_{\mathbb{F}}+1_{\mathbb{F}}).(\lambda.1)=2(\lambda.1)$, another contradiction since there is no integer with this property.

Example 5.28. In Example 3.10 we saw that $\mathbb{Z}$ had no proper subrings. On the other hand, if R is a ring with no proper subrings then the image of the unique ring homomorphism $\phi : \mathbb{Z} \to R$ afforded by Proposition 3.7 must be R (since the image is a subring by Proposition 2.9) and so by the First Isomorphism Theorem R is isomorphic to a quotient of $\mathbb{Z}$. In Example 5.9 we saw every ideal of $\mathbb{Z}$ has the form $\langle N \rangle$ for $N \in \mathbb{N}_0$ and so that a ring with no proper subrings must be isomorphic to $\mathbb{Z}$ (recall Example 5.23) or $\mathbb{Z}_N$ for $N \in \mathbb{N}^*$. A short check confirms that these rings really do not have any proper subrings (and the existence of proper subrings is a property that is preserved by isomorphisms).

One can further test our understanding of rings by asking which (commutative) rings have exactly one or two proper subrings, and this has been investigated in [ZD16].

Given an ideal I in R we write $\mathrm{Ideals}_I(R)$ for the set of ideals in R containing I, and $\mathrm{Ideals}(R)(= \mathrm{Ideals}_{\{0\}}(R))$ for the set of ideals of R.

Theorem 5.29 (Relationship between ideals in R and R/I)**.** *Suppose that R is a ring and I is an ideal in R. Then the map*

$$\phi : \mathrm{Ideals}_I(R) \to \mathrm{Ideals}(R/I); I' \mapsto \{x + I : x \in I'\}.$$

is a well-defined inclusion-preserving bijection.

Proof. First, we show the map is well-defined. Suppose that $I' \in \mathrm{Ideals}_I(R)$, and $S, T \in \phi(I')$. Then there are elements $x, y \in I'$ such that $S = x + I$ and $T = y + I$ so $S + (-T) = (x + I) + ((-y) + I) = (x + (-y)) + I \in \phi(I')$. Since $\phi(I')$ is non-empty, the subgroup test tells us that $\phi(I')$ is an additive subgroup of R/I. Furthermore, if $x + I \in R/I$ and $y + I \in \phi(I')$ for some $y \in I'$ then $xy, yx \in I'$ and so $(x + I) \times (y + I) = (xy) + I \in \phi(I')$ and $(y + I) \times (x + I) = (yx) + I \in \phi(I')$. Thus $\phi(I')$ is genuinely an ideal in R/I.

ϕ is visibly inclusion-preserving; it is an injection since $I' = \bigcup \{x + I : x + I \in \phi(I')\}$ in view of the fact that $I \subset I'$.

Finally, if $J \in \mathrm{Ideals}(R/I)$ then put $I' := \bigcup_{K \in J} K$. $I \subset I'$ since $I = 0_{R/I} \in J$. If $x, y \in I'$ then $x + I, y + I \in J$ and so $(x + (-y)) + I \in J$ (since J is an additive group) and hence $x + (-y) \in I'$. It follows that I' is an additive group by the subgroup test. If $x \in R$ and $y \in I'$ then $(x + I) \times (y + I) \in J$ and so $(xy) + I \in J$ and $xy \in I'$, and similarly $yx \in I'$ so we see that I' is an ideal. Moreover $\phi(I') = J$, and ϕ is a surjection. □

This result also goes by the name of the Correspondence Theorem and sometimes the Fourth Isomorphism Theorem for rings.

6 Proper, prime, and maximal ideals

Suppose that R is a commutative ring. We say that an ideal I in R is **proper** if $I \neq R$, and have the following immediate fact.

Lemma 6.1. *Suppose that R is a commutative ring and I is an ideal in R. Then I is proper if and only if R/I is non-trivial.*

Remark 6.2. If R is a non-trivial ring then R always has a proper ideal – $\{0_R\}$ – while we saw in Example 3.10 that there are non-trivial rings (*e.g.* $\mathbb{Z}$) with no proper subrings.

We say that an ideal I is **prime** if it is proper and whenever $ab \in I$ we have either $a \in I$ or $b \in I$.

Proposition 6.3. *Suppose that R is a commutative ring and I is an ideal in R. Then I is a prime if and only if R/I is an integral domain. In particular R is an integral domain if and only if $\{0_R\}$ is prime.*

Proof. For 'only if' we have $(a+I)(b+I) = 0_{R/I} = I$, so $ab \in I$ and therefore $a \in I$ or $b \in I$ by primality. Consequently $a + I = I = 0_{R/I}$ or $b + I = I = 0_{R/I}$ *i.e.* R/I is an integral domain. (R/I is non-trivial since I is proper.) In the other direction, I is proper since R/I is non-trivial, and if $ab \in I$ then $(a+I)(b+I) = 0_{R/I}$, and $a + I = 0_{R/I} = I$ or $b + I = 0_{R/I} = I$. We conclude $a \in I$ or $b \in I$ as required. □

Example 6.4. The ideal $\langle X\rangle$ is prime in $R[X]$ when R is an integral domain. To see this, suppose that $p(X)q(X) \in \langle X\rangle$. We can write $p(X) = a + Xg(X)$ and $q(X) = b + Xh(X)$ for $g, h \in R[X]$, and hence $ab + X(g(X) + h(X) + Xg(X)h(X)) = Xr(X)$ for some $r \in R[X]$. It follows that $ab = 0_R$ and since R is an integral domain either $a = 0_R$ and so $p \in \langle X\rangle$, or $b = 0_R$ and $q \in \langle X\rangle$.

We say that an ideal I is **maximal** if I is proper and whenever $I \subset J \subset R$ for some ideal J we have $J = I$ or $J = R$.

Remark 6.5. ⚠Maximal here is maximal with respect to inclusion amongst *proper* ideals; all ideals in R are contained in R.

Proposition 6.6. *Suppose that R is a commutative ring and I is an ideal in R. Then I is maximal if and only if R/I is a field. In particular R is a field if and only if $\{0\}$ is maximal.*

Proof. Suppose that R/I is a field. Then R/I is non-trivial and so I is proper; suppose J is an ideal with $I \subsetneq J \subset R$. Then there is $x \in J\backslash I$ and since R/I is a field some $y \in R$ such that $xy + I = 1 + I$ whence $1 \in xR + I \subset J$ and so $J = R$, whence I is maximal as claimed.

Conversely, if I is maximal and $x \in R$ has $x + I \neq I$ then $I + xR$ is an ideal properly containing I and so by maximality equals R. It follows that there is some $y \in R$ such that $1 \in xy + I$ whence $(x+I)(y+I) = 1_{R/I}$ so that $U(R/I) = (R/I)^*$ and R/I is a field as required. (R/I is non-trivial as I is proper.) □

Example 6.7. Given a prime p then $\langle p\rangle$ is maximal in $\mathbb{Z}$ since $\mathbb{Z}/\langle p\rangle$ is a field (as we saw in Example 4.3).

Remark 6.8. It follows immediately from this and Proposition 6.3 that every maximal ideal is prime, but this can also be proved directly. ⚠Although it will turn out that in PIDs all non-zero prime ideals are maximal (essentially Proposition 7.11), this is not true in general *e.g.* $\langle X\rangle$ in $\mathbb{Z}[X]$ is prime, and properly contained in the proper ideal $\langle 2, X\rangle$.

It is not immediately obvious that a non-trivial commutative ring, R, should have a maximal proper ideal. If R is finite then we might proceed iteratively: note that $\{0\}$ is a proper ideal (since R is non-trivial). Suppose we have constructed some proper ideal I. If this is maximal then stop; if not then there is some proper ideal strictly containing I. In the second case replace I by this new ideal. The new ideal is strictly larger, and since R is finite this process must terminate.

If R is infinite this process might not terminate, but we still have the intuition that we should be able to keep going until we exhaust all the element of R. This intuition can be formalised through a transfinite induction, but the conclusion (in a slightly generalised form which follows) is more commonly established via Zorn's Lemma following [Zor35].

Theorem 6.9 (Krull's Theorem)**.** *Suppose that R is a commutative ring and I is a proper ideal in R. Then there is a maximal ideal J in R containing I.*

We shall not prove this here, though it is not particularly involved. In fact we could take it an an axiom – it is known to be equivalent to the Axiom of Choice or Zorn's Lemma [Hod79].

7 Divisibility

Suppose that R is a commutative ring. Principal ideals capture a notion of divisibility: we say that a **divides** b or b is a **multiple** of a, and write $a \mid b$ if any of the following equivalent properties holds:

$$b \in \langle a\rangle;\ \text{ or } \langle b\rangle \subset \langle a\rangle;\ \text{ or there is some } x \in R \text{ such that } b = xa.$$

⚠While the first two of these are equivalent even if the ring is not commutative, returning to Example 5.3 we see that the third is not equivalent to the others: In the ring $M_2(\mathbb{F})$, $\langle A\rangle = \langle I\rangle$, but for reasons of rank I is not a product of A with some other element.

Remark 7.1. The structure of ideals means that if $a \mid b_i$ for all $1 \leqslant i \leqslant n$, and $r_1, \dots, r_n \in R$ then $a \mid b_1r_1 + \cdots + b_nr_n$.

We say that a and b are **associates** and write $a \sim b$ if $\langle a\rangle = \langle b\rangle$.

Proposition 7.2. *Suppose that R is a commutative ring. Then $\mid$ is reflexive and transitive, and if $x \mid x'$ and $y \mid y'$ then $xy \mid x'y'$. Hence $\sim$ is an equivalence relation, and if $x \sim x'$ and $y \sim y'$ then $xy \sim x'y'$. Furthermore, $x \sim 0$ if and only if $x = 0$, and $x \sim 1$ if and only if $x \in U(R)$.*

Proof. Reflexivity and transitivity follow immediately from the corresponding facts for subset inclusion. If $x \mid x'$ and $y \mid y'$ then there are elements $a, b \in R$ such that $x' = ax$ and $y' = by$ so $x'y' = (ab)(xy)$, and $xy \mid x'y'$.

Furthermore, $0 \sim 0$, and if $x \sim 0$ then there is $a \in R$ such that $x = 0a = 0$. If $x \in U(R)$ then $xx^{-1} = 1$ so $x \mid 1$, and $x = 1x$ so $x \sim 1$; and if $x \sim 1$ then there is $a \in R$ such that $1 = xa$ and hence $x \in U(R)$. □

Remark 7.3. ⚠Ideals depend on the ambient ring and so do $\mid$ and $\sim$ *e.g.* $2 \nmid 3$ in $\mathbb{Z}$, but $2 \mid 3$ in $\mathbb{Q}$.

Remark 7.4. For $r \in R$ and $p(X) = a_0 + a_1X + \cdots + a_dX^d \in R[X]$ we have $r|p$ in $R[X]$ if and only if $r \mid a_i$ in R for all $0 \leqslant i \leqslant d$ by equating coefficients.

We say that c is a **common divisor** of a and b if $c \mid a$ and $c \mid b$, and d is a **greatest common divisor (gcd)** if it is a common divisor, and every common divisor of a and b is a divisor of d. It follows immediately that if d and d' are gcds of a and b then $d \sim d'$.

Remark 7.5. All of this terminology coincides with its usual meaning in $\mathbb{Z}$.

Proposition 7.6. *Suppose that R is commutative ring in which every ideal is principal. Then every pair $a, b \in R$ has a greatest common divisor.*

Proof. Since every ideal in R is principal there is some $d \in R$ such that $\langle a, b\rangle = \langle d\rangle$, and d is a common divisor of a and b. Now if c is a common divisor of a and b, then $a, b \in \langle c\rangle$ and so $\langle d\rangle = \langle a, b\rangle \subset \langle c\rangle$ as required. □

We say that an element $x \in R$ is **prime** if $\langle x\rangle$ is a prime ideal; in other notation if ($x \not\sim 1$ and) $x \mid ab$ implies $x \mid a$ or $x \mid b$. In particular, if R is an integral domain then Example 6.4 tells us that X is prime in $R[X]$.

Remark 7.7. By induction, given a prime x and a finite list of elements $(y_i)_{i\in I}$ such that $x \mid \prod_{i\in I} y_i$, there is some $i \in I$ such that $x \mid y_i$.

Proposition 7.8. *Suppose that R is an integral domain and $r \in R$ is prime as an element of R. Then r is also prime as an element of $R[X]$.*

Proof. Suppose that $p(X) = a_0 + a_1X + \cdots + a_nX^n$ and $q(X) = b_0 + b_1X + \cdots + b_mX^m$ are such that $r \mid pq$ in $R[X]$ and $r \nmid p$ in $R[X]$ so that there is some minimal $k \in \mathbb{N}_0$ such that $r \nmid a_k$ in R. Suppose that $l \geqslant 0$ and that we have shown $r \mid b_j$ in R for all $j < l$. The coefficient of X^{k+l} in pq is

$$\sum_{j=0}^{k+l} a_jb_{k+l-j} = \sum_{j=0}^{k-1} a_jb_{k+l-j} + a_kb_l + \sum_{j=0}^{l-1} a_{k+l-j}b_j.$$

r divides the left hand side (in R) by hypothesis; it divides the first summand on the right (in R) since $r \mid a_i$ in R for all $0 \leqslant i < k$ by minimality of k; and it divides the last summand (in R) since $r \mid b_j$ in R for all $0 \leqslant j < l$ by the inductive hypothesis. It follows that $r \mid a_k b_l$ in R. But r is prime in R and $r \nmid a_k$ in R by hypothesis, so we conclude $r \mid b_l$ in R. Thus by induction $r \mid b_l$ in R for all $l \in \mathbb{N}_0$ so that $r \mid q$ in $R[X]$ as required. □

Remark 7.9. ⚠Note that primality is not in general preserved on passage from a subring to a ring: every integral domain is a subring of a field and the only prime in a field is 0.

We say that $x \in R$ is **irreducible** if whenever $a \mid x$ we have $a \sim x$ or $a \sim 1$ but *not* both[25]. This is equivalent to saying that $\langle x \rangle$ is maximal amongst proper *principal* ideals.

Irreducible elements can behave in unexpected ways, for example 3 is irreducible in $\mathbb{Z}_6$ but $3^2 = 3$ in that ring. The next lemma is useful for showing that irreducible elements behave better in integral domains.

Lemma 7.10 (Cancellation). *Suppose that R is an integral domain, $w \mid z$ (and z non-zero), and $xz \mid yw$. Then $x \mid y$, and in particular, if $z \sim w$ (are both non-zero) then $xz \sim yw$ if and only if $x \sim y$.*

Proof. Since $w \mid z$ and $xz \mid yw$ there are elements a and b such that $z = aw$ and $bxz = yw$ so $bxaw = yw$ and since w is not a zero-divisor right multiplication by w is injective and so $(ba)x = bxa = y$ and $x \mid y$. □

Proposition 7.11. *Suppose that R is an integral domain and $x \in R^*$ is prime. Then x is irreducible.*

Proof. Suppose that $x \in R^*$ is prime (so that $x \not\sim 1$) and $a \mid x$. Then there is $b \in R$ such that $x = ab$. By primality either $x \mid a$ and so $x \sim a$ and we are done; or $x \mid b$ so that $ax \mid ab = x$, and by cancellation $a \mid 1$ since $x \in R^*$, ensuring $a \sim 1$. □

Remark 7.12. Exercise II.3 gives examples to show that even in integral domains, irreducible elements need not be primes.

Remark 7.13. Note that 0_R is always prime in an integral domain R, but it is irreducible if and only if $\langle x \rangle = R$ for all $x \in R^*$, which is true if and only if R is a field.

Proposition 7.14. *Suppose that R is an integral domain such that every pair of elements has a greatest common divisor and $x \in R$ is irreducible. Then x is prime.*

Proof. We show that if $x \mid ab$ has $x \nmid a$, then $x \mid b$. If $b = 0$ then $x \mid b$ as required, so we may suppose $b \in R^*$. By hypothesis xb and ab have a gcd, call it c. Since $b \mid xb$ and $b \mid ab$ we have $b \mid c$, so that $c = db$ for some $d \in R$. Since $db = c \mid xb$ and $db = c \mid ab$, by cancellation we have $d \mid x$ and $d \mid a$. Irreducibility of x tells us that either $d \sim x$ or $d \sim 1$; we cannot have the former since $d \mid a$, but $d \sim x \nmid a$. Hence $d \sim 1$ and so $d \in U(R)$ and

[25] ⚠Note, in particular, that units are *not* irreducible since if x is a unit then $x \sim 1$.

$d^{-1}c = b$; in particular, $c \mid b$. But then x is a common factor of xb and ab and so $x \mid c \mid b$ as required. □

Remark 7.15. Usually a positive integer is said to be prime if it is irreducible in the sense of this section. Since $\mathbb{Z}$ is a PID it follows by Propositions 7.6, 7.11 and 7.14 that a positive integer is prime in the usual sense if and only if it is prime in the sense of this section, and there is no conflict in nomenclature.

Primes are particularly important because they ensure a uniqueness of factorisation. To be precise a (possibly empty) vector $(x_1, \ldots, x_r)$ is a **factorisation** of an element x if $x \sim x_1 \cdots x_r$; the x_is are called the **factors**, and if all the factors are irreducible then we say that x has a **factorisation into irreducibles**. We say that a factorisation $(x_1, \ldots, x_r)$ of x into irreducibles is **unique** if whenever $(y_1, \ldots, y_s)$ is factorisation of x into irreducibles there is a bijection $\pi : \{1, \ldots, r\} \to \{1, \ldots, s\}$ such that $x_i \sim y_{\pi(i)}$ for all $1 \leqslant i \leqslant r$.

Remark 7.16. ⚠In particular, every unit has a unique factorisation into irreducibles.

Proposition 7.17. *Suppose that R is an integral domain and $x \in R^*$ has a (possibly empty) factorisation in which every factor is prime. Then x has a unique factorisation into irreducibles.*

Proof. Let $(x_1, \ldots, x_r)$ be a factorisation of x in which every factor is prime. Since $x \in R^*$, we have $x_1, \ldots, x_r \in R^*$, and so by Proposition 7.11 we have that x has a factorisation into irreducibles. We shall prove that if $(y_i)_{i \in I}$ are irreducible elements indexed by a finite set I such that $x \sim \prod_{i \in I} y_i$ then there is a bijection $\pi : \{1, \ldots, r\} \to I$ such that $x_i \sim y_{\pi(i)}$ for all $1 \leqslant i \leqslant r$.

We proceed by induction on r. For $r = 0$ we have $\prod_{i \in I} y_i \sim 1$ (by definition of the empty product) and so there is $u \in U(R)$ such that $\prod_{i \in I} y_i = u$. Hence for all $j \in I$, we have $y_j \left(u^{-1} \prod_{i \in I \setminus \{j\}} y_i\right) = 1$ and $y_j \in U(R)$. It follows that I is empty since no unit is irreducible, and we have the base case.

Now, suppose that $r > 0$. Then x_r is prime and $x_r \mid \prod_{i \in I} y_i$ whence there is some $j \in I$ such that $x_r \mid y_j$. But y_j is irreducible and $x_r \not\sim 1$ and so $x_r \sim y_j$. But then $x_1 \cdots x_{r-1} \sim \prod_{i \in I \setminus \{j\}} y_i$ by cancellation, and by the inductive hypothesis there is a bijection $\tilde{\pi} : \{1, \ldots, r-1\} \to I \setminus \{j\}$ such that $x_i \sim y_{\tilde{\pi}(i)}$ for all $1 \leqslant i \leqslant r-1$. Extend this to a bijection $\{1, \ldots, r\} \to I$ by setting $\pi(r) = j$ and the result is proved. □

We turn now to the problem of finding factorisations into irreducibles (Proposition 7.14 will then turn these into factorisations in which every factor is prime for use in Proposition 7.17).

We say that a commutative ring R has the **ascending chain condition on principal**

ideals[26] or **ACCP** if for every sequence $(d_n)_{n=0}^\infty$ of elements of R with $d_{n+1} \mid d_n$ for all $n \in \mathbb{N}_0$, there is some $N \in \mathbb{N}_0$ such that $d_n \sim d_N$ for all $n \geqslant N$. The idea this captures is that we cannot 'keep dividing indefinitely'.

Proposition 7.18. *Suppose that R is a PID. Then R has the ACCP.*

Proof. Suppose that $(d_n)_{n=0}^\infty$ has $d_{n+1} \mid d_n$ for all $n \in \mathbb{N}_0$ and let

$$I := \{r \in R : d_n \mid r \text{ for some } n \in \mathbb{N}_0\}.$$

This is an ideal: if $r, s \in I$ then there are $n, m \in \mathbb{N}_0$ such that $d_m \mid r$ and $d_n \mid s$, but $d_{m+n} \mid d_n \mid r$ and $d_{n+m} \mid d_m \mid s$ so $d_{n+m} \mid r - s$ and $r - s \in I$; if $r \in I$ and $s \in R$ then there is $n \in \mathbb{N}_0$ such that $d_n \mid r$ so $d_n \mid rs$ and hence $rs, sr \in I$; and finally $0 \in I$.

Since R is a PID there is some $d \in I$ such that $I = \langle d \rangle$. Since $d \in I$ there is some $N \in \mathbb{N}_0$ such that $d_N \mid d$, but then $d_n \in I$ for all $n \in \mathbb{N}_0$ and so $d_N \mid d \mid d_n$ for all $n \in \mathbb{N}_0$ and hence $d_n \sim d_N$ for all $n \geqslant N$. The result is proved. □

Proposition 7.19. *Suppose that R is an integral domain with the ACCP. Then every $x \in R^*$ has a factorisation into irreducibles.*

Proof. Write $\mathcal{F}$ for the set of elements in R^* that have factorisation into irreducibles so that all units and irreducible elements are in $\mathcal{F}$. $\mathcal{F}$ is closed under multiplication, by design and since R is an integral domain.

Were $\mathcal{F}$ not to be the whole of R^* then there would be some $x_0 \in R^* \backslash \mathcal{F}$. Now create a chain iteratively: at step i suppose we have $x_i \in R^* \backslash \mathcal{F}$. Since x_i is not irreducible and not a unit there is $y_i \mid x_i$ with $y_i \not\sim 1$ and $y_i \not\sim x_i$; let $z_i \in R^*$ be such that $x_i = y_i z_i$. If $z_i \sim x_i$, then $z_i \sim y_i z_i$ and by cancellation $1 \sim y_i$, a contradiction. We conclude $y_i, z_i \not\sim x_i$.

Since $\mathcal{F}$ is closed under multiplication we cannot have both y_i and z_i in $\mathcal{F}$. Let $x_{i+1} \in \{y_i, z_i\}$ such that $x_{i+1} \notin \mathcal{F}$; by design $x_{i+1} \mid x_i$ and $x_{i+1} \not\sim x_i$. This process produces a sequence $\cdots \mid x_2 \mid x_1 \mid x_0$ in which $x_i \not\sim x_{i+1}$ for all $i \in \mathbb{N}_0$ contradicting the ACCP. □

Remark 7.20. Integral domains in which every non-zero element has a factorisation into irreducibles are called **factorisation domains** or **atomic domains**. There are factorisation domains not having the ACCP but these are not easy to construct; the first example was given by Grams in [Gra74].

Finally, a **unique factorisation domain** or **UFD** is an integral domain in which every $x \in R^*$ has a unique factorisation into irreducibles.

[26] The reason for the name is that it can also be formulated as saying if $(I_i)_{i \in \mathbb{N}_0}$ is an ascending chain (meaning $I_i \subset I_{i+1}$ for all $i \in \mathbb{N}_0$) of principal ideals then there is some $N \in \mathbb{N}_0$ such that $I_n = I_N$ for all $n \geqslant N$.

Theorem 7.21. *Suppose that R is a PID. Then R is a UFD.*

Proof. By Propositions 7.18 and 7.19 we have that every $x \in R^*$ has a factorisation into irreducibles. But then every irreducible is prime by Propositions 7.6 and 7.14. The result then follows by Proposition 7.17. □

Remark 7.22. In particular, since $\mathbb{Z}$ is a PID the above gives the Fundamental Theorem of Arithmetic.

Remark 7.23. $\mathbb{Z}[X]$ is an example of a UFD that is not a PID; see Exercise II.8 for details.

The division algorithm and Euclidean domains

A **Euclidean function** on R is a function $f : R^* \to \mathbb{N}_0$ such that if $a, b \in R^*$ then either $b \mid a$; or there are $q \in R$, $r \in R^*$ such that $a = bq + r$ and $f(r) < f(b)$. We say that an integral domain R is a **Euclidean domain** if R supports at least one Euclidean function.

Remark 7.24. ⚠There are some variations on the definition of Euclidean function. Sometimes (*e.g.* [Gal13, p337]) a Euclidean function f is required to have the additional property that $f(a) \leqslant f(ab)$ for all $a, b \in R^*$. (Exercise E.6 asks for a proof of this.)

On [Kea98, p17] Keating uses an even stronger definition of Euclidean function f requiring that $f(ab) = f(a)f(b)$ whenever $a, b \in R^*$. This is a genuinely stronger definition, meaning there are Euclidean domains in our sense but not in the sense of Keating, though this is a recent discovery: [CNT19, Theorem 1.3]. We do *not* assume this stronger property though many of our Euclidean functions will happen to satisfy it.

Example 7.25. Suppose that $\mathbb{F}$ is a field and let $f : \mathbb{F}^* \to \mathbb{N}_0$ be any function. Since every two non-zero units divide each other in a field, f is a Euclidean function for $\mathbb{F}$ and so $\mathbb{F}$ is a Euclidean domain. The function $f : \mathbb{F}^* \to \mathbb{N}_0$ taking all non-zero elements of $\mathbb{F}$ to 1 is a Euclidean function in Keating's sense from Remark 7.24 and is perhaps a slightly more natural choice.

Example 7.26 (Division algorithm for $\mathbb{Z}$)**.** If $a, b \in \mathbb{Z}^*$ and $b \nmid a$ then let bq be (one of) the multiple(s) of b nearest to a. Then $r := a - bq$ has $|r| < |b|$, and $|\cdot|$ is a Euclidean function on $\mathbb{Z}$.

⚠Note that there were *two* choices for bq and hence for r in the case that $b \nmid a$.

Example 7.27 (Division algorithm for $\mathbb{F}[X]$)**.** Suppose that $\mathbb{F}$ is a field and $a, b \in \mathbb{F}[X]^*$. Let $\mathcal{P} := \{a + bq : q \in \mathbb{F}[X]\}$, and note that if $b \nmid a$ then $\mathcal{P}$ does not include the zero polynomial.

If $b \nmid a$, we let $r \in \mathcal{P}$ be a polynomial of minimal degree. If $\deg r \geqslant \deg b$, then we may let λ be the ratio of the lead coefficient of r to that of b and note that $r(X) - \lambda X^{\deg r - \deg b} b(X) \in$

$\mathcal{P}$ and has strictly smaller degree than r, a contradiction. It follows that $\deg r < \deg b$ and deg is a Euclidean function for $\mathbb{F}[X]$.

Remark 7.28. Suppose that f is a Euclidean function on an integral domain R such that $f(a) \leqslant f(ab)$ for all $a, b \in R^*$, and for all $a, b \in R^*$ either $b \mid a$ or there is a unique pair $(q, r) \in R \times R^*$ with $a = bq + r$ and $f(r) < f(b)$. Then either R is itself a field or $R = \mathbb{F}[X]$ for a field $\mathbb{F}$. Exercise E.6 develops a proof of this, and in particular since $\mathbb{Z}$ is neither a field nor a polynomial ring over a field the choice mentioned in the warning in Example 7.26 was in fact necessary.

Proposition 7.29. *Suppose that R is a Euclidean domain. Then R is a PID.*

Proof. Let f be a Euclidean function on R and suppose I is a non-zero ideal. Let $x \in I$ have $f(x)$ minimal, and suppose that $y \in I$. If $y \notin \langle x \rangle$ then there is $q \in R$ and $r \in R^*$ with $y = qx + r$ and $f(r) < f(x)$ so that $r \in I$, contradicting minimality of $f(x)$. □

Remark 7.30. In particular if $\mathbb{F}$ is a field then the ring $\mathbb{F}[X]$ is a PID.

Remark 7.31. There are examples of PIDs which are not Euclidean domains, one of which is developed in Exercise III.9.

Remark 7.32. A **Dedekind-Hasse function** is a map $N : R^* \to \mathbb{N}_0$ such that whenever $a, b \in R^*$ either $b \mid a$; or there are elements $p, q \in R$, $c \in R^*$ such that $ap = bq + c$ and $N(c) < N(b)$. The definition of Euclidean function places the additional requirement that $p = 1$, so in particular any ring supporting a Euclidean function supports a Dedekind-Hasse function.

It can be shown (see *e.g.* [Cla10, Theorem 49]) that an integral domain is a PID if and only if it supports a Dedekind-Hasse function. In particular, from this point of view PIDs and Euclidean domains may not seem to very different despite Remark 7.31. The important feature of Euclidean functions is that they are often in some sense easy to compute without knowing the factorisation of an element into primes. By contrast the construction of a Dedekind-Hasse function for an arbitrary PID is usually done by letting $N(a)$ be the number of prime factors of a (well-defined by Theorem 7.21).

8 Fields and adjoining elements

A field $\mathbb{K}$ is a **field extension** of a field $\mathbb{F}$ if there is a ring homomorphism $\phi : \mathbb{F} \to \mathbb{K}$. ⚠Despite the fact we speak of $\mathbb{K}$ as a field extension of $\mathbb{F}$ without mentioning ϕ, in any given instance we will have a particular ϕ in mind.[27] Often this will just the inclusion map because $\mathbb{F}$ will be a subfield of $\mathbb{K}$. Indeed, by relabelling the elements of $\mathbb{F}$ we can *always* assume that $\mathbb{F}$ is a subfield of $\mathbb{K}$ because ring homomorphisms between fields are injective (Proposition 2.4).

Proposition 2.11 shows us how to use ϕ to endow $\mathbb{K}$ with the structure of a vector space over $\mathbb{F}$ such that multiplication is bilinear. We call the $\mathbb{F}$-dimension of $\mathbb{K}$ w.r.t. *this* vector space structure the **degree** of the field extension, denoted $|\mathbb{K} : \mathbb{F}|$.

Theorem 8.1. *Suppose that $\mathbb{F}$ is a field and $f \in \mathbb{F}[X]$ is irreducible of degree d. Then $\mathbb{K} := \mathbb{F}[X]/\langle f\rangle$ is a field extension of $\mathbb{F}$ by the map $\mathbb{F} \to \mathbb{K}; \lambda \mapsto \lambda + \langle f\rangle$, and writing $\alpha := X + \langle f\rangle$, $1_{\mathbb{K}}, \alpha, \dots, \alpha^{d-1}$ is a basis for $\mathbb{K}$ in this $\mathbb{F}$-vector space structure.*

Proof. $\mathbb{F}[X]$ is a PID (Remark 7.30) and hence the fact that $\langle f\rangle$ is maximal amongst proper principal ideals means it is maximal amongst *all* proper ideals and Proposition 6.6 tells us that $\mathbb{K} = \mathbb{F}[X]/\langle f\rangle$ is a field. The given map is formed by composing the inclusion map $\mathbb{F} \to \mathbb{F}[X]$ and the quotient map $\mathbb{F}[X] \to \mathbb{F}[X]/\langle f\rangle$ and so is a ring homomorphism, and hence a field extension.

The elements $1_{\mathbb{K}}, \alpha, \dots, \alpha^{d-1}$ are $\mathbb{F}$-independent in $\mathbb{K}$: indeed, suppose that $a_0, \dots, a_{d-1} \in \mathbb{F}$ have $0_{\mathbb{K}} = a_0.1_{\mathbb{K}} + a_1.\alpha + \dots + a_{d-1}.\alpha^{d-1}$. This says exactly that $f \mid a_0 + a_1X + \dots + a_{d-1}X^{d-1}$. If the right hand side is non-zero then it has degree strictly smaller than d; a contradiction. Hence the right is $0_{\mathbb{F}[X]}$ and so $a_0, \dots, a_{d-1} = 0_{\mathbb{F}}$ as required.

On the other hand, if $f(X) = a_0 + a_1X + \dots + a_dX^d$ then every $\beta \in \mathbb{K}$ has a polynomial $p(X) = b_0 + b_1X + \dots + b_nX^n \in \mathbb{F}[X]$ such that $\beta = p(X) + \langle f\rangle$. By the division algorithm for $\mathbb{F}[X]$ (Example 7.27), either $p \in \langle f\rangle$ (and so $\beta = 0_{\mathbb{K}}$) or there is some $q \in \mathbb{F}[X]$ and $r \in \mathbb{F}[X]^*$ with $\deg r < \deg f = d$ such that $p(X) = q(X)f(X) + r(X)$. Then $\beta = r(X) + \langle f\rangle$, and writing $r(X) = c_0 + c_1X + \dots + c_{d-1}X^{d-1}$ for $c_0, \dots, c_{d-1} \in \mathbb{F}$ we have $\beta = c_0.1_{\mathbb{K}} + c_1.\alpha + \dots + c_{d-1}.\alpha^{d-1}$, and hence $1_{\mathbb{K}}, \alpha, \dots, \alpha^{d-1}$ is a spanning set.

It follows that $1_{\mathbb{K}}, \alpha, \dots, \alpha^{d-1}$ is a basis and the result is proved. □

We say that $\mathbb{K}$ is the field $\mathbb{F}$ with the element α **adjoined**.

Remark 8.2. In view of the above it becomes important to identify irreducible polynomials in $\mathbb{F}[X]$. Every degree 1 polynomial in $\mathbb{F}[X]$ is irreducible. First $f \not\sim 1$ since $\deg f \neq 0$. Now, if $g \mid f$ then let $h \in \mathbb{F}[X]^*$ be such that $f = gh$ and $1 = \deg f = \deg g + \deg h$, so either $\deg h = 0$ so h is a unit, and $g \sim f$; or $\deg h = 1$ and so $\deg g = 0$ and so g is a unit and $g \sim 1$ as required.

[27] Appearances are a bit deceptive here because if $\mathbb{F}$ is a field extension of $\mathbb{Q}$ or $\mathbb{F}_p$ then it is so uniquely (this essentially follows from Proposition 3.7), so in fact we can identify the ring homomorphism just from the fields. However, there are fields with non-trivial automorphisms (*e.g.* $\mathbb{C} \to \mathbb{C}; z \mapsto \overline{z}$) and so $\mathbb{C}$ is a field extension of $\mathbb{C}$ in multiple ways. Tutors may discuss this when considering Exercise I.5.

For some fields these are the *only* irreducible elements of the polynomial ring. Indeed, in $\mathbb{C}[X]$ the Fundamental Theorem of Algebra tells us that every non-constant polynomial has a root in $\mathbb{C}$. We are done once we note the general fact that if $f \in \mathbb{F}[X]$ of degree $n \geqslant 2$ has a root then it is not irreducible. Indeed, if f has a root λ then $X - \lambda \mid f$, but $X - \lambda \not\sim f$ and $X - \lambda \not\sim 1$ since $\deg(X - \lambda) \neq \deg f$ and $\deg(X - \lambda) \neq \deg 1$; we conclude f is not irreducible.

While there are polynomials of degree 4 or more without roots that are not irreducible (*e.g.* $(X^2 + 1)^2$ in $\mathbb{R}[X]$), if $f \in \mathbb{F}[X]$ is non-constant of degree at most 3 and no root then it *is* irreducible. First, $f \not\sim 1$ since f is non-constant. Now, suppose $g \mid f$ has $g \not\sim 1$ and $g \not\sim f$, and write $f = gh$ for some $h \in \mathbb{F}[X]$. Since $g \not\sim 1$ we have that $\deg g \neq 0$ and since $g \not\sim f$ we have $\deg h \neq 0$. Since $\deg g + \deg h \leqslant 3$ it follows that $\deg g = 1$ or $\deg h = 1$; in the former case there is $a \in \mathbb{F}^*$ and $b \in \mathbb{F}$ such that $g(X) = aX + b$ and hence $-ba^{-1}$ is a root of g and so a root of f; and similarly in the latter.

Example 8.3. The polynomial $X^2 + 1$ is irreducible over $\mathbb{R}$ since it has no root, and hence $\mathbb{R}[X]/\langle X^2 + 1\rangle$ is a field and a 2-dimensional vector space over $\mathbb{R}$. The map $\phi : \mathbb{R}[X] \to \mathbb{C}; p \mapsto p(i)$ (as defined in Proposition 4.21) is a ring homomorphism. It is surjective as a basic property of $\mathbb{C}$ (Example 4.11). The kernel is principal (since $\mathbb{R}[X]$ is a PID) and contains $X^2 + 1$ by definition of i. Since $\mathbb{C}$ (and therefore $\mathbb{R}[X]/\ker\phi$) is not trivial and $X^2 + 1$ is irreducible we conclude that $X^2 + 1$ is a generator of the kernel and by the First Isomorphism Theorem we have $\mathbb{R}[X]/\langle X^2 + 1\rangle \cong \mathbb{C}$ as rings. In fact this is one way of constructing $\mathbb{C}$.

Example 8.4. The polynomial $X^2 + X + 1$ is the only irreducible polynomial of degree 2 in $\mathbb{F}_2[X]$. Indeed, neither 0 nor 1 are roots so $X^2 + X + 1$ is irreducible. On the other hand there are only four degree 2 polynomials in $\mathbb{F}_2[X]$, and the other three are X^2, $X^2 + X$ and $X^2 + 1$ which visibly have roots of 0, 0 (and 1), and 1 respectively. Hence these are not irreducible.

The ring $\mathbb{F}_2[X]/\langle X^2 + X + 1\rangle$ is then a field of order 4 which is dentoed $\mathbb{F}_4$. ⚠This field is *not* equal to the ring $\mathbb{Z}_4$ – indeed the latter is not even an integral domain since $2^2 = 0$ but $2 \neq 0$.

Finding irreducible polynomials is somewhat like finding primes in the integers, and there are various tests for irreducibility which can help in this endeavour.

We say that $f \in \mathbb{Z}[X]$ is **primitive** if there is no prime dividing all the coefficients of f.

Remark 8.5. Note that if f is primitive and of degree 0 then f is a unit in $\mathbb{Z}[X]$ since $\mathbb{Z}$ is a UFD (and so every non-unit has a prime factor).

Theorem 8.6 (Gauss' Lemma)**.** *Suppose that $f \in \mathbb{Z}[X]$. Then f is non-constant and irreducible in $\mathbb{Z}[X]$ if and only if f is primitive and irreducible in $\mathbb{Q}[X]$.*

Proof. Suppose that f is irreducible in $\mathbb{Z}[X]$. This immediately tells us that f is primitive since if p were a prime dividing all the coefficients of f then $p \mid f$ in $\mathbb{Z}[X]$. Since $p \not\sim 1$ we conclude that $p \sim f$ (in $\mathbb{Z}[X]$) by irreducibility of f, contradicting the fact that f is non-constant.

Now, suppose that $f = gh$ for $g, h \in \mathbb{Q}[X]$. Then let $\lambda \in \mathbb{N}^*$ be minimal such that there is $q \in \mathbb{Q}^*$ with $\lambda q^{-1}g$ and qh both in $\mathbb{Z}[X]$. Suppose that $p \in \mathbb{Z}$ is prime with $p \mid \lambda$. Then p is prime as a constant polynomial in $\mathbb{Z}[X]$ and since $p \mid \lambda f = (\lambda q^{-1}g)(qh)$, we have $p \mid \lambda q^{-1}g$ or $p \mid qh$ (both in $\mathbb{Z}[X]$). The former contradicts minimality of λ directly, and the latter once we note that $(q/p)h \in \mathbb{Z}[X]$ and $(\lambda/p)(q/p)^{-1}g = \lambda q^{-1}g \in \mathbb{Z}[X]$. We conclude that λ has no prime factors and hence (since $\mathbb{Z}$ is a UFD) is a unit. Thus $q^{-1}g \mid f$ in $\mathbb{Z}[X]$ and so by irreducibility of f in $\mathbb{Z}[X]$ we conclude that either $q^{-1}g \sim 1$ or $q^{-1}g \sim f$ in $\mathbb{Z}[X]$. Hence either $g \sim 1$ in $\mathbb{Q}[X]$ or $g \sim f$ in $\mathbb{Q}[X]$ and finally, since f is non-constant we have $f \not\sim 1$ in $\mathbb{Q}[X]$ and so f is irreducible in $\mathbb{Q}[X]$.

Conversely, suppose $f \in \mathbb{Z}[X]$ is primitive and irreducible in $\mathbb{Q}[X]$. First, $f \not\sim 1$ in $\mathbb{Q}[X]$ and so f is non-constant. Suppose $g \mid f$ in $\mathbb{Z}[X]$. By irreducibility of f in $\mathbb{Q}[X]$, either $g \sim 1$ in $\mathbb{Q}[X]$ so $\deg g = 0$, and since f is primitive $g \sim 1$ in $\mathbb{Z}[X]$; or $g \sim f$ in $\mathbb{Q}[X]$, then $\deg g = \deg f$ and writing $f = gh$ for $h \in \mathbb{Z}[X]$ we have $\deg h = 0$, and since f is primitive $h \sim 1$ in $\mathbb{Z}[X]$, whence $g \sim f$ in $\mathbb{Z}[X]$. The result is proved. □

Proposition 8.7 (Eisenstein's Criterion)**.** *Suppose that $f(X) = a_nX^n + \cdots + a_1X + a_0$ is a primitive polynomial in $\mathbb{Z}[X]$ and p is a prime such that $p \mid a_i$ for all $0 \leqslant i < n$; $p \nmid a_n$; and $p^2 \nmid a_0$. Then f is irreducible in $\mathbb{Z}[X]$.*

Proof. Suppose that $f = gh$ for $g, h \in \mathbb{Z}[X]$. The quotient map $\mathbb{Z} \to \mathbb{F}_p$ and the inclusion $\mathbb{F}_p \to \mathbb{F}_p[X]$ compose to give a homomorphism $\mathbb{Z} \to \mathbb{F}_p[X]$, so there is an evaluation homomorphism $\phi : \mathbb{Z}[X] \to \mathbb{F}_p[X]$ taking X to X. In particular, note that

$$\phi(f) = \phi(g)\phi(h) \text{ and } \deg q \geqslant \deg \phi(q) \text{ whenever } \phi(q) \in \mathbb{F}_p[X]^*.$$

Since $p \mid a_i$ for all $i < n$ and $p \nmid a_n$ we have $\phi(f) \sim X^n$.

Since $X \in \mathbb{F}_p[X]$ is prime it follows that $\phi(g) \sim X^i$ and $\phi(h) \sim X^{n-i}$ (either by induction, or because $\mathbb{F}_p[X]$ is a UFD). If $i > 0$ then $\phi(g)$ has zero constant coefficient and so p divides the constant coefficient of g. a_0 is the product of the constant coefficients of g and h and since $p^2 \nmid a_0$ we conclude that p does not divide the constant coefficient of h *i.e.* $i = n$. But then $\deg g \geqslant \deg \phi(g) = n$, and $n = \deg f = \deg g + \deg h$, so $\deg h = 0$. Since f is primitive, h is then a unit and so $g \sim f$. The case $i = 0$ is handled similarly and has $g \sim 1$ □

Example 8.8. The polynomial $X^3 - 2$ is irreducible in $\mathbb{Z}[X]$ by Eisenstein's Criterion with the prime 2 since it is visibly primitive (with the lead coefficient being 1). It is non-constant and so by Gauss' Lemma is irreducible in $\mathbb{Q}[X]$. By Theorem 8.1, $\mathbb{Q}[X]/\langle X^3-2\rangle$ is a degree 3 field extension of $\mathbb{Q}$.

The reals contain a unique positive root to X^3-2 denoted element $\sqrt[3]{2}$ (this is from Prelims Analysis) so we get an evaluation homomorphism $\mathbb{Q}[X]\to\mathbb{R}; p\mapsto p(\sqrt[3]{2})$. The kernel of this map is principal (again $\mathbb{Q}[X]$ is a PID), and since $\mathbb{Q}[\sqrt[3]{2}]$ is not trivial and X^3-2 is irreducible we see that the kernel is $\langle X^3-2\rangle$. It follows from the First Isomorphism Theorem that $\mathbb{Q}[\sqrt[3]{2}]$ is isomorphic to $\mathbb{Q}[X]/\langle X^3-2\rangle$ as a ring and, in particular, it is a field.[28]

Finally, the evaluation map $\mathbb{Q}[X]\to\mathbb{R}$ above is $\mathbb{Q}$-linear w.r.t. the two $\mathbb{Q}$-vector space structures on $\mathbb{Q}[X]$ and $\mathbb{R}$ induced by Proposition 2.11 and the inclusions $\mathbb{Q}\hookrightarrow\mathbb{Q}[X]$ and $\mathbb{Q}\hookrightarrow\mathbb{R}$. Thus by the First Isomorphism Theorem for vector spaces the $\mathbb{Q}$-vector spaces $\mathbb{Q}[X]/\langle X^3-2\rangle$ and $\mathbb{Q}[\sqrt[3]{2}]$ have the same dimension and so $\mathbb{Q}[\sqrt[3]{2}]$ is a degree 3 field extension of $\mathbb{Q}$ (where the extension homomorphism is the inclusion map).

Theorem 8.9 (Tower Law). *Suppose that $\phi:\mathbb{K}\to\mathbb{L}$ and $\psi:\mathbb{F}\to\mathbb{K}$ are field extensions. Then $\phi\circ\psi:\mathbb{F}\to\mathbb{L}$ is a field extension and if either $|\mathbb{L}:\mathbb{F}|<\infty$ or $|\mathbb{L}:\mathbb{K}|,|\mathbb{K}:\mathbb{F}|<\infty$ then $|\mathbb{L}:\mathbb{F}|=|\mathbb{L}:\mathbb{K}||\mathbb{K}:\mathbb{F}|$.*

Proof. First, the composition of homomorphisms is a homomorphism so that $\phi\circ\psi$ is a field extension. Since all ring homomorphisms between fields are injective (Proposition 2.4), by relabelling we may assume that $\mathbb{F}$ is a subfield of $\mathbb{K}$ and $\mathbb{K}$ is a subfield of $\mathbb{L}$. We do this to make the notation simpler.

Let $e_1,\dots,e_n$ be a basis for $\mathbb{L}$ as a vector space over $\mathbb{K}$, and let $f_1,\dots,f_m$ be a basis for $\mathbb{K}$ as a vector space over $\mathbb{F}$. Now, for $x\in\mathbb{L}$ there are scalars $\lambda_1,\dots,\lambda_n\in\mathbb{K}$ such that $x=\lambda_1e_1+\cdots+\lambda_ne_n$, and since $f_1,\dots,f_m$ is spanning, for each $1\leqslant j\leqslant n$ there are scalars $\mu_{1,j},\dots,\mu_{m,j}\in\mathbb{F}$ such that $\lambda_j=\mu_{1,j}f_1+\cdots+\mu_{m,j}f_m$. Hence $x=\sum_{j=1}^n\sum_{i=1}^m\mu_{i,j}f_ie_j$, so by have that $(f_ie_j)_{i=1,j=1}^{m,n}$ is an $\mathbb{F}$-spanning subset of $\mathbb{K}$. Now suppose $\mu_{1,1},\dots,\mu_{m,n}\in\mathbb{F}$ are such that $\sum_{j=1}^n\sum_{i=1}^m\mu_{i,j}f_ie_j=0_{\mathbb{L}}$. Then $\sum_{j=1}^n\left(\sum_{i=1}^m\mu_{i,j}f_i\right)e_j=0_{\mathbb{L}}$, but $\sum_{i=1}^m\mu_{i,j}f_i\in\mathbb{K}$ for each $1\leqslant j\leqslant n$ and since $e_1,\dots,e_n$ are $\mathbb{K}$-linearly independent we have $\sum_{i=1}^m\mu_{i,j}f_i=0_{\mathbb{K}}$ for all $1\leqslant j\leqslant n$. But now $f_1,\dots,f_m$ are $\mathbb{F}$-linearly independent and so $\mu_{i,j}=0_{\mathbb{F}}$ for all $1\leqslant i\leqslant m$ and $1\leqslant j\leqslant n$. It follows that $(f_ie_j)_{i=1,j=1}^{m,n}$ is a basis for $\mathbb{L}$ as an $\mathbb{F}$-vector space and the result follows. □

Example 8.10. We can use the Tower Law to show that $\sqrt{2}$ is *not* a $\mathbb{Q}$-linear combination of 1, $\sqrt[3]{2}$, and $\sqrt[3]{2}^2$. Indeed, if it were then $\mathbb{Q}[\sqrt{2}]$ would be a subfield of $\mathbb{Q}[\sqrt[3]{2}]$, and the inclusions $\mathbb{Q}\hookrightarrow\mathbb{Q}[\sqrt{2}]$ and $\mathbb{Q}[\sqrt{2}]\hookrightarrow\mathbb{Q}[\sqrt[3]{2}]$ would be field extensions. By the Tower Law we would have $3=|\mathbb{Q}[\sqrt[3]{2}]:\mathbb{Q}|=|\mathbb{Q}[\sqrt[3]{2}]:\mathbb{Q}[\sqrt{2}]||\mathbb{Q}[\sqrt{2}]:\mathbb{Q}|=|\mathbb{Q}[\sqrt[3]{2}]:\mathbb{Q}[\sqrt{2}]|\times 2$, but 3 is not even.

[28] *c.f.* Example 4.10 where the fact that $\mathbb{Q}[\sqrt{2}]$ is a field is proved directly by producing inverse elements rather than through the irreducibility of the polynomial X^2-2.

9 Modules

Modules can be viewed in a variety of ways. First, we shall think of them as vector spaces with the field replaced by a ring. Concretely, suppose that R is a commutative ring. A **(left) R-module** is a commutative group, also denoted M and called the **additive group**, and a map $. : R \times M \to M; (r, x) \mapsto r.x$ such that

(M1) $1.x = x$ for all $x \in M$;

(M2) $r.(s.x) = (rs).x$ for all $r, s \in R$ and $x \in M$;

(M3) $(r + s).x = (r.x) + (s.x)$ for all $r, s \in R$ and $x \in M$;

(M4) $r.(x + y) = (r.x) + (r.y)$ for all $r \in R$ and $x, y \in M$.

The identity of M is denoted 0 (or 0_M is disambiguation is called for) and is called the **zero** of the module, and the map $.$ is called the **scalar multiplication** of the module. If the latter is clear we simply speak of the R-module M. ⚠Sometimes the scalar multiplication really does need to be spelled out. See Example III.1.

Remark 9.1. Some quick checks reveal that $0_R.x = 0_M$ and $(-1).x = -x$ for all $x \in M$.

⚠We take R to be commutative, but this is not necessary at this stage though it will be for a number of our later results.

Example 9.2 (Vector spaces)**.** Given a field $\mathbb{F}$, a vector space V is exactly a (left) $\mathbb{F}$-module, with the two notions of scalar multiplication coinciding.

Example 9.3 (Zero module)**.** For any commutative ring R the trivial group – usually denoted $\{0\}$ in this context – and the scalar multiplication defined by $r.0 := 0$ for all $r \in R$ is a module called the **zero (R-)module**.

Groups arise naturally from considering the set of bijections from a set to itself; modules arise naturally from considering the set of homomorphisms from a commutative group to itself. To understand this we shall need a few facts about algebra of groups homomorphisms of commutative groups.

By default we write $+$ for the binary operation on a commutative group; $-x$ for the additive inverse of x; and 0 for the identity of the group. As with rings we may use subscripts to disambiguate if there are multiple groups *i.e.* we may write $+_N$ for the group operation on the commutative group N.

Proposition 9.4. *Suppose that M and N are commutative groups. Then* $\mathrm{Hom}(M, N)$*, the set of group homomorphisms $M \to N$, is itself a commutative group when endowed with the operation $\widehat{+}$, called* ***pointwise addition****, and defined by $(\phi \widehat{+} \psi)(x) := \phi(x) + \psi(x)$ for all*

$x \in M$; identity $M \to N; x \mapsto 0_N$; and the inverse of a homomorphism ϕ being the map $M \to N; x \mapsto -\phi(x)$.

Suppose P is a further commutative group and $\phi \in \mathrm{Hom}(M,N)$ and $\psi \in \mathrm{Hom}(N,P)$, then $\psi \circ \phi \in \mathrm{Hom}(M,P)$; and if $\pi \in \mathrm{Hom}(M,N)$ then $\psi \circ (\phi \hat{+} \pi) = (\psi \circ \phi)\hat{+}(\psi \circ \pi)$; and if $\pi \in \mathrm{Hom}(N,P)$ then $(\psi \hat{+} \pi) \circ \phi = (\psi \circ \phi)\hat{+}(\pi \circ \phi)$.

Proof. The commutativity (and associativity) of N here is crucial for ensuring that $\hat{+}$ is well-defined: In particular, suppose that $\phi, \psi \in \mathrm{Hom}(M,N)$ then for all $x, y \in M$ we have

$$\begin{aligned}
(\phi \hat{+} \psi)(x +_M y) &= \phi(x +_M y) +_N \psi(x +_M y) \\
&= (\phi(x) +_N \phi(y)) +_N (\psi(x) +_N \psi(y)) && \text{ϕ and ψ are homomorphisms} \\
&= (\phi(x) +_N \psi(x)) +_N (\phi(y) +_N \psi(y)) && \text{associativity and commutativity of $+_N$} \\
&= (\phi \hat{+} \psi)(x) +_N (\phi \hat{+} \psi)(y). && \text{definition of $\hat{+}$}
\end{aligned}$$

It follows that $\phi \hat{+} \psi \in \mathrm{Hom}(M,N)$. Since the operation $+_N$ is associative and commutative, so is the operation $\hat{+}$. The map $M \to N; x \mapsto 0_N$ is a homomorphism because $0_N + 0_N = 0_N$, and it is an identity for $\hat{+}$ because 0_N is an identity for $+_N$. Finally, if $\phi \in \mathrm{Hom}(M,N)$ then the map $M \to N; x \mapsto -\phi(x)$ is a homomorphism because $-\phi(x +_M y) = -(\phi(x) +_N \phi(y)) = (-\phi(y)) +_N (-\phi(x)) = (-\phi(x)) +_N (-\phi(y))$ for all $x, y \in M$ since $+_N$ is associative *and* commutative, and it is an additive inverse for ϕ w.r.t. $\hat{+}$ since $-\phi(x)$ is an additive inverse for $\phi(x)$ w.r.t. $+_N$. The first part follows.

For the second the composition of homomorphisms is a homomorphism[29] says exactly that if $\phi \in \mathrm{Hom}(M,N)$ and $\psi \in \mathrm{Hom}(N,P)$, then $\psi \circ \phi \in \mathrm{Hom}(M,P)$. Now, if $\phi, \pi \in \mathrm{Hom}(M,N)$ and $\psi \in \mathrm{Hom}(N,P)$, then

$$\psi \circ (\phi \hat{+} \pi)(x) = \psi(\phi(x) + \pi(x)) = \psi(\phi(x)) + \psi(\pi(x)) = ((\psi \circ \phi)\hat{+}(\psi \circ \pi))(x)$$

by definition and the fact that ψ is a homomorphism, and we have that $\psi \circ (\phi \hat{+} \pi) = (\psi \circ \phi)\hat{+}(\psi \circ \pi)$ as claimed. On the other hand, if $\phi \in \mathrm{Hom}(M,N)$ and $\psi, \pi \in \mathrm{Hom}(N,P)$, then

$$(\psi \hat{+} \pi) \circ \phi(x) = \psi(\phi(x)) + \pi(\phi(x)) = ((\psi \circ \phi)\hat{+}(\pi \circ \phi))(x)$$

by definition.[30] □

Remark 9.5. We use the notation $\hat{+}$ for clarity in the proof above, and from now on we

[29] We have used this fact before – and the proof is barely a line – we have written it out again here to show that the binary operation $\hat{+}$ can take arguments like $\psi \circ \phi$.

[30] ⚠For the identity $\psi \circ (\phi \hat{+} \pi) = (\psi \circ \phi)\hat{+}(\psi \circ \pi)$ we used the homomorphism property of ψ, while the identity $(\psi \hat{+} \pi) \circ \phi = (\psi \circ \phi)\hat{+}(\pi \circ \phi)$ followed simply from the definition. It may be instructive to recall the first part of Exercise I.3.

extend the convention of writing + for the group operation on a commutative group to the group $\mathrm{Hom}(M,N)$.

Remark 9.6. The addition of commutative groups extends to iterated sums in the same way as addition in a ring as discussed in §3 and we shall not revisit that here.

The second part of Proposition 9.4 is a pair of identities which look a great deal like the distributivity axiom for a ring, and indeed there is an important ring lurking here.

Theorem 9.7. *Suppose that M is a commutative group. Then* $\mathrm{Hom}(M,M)$ *equipped with pointwise addition as its addition and functional composition as its multiplication is a ring whose multiplicative identity is the map $M \to M; x \mapsto x$ and where $U(\mathrm{Hom}(M,M))$ is the set of bijective homomorphisms $M \to M$, with the multiplicative inverse of $\phi \in U(\mathrm{Hom}(M,M))$ being the inverse function.*

Proof. Most of this follows from Proposition 9.4. In particular, $\mathrm{Hom}(M,M)$ is a commutative group under this addition by the first part of that proposition, and the proposed multiplication distributes by the second part. It remains to recall that composition of functions is associative so the proposed multiplication is associative, and the map $M \to M; x \mapsto x$ is certainly a homomorphism and an identity for composition.

Suppose that $\phi \in U(\mathrm{Hom}(M,M))$. Then there is $\psi \in \mathrm{Hom}(M,M)$ such that $\phi(\psi(x)) = x = \psi(\phi(x))$ for all $x \in M$, and it follows that ϕ is a bijection. Conversely, if $\phi \in \mathrm{Hom}(M,M)$ is a bijection then the map taking each element of M to its unique preimage under ϕ, is a homomorphism since ϕ is a homomorphism. Moreover it is an inverse for ϕ w.r.t. the given multiplication because it is an inverse for ϕ under functional composition, and given multiplication is just functional composition restricted to $\mathrm{Hom}(M,M)$. □

Remark 9.8. Note that if $\phi \in U(\mathrm{Hom}(M,M))$ then the two possible meanings of ϕ^{-1} – one as the functional inverse, and the other as the inverse with respect to the multiplication on the ring $\mathrm{Hom}(M,M)$ – coincide.

Specifying the scalar multiplication of an R-module on a commutative group M turns out to be equivalent to specifying a ring homomorphism $R \to \mathrm{Hom}(M,M)$ by a process called currying:

Proposition 9.9 (Currying)**.** *Suppose that R is a commutative ring and M is a commutative group. Let* $\mathrm{Scalar}_R(M)$ *denote the set of functions $. : R \times M \to M$ satisfying the axioms (M1)–(M4); and* $\mathrm{RingHom}(R,\mathrm{Hom}(M,M))$ *the set of ring homomorphisms $R \to \mathrm{Hom}(M,M)$. Then the* ***currying*** *map*

$$\mathrm{Scalar}_R(M) \to \mathrm{RingHom}(R,\mathrm{Hom}(M,M))$$

$$. : R \times M \to M \mapsto \begin{array}{ccc} R & \to & \mathrm{Hom}(M,M) \\ r & \mapsto & (M \to M; x \mapsto r.x) \end{array}$$

and ***uncurrying*** *map*

$$\mathrm{RingHom}(R,\mathrm{Hom}(M,M)) \to \mathrm{Scalar}_R(M)$$
$$\phi : R \to \mathrm{Hom}(M,M) \mapsto \begin{array}{ccc} R \times M & \to & M \\ (r,x) & \mapsto & \phi(r)(x) \end{array}$$

are well-defined and inverses of each other.

Proof. If $. : R \times M \to M$ is a scalar multiplication then $R \to \mathrm{Hom}(M,M); r \mapsto (M \to M; x \mapsto r.x)$ is well-defined since scalar multiplication distributes over addition (M4), and a ring homomorphism because of (M1)–(M3). Hence the first map in the proposition is well-defined. In the other direction, if $\phi : R \to \mathrm{Hom}(M,M)$ is a ring homomorphism then the map $R \times M \to M; (r,x) \mapsto \phi(r)(x)$ enjoys (M4) since $\phi(r)$ is a homomorphism of M, and (M1)–(M3) since ϕ is a ring homomorphism. It follows that the second map is well-defined.

Finally a short check reveals that these maps are inverses of each other. □

Remark 9.10. In the light of this proposition we shall often specify an R-module structure on a commutative group M simply by defining a ring homomorphism $R \to \mathrm{Hom}(M,M)$.

Example 9.11 (Abelian groups as modules)**.** Suppose that M is a commutative[31] group with identity 0_M. Then by Proposition 3.7 M there is a (unique) homomorphism $\mathbb{Z} \to \mathrm{Hom}(M,M)$ which by uncurrying endows M with the structure of a $\mathbb{Z}$-module.

Example 9.12 (Vector spaces with an endomorphism as modules)**.** Suppose that V is an $\mathbb{F}$-vector space and $T : V \to V$ is $\mathbb{F}$-linear (this is the eponymous endomorphism). Then by currying the vector space structure gives a homomorphism $\phi : \mathbb{F} \to \mathrm{Hom}(V,V)$. Since T is a group homomorphism of the additive group of V we have $T \in \mathrm{Hom}(V,V)$, and since $T(\lambda.v) = \lambda.T(v)$ for all $\lambda \in \mathbb{F}$ and $v \in V$ we have that T commutes with the image of ϕ and so by Proposition 4.21 there is an evaluation homomorphism $\mathbb{F}[X] \to \mathrm{Hom}(V,V)$ taking X to T. By uncurrying this gives V the structure of an $\mathbb{F}[X]$-module. Concretely the scalar multiplication has $p.v = p(T)v$ for $p \in \mathbb{F}[X]$ and $v \in V$.

Remark 9.13. In the above example we write $\mathrm{End}_{\mathbb{F}}(V)$ for the set of $\mathbb{F}$-linear maps $V \to V$. All linear maps are, in particular, homomorphisms of the additive group, so this is a subset of $\mathrm{Hom}(V,V)$ where V is just considered as the additive group of the vector space. The sum and composition of two linear maps is linear; if T is linear then $-T$ is linear; and the identity map is linear. Hence by the subring test $\mathrm{End}_{\mathbb{F}}(V)$ is a ring.

⚠Note that $\mathrm{End}_{\mathbb{C}}(\mathbb{C})$ does not include complex conjugation, but $\mathrm{Hom}(\mathbb{C},\mathbb{C})$ does.

[31] In these notes we use the word commutative in place of Abelian.

10 Linear maps, isomorphisms, and submodules

As with rings we shall be interested in the structure-preserving maps for modules: An R-**linear map** between two R-modules M and N is a group homomorphism $\phi : M \to N$ with

$$\phi(r.x) = r.\phi(x) \text{ for all } x \in M, r \in R.$$

⚠The . on the left is the scalar multiplication on M and the . is the scalar multiplication on N.

Remark 10.1. If $\mathbb{F}$ is a field this has the same meaning as $\mathbb{F}$-linear for vector spaces.

The linear maps between modules are themselves structured.

Proposition 10.2. *Suppose that R is a commutative ring and M and N are R-modules. Then $L(M,N)$, the set of R-linear maps $M \to N$, is a commutative group under pointwise addition, and the map $\hat{.} : R \times L(M,N) \to L(M,N)$ defined by $(r \hat{.} \psi)(x) = r.\psi(x)$ for all $x \in M$ is well-defined and gives the group $L(M,N)$ the structure of an R-module.*

Proof. The zero map, $M \to N; x \mapsto 0_N$ is linear and so $L(M,N)$ is a non-empty subset of $\operatorname{Hom}(M,N)$. If $\phi, \psi \in L(M,N)$ then $\phi - \psi$ is a homomorphism (since $\operatorname{Hom}(M,N)$ is a commutative group under pointwise addition) and

$$(\phi - \psi)(r.x) = \phi(r.x) - \psi(r.x) = r.\phi(x) - r.\psi(x) = r.(\phi(x) - \psi(x)) = r.((\phi - \psi)(x))$$

for all $r \in R$ and $x \in M$. We conclude that $\phi - \psi$ is linear and hence by the subgroup test $L(M,N)$ is a commutative group.

To see that $\hat{.}$ is well-defined, first note that

$$(r \hat{.} \psi)(x + y) = r.(\psi(x + y)) = r.(\psi(x) + \psi(y)) = r.\psi(x) + r.\psi(y) = (r \hat{.} \psi)(x) + (r \hat{.} \psi)(y)$$

for $r \in R$, $\psi \in L(M,N)$ and $x, y \in M$, since ψ is a homomorphism and (M4) holds for ., so that $r \hat{.} \psi$ is a homomorphism. Secondly,

$$(r \hat{.} \psi)(s.x) = r.\psi(s.x) = r.(s.\psi(x)) = (rs).\psi(x) = (sr).\psi(x) = s.(r.\psi(x)) = s.(r \hat{.} \psi)(x)$$

for all $r, s \in R$, $\psi \in L(M,N)$ and $x \in M$, by linearity of ψ, (M2) for ., and commutativity of R. It follows that $r \hat{.} \psi$ is linear. It remains to check (M1)–(M4) for $\hat{.}$. (M1)–(M3) follow from the corresponding axioms for ., and (M4) follows from (M4) for . and the definition of pointwise addition. □

Remark 10.3. Conventionally we drop the circumflex from $\hat{.}$; we used it above to make the argument clearer.

Example 10.4. Suppose that M and N are commutative groups. By Example 9.11 M and N are uniquely equipped with the structure of a $\mathbb{Z}$-module, and an induction shows that *any* $\phi \in \text{Hom}(M,N)$ is $\mathbb{Z}$-linear with respect to this module structure so that $L(M,N) = \text{Hom}(M,N)$ in this case. This is not typical: the example at the end of Remark 9.13 shows that in general $L(M,N)$ may be strictly contained in $\text{Hom}(M,N)$.

Lemma 10.5. *Suppose that M, N, and P are R-modules and $\phi : M \to N$ and $\psi : N \to P$ are R-linear. Then $\psi \circ \phi : M \to P$ is R-linear.*

Proof. The composition of group homomorphisms is a group homomorphism, and $(\psi \circ \phi)(r.x) = \psi(\phi(r.x)) = \psi(r.\phi(x)) = r.\psi(\phi(x)) = r.(\psi \circ \phi)(x)$ for all $r \in R$ and $x \in M$. The result is proved. □

Remark 10.6. Given an R-module M we extend the notation of Remark 9.13 and write $\text{End}_R(M)$ for the set of R-linear maps $M \to M$. (Note that $\text{End}_R(M) = L(M,M)$.) Again by the subring test (the identity map is R-linear, Proposition 10.2 gives sums and additive inverses, and Lemma 10.5 gives products) this is a ring.

Lemma 10.7. *Suppose that $\phi : M \to N$ is an R-linear map between R-modules. Then $\phi(0_M) = 0_N$ and $\phi(-x) = -\phi(x)$ for all $x \in M$.*

Proof. This already follows from the fact that ϕ is a group homomorphism. □

An R-module N is a **submodule** of an R-module M if the map $j : N \to M; x \mapsto x$ is a well-defined R-linear map. We write $N \leqslant M$ when N is a submodule of M.

Lemma 10.8 (Submodule test)**.** *Suppose that M is an R-module and $\varnothing \neq N \subset M$ has $x + y \in N$ for all $x, y \in N$, and $r.x \in N$ whenever $x \in N$ and $r \in R$. Then addition on M and scalar multiplication of R on M restrict to well-defined operations on N giving it the structure of a submodule of M.*

Proof. First, $-1 \in R$ and $(-1).x = x$ for all $x \in M$ so that by the hypotheses, N is non-empty and $x - y \in N$ whenever $x, y \in N$. It follows that N with binary operation the addition on M restricted to N, is a subgroup of M by the subgroup test. The last hypothesis ensures that scalar multiplication of R on M restricts to a well-defined function $R \times N \to N$ which *a fortiori* satisfies (M1)–(M4). Finally, the inclusion map is R-linear and the result is proved. □

Remark 10.9. As with rings (see the comment immediately after Lemma 2.7), given a subset satisfying the hypotheses of the above lemma, we make the common abuse of calling it a submodule on the understanding that we are referring to the induced operations.

Example 10.10. Given an R-module M, the zero R-module $\{0_M\}$ and M itself are submodules of M. In this way modules are more like ideals (Example 5.2) than subrings (see Example 3.10). This foreshadows the fact in Example 10.14 that ideals can be viewed as submodules.

Given an R-linear map $\phi : M \to N$, its **kernel** is its kernel as a homomorphism of groups.

Proposition 10.11. *Suppose that $\phi : M \to N$ is R-linear. Then $\ker \phi$ is a submodule of M and $\operatorname{Im} \phi$ is a submodule of N.*

Proof. Both are subgroups of the relevant groups by the corresponding result for groups, so by the submodule test it is enough to note that if $x \in \ker \phi$ then $0_N = r.0_N = r.\phi(x) = \phi(r.x)$ and so $r.x \in \ker \phi$, and if $x \in \operatorname{Im} \phi$ then there is $y \in M$ such that $x = \phi(y)$ and so $r.x = r.\phi(y) = \phi(r.y) \in \operatorname{Im} \phi$. □

Remark 10.12. ⚠While kernels of ring homomorphisms need not be subrings, kernels of module linear maps *are* submodules.

Proposition 2.11 showed how ring homomorphisms from fields give rise to vector space structure. This is a special case of the fact that ring homomorphisms from commutative rings give rise to module structure.

Proposition 10.13. *Suppose that $\phi : R \to S$ is a ring homomorphism from a commutative ring R. Then the map $R \times S \to S; (r, v) \mapsto r.v := \phi(r)v$ gives the additive group of S the structure of an R-module such that right multiplication on S is R-linear and if ϕ maps into the centre of S then left multiplication is R-linear too.*

Proof. (M1) follows since $\phi(1_R) = 1_S$; (M2), since $\phi(rr') = \phi(r)\phi(r')$ and multiplication in S is associative; (M3) since both ϕ and multiplication on the right in S are additive homomorphisms; and (M4) since multiplication on the left in S is an additive homomorphism. Linearity of right multiplication follows since multiplication on the right in S is an additive homomorphism, and multiplication in S is associative (so $(r.x)y = (\phi(r)x)y = \phi(r)(xy) = r.(xy)$). Finally, left multiplication in S is an additive homomorphism, and if ϕ maps into the centre of S then $x(r.y) = x(\phi(r)y) = (x\phi(r))y = (\phi(r)x)y = \phi(r)(xy) = r.(xy)$ so that left multiplication is R-linear. □

We say that the ring multiplication in S above is **bilinear** if multiplication on the left *and* right is linear.

Example 10.14. Suppose that R is a commutative ring. Then R is an R-module by the above proposition applied to the identity map. Furthermore, every submodule of this R-module is an ideal in R, and conversely by the submodule test every ideal is a submodule.

⚠Many of the familiar rings ($.e.g$ $\mathbb{Z}$, $\mathbb{F}_p$, $\mathbb{Q}$, and $\mathbb{R}$) only have one ring homomorphism from the ring to itself, so that there is a unique way that Proposition 10.13 can be used to give R the structure of an R-module in these cases. However, caution is warranted because more generally there may be many. (See Exercise III.1.)

Example 10.15. Suppose that R is a commutative ring and $\phi : R \to S$ is a ring homomorphism with $\lambda \in S$ commuting with all elements of the image of ϕ. Then (by Proposition 4.21) there is an evaluation homomorphism $R[X] \to R[\lambda]$, and the above proposition gives the ring $R[\lambda]$ the structure of an $R[X]$-module such that multiplication is bilinear. When ϕ is clear we shall speak of the $R[X]$-module $R[\lambda]$.

⚠Given a field $\mathbb{F}$ and a matrix $A \in M_n(\mathbb{F})$ there are two $\mathbb{F}[X]$-modules naturally associated with A: the first is the $\mathbb{F}[X]$-module arising by the construction in Example 9.12 applied to the linear map $\mathbb{F}^n \to \mathbb{F}^n; v \mapsto vA$. The second, recalling the conventional meaning of $\mathbb{F}[A]$ from Example 4.28, is the $\mathbb{F}[X]$-module $\mathbb{F}[A]$ defined in the present example.

That being said, $\mathbb{F}[A]$ is itself an $\mathbb{F}$-vector space and the map $\tilde{A} : \mathbb{F}[A] \to \mathbb{F}[A]; p(A) \mapsto Ap(A)$ is a well-defined $\mathbb{F}$-linear map. The $\mathbb{F}[X]$-module $\mathbb{F}[A]$ defined in this example is the same as the vector-space-with-endomorphism module defined by $\tilde{A}$ on the $\mathbb{F}$-vector space $\mathbb{F}[A]$ by Example 4.28.

Isomorphisms of modules

We say that $\phi : M \to N$ is an **R-linear isomorphism** if it is an R-linear bijection.

Lemma 10.16. *Suppose that $\phi : M \to N$ is an R-linear isomorphism. Then ϕ^{-1} is R-linear, and hence an R-linear isomorphism.*

Proof. ϕ^{-1} is a group homomorphism since ϕ is a bijective group homomorphism. Hence it is enough to show that: $\phi^{-1}(r.x) = \phi^{-1}(r.\phi(\phi^{-1}(x))) = \phi^{-1}(\phi(r.\phi^{-1}(x))) = r.\phi^{-1}(x)$ for all $x \in M$ and $r \in R$ by the R-linearity of ϕ and the fact that ϕ^{-1} is a left and right inverse for ϕ. □

We write $M \cong N$ if there is an R-linear isomorphism $M \to N$.

Proposition 10.17. *$\cong$ is an equivalence relation.*

Proof. The identity map on an R-module is an R-linear isomorphism so $\cong$ is reflexive. $\cong$ is symmetric in view of Lemma 10.16. Finally, $\cong$ is transitive since the composition of bijections is a bijection, and composition of R-linear maps is R-linear – this is Lemma 10.5. □

Remark 10.18. ⚠Note that there are rings which also have a module structure that are isomorphic as modules but not as rings (see Exercises IV.3 and I.8) and vice-versa (see Exercises III.3 and I.6).

Quotient modules and the First Isomorphism Theorem

Proposition 10.19 (Quotient modules)**.** *Suppose that M is an R-module and N is a submodule of M. Then the commutative group M/N may be endowed with the structure of an R-module such that $q : M \rightarrow M/N; x \mapsto x + N$ is an R-linear surjection.*

Proof. Since N is a commutative subgroup of M we have that M/N is a commutative group and the map q is a homomorphism by definition of the quotient group construction. Write . for the scalar multiplication on M and define a map $\hat{.} : R \times M/N \rightarrow M/N$ by $r\hat{.}q(x) := q(r.x)$ for all $r \in R$ and $x \in M$. This is well-defined, first since q is surjective so that for every $W \in M/N$ has $W = q(x)$ for some x; and since if $q(x) = q(y)$ so that $x + N = y + N$, then $x - y \in N$ and hence $r.(x - y) \in N$ and so $r.x + N = r.y + N$ *i.e.* $q(r.x) = q(r.y)$.

(M1) follows since $1\hat{.}q(x) = q(1.x) = q(x)$ for all $x \in M$ by (M1) for .. (M2) follows since $r\hat{.}(s\hat{.}q(x)) = r\hat{.}q(s.x) = q(r.(s.x)) = q((rs).x) = (rs)\hat{.}q(x)$ for all $r, s \in R$ and $x \in M$ by (M2) for .. (M3) follows since q is a homomorphism so $(r + s)\hat{.}q(x) = q((r + s).x) = q(r.x + s.x) = q(r.x) + q(s.x) = r\hat{.}q(x) + s\hat{.}q(x)$ for all $r, s \in R$ and $x \in M$ by (M3). Finally, (M4) follows since q is a homomorphism so $r\hat{.}(q(x) + q(y)) = r\hat{.}q(x + y) = q(r.(x + y)) = q(r.x + r.y) = q(r.x) + q(r.y) = r\hat{.}q(x) + r\hat{.}q(y)$ for all $r \in R$ and $x, y \in M$ by (M4).

Finally, it remains to note that q is R-linear by definition and the result is proved. □

Example 10.20 (Example 10.14, continued)**.** Suppose that R is a commutative ring and I is an ideal in R. Then I is a submodule of R (as noted in Example 10.14) and hence R/I is an R-module. Of course R/I is also a ring and Proposition 10.13 applied to the quotient map endows R/I with the same module structure as the aforementioned one and additionally gives that the multiplication on the ring R/I is bilinear. Put another way, the ring structure and module structure on R/I are 'compatible'.

Theorem 10.21 (First Isomorphism Theorem)**.** *Suppose that $\phi : M \rightarrow N$ is an R-linear map between R-modules M and N. Then $\ker\phi$ is a submodule of M; $\operatorname{Im}\phi$ is a submodule of N; and the map*

$$\widetilde{\phi} : M/\ker\phi \rightarrow N; x + \ker\phi \mapsto \phi(x)$$

is an injective R-linear map with image $\operatorname{Im}\phi$.

Proof. The first two conclusions are in Proposition 10.11. By Proposition 10.19 $M/\ker\phi$ is an R-module. The map is injective and well-defined since $x+\ker\phi = y+\ker\phi$ iff $x-y \in \ker\phi$ iff $\phi(x - y) = 0$ iff $\phi(x) = \phi(y)$. It is a homomorphism by the First Isomorphism Theorem for groups and so it remains to check that

$$\widetilde{\phi}(r.(x + \ker\phi)) = \widetilde{\phi}((r.x) + \ker\phi) = \phi(r.x) = r.\phi(x) = r.\widetilde{\phi}(x + \ker\phi).$$

The result is proved. □

Example 10.22. The First Isomorphism Theorem applied to the R-linear map $M \to M; x \mapsto x$ gives the isomorphism $M/\{0\} \cong M$; *c.f.* Example 5.23.

11 Direct sums of modules

One may of generating new modules from old is through direct sums.

Proposition 11.1. *Suppose that R is a commutative ring and $(M_i)_{i\in I}$ is a family of R modules. Then the **direct sum** is the set $\bigoplus_{i\in I} M_i$, of families $(x_i)_{i\in I}$ with $x_i \in M_i$ for all $i \in I$, and $x_i = 0_{M_i}$ for all but finitely many $i \in I$, endowed with the structure of an R-module with addition and scalar multiplication defined by*

$$x + y := (x_i + y_i)_{i\in I} \text{ and } r.x := (r.x_i)_{i\in I} \text{ for all } x, y \in \bigoplus_{i\in I} M_i \text{ and } r \in R.$$

The zero of this module is $(0_{M_i})_{i\in I}$, the additive inverse of $(x_i)_{i\in I}$ is $(-x_i)_{i\in I}$. The embeddings

$$\iota_j : M_j \to \bigoplus_{i\in I} M_i \text{ where } \iota_j(x)_i = \begin{cases} x_j & \text{if } i = j \\ 0_{M_i} & \text{otherwise} \end{cases}$$

are R-linear.

Proof. The direct sum of commutative groups is a commutative group with the given identity and additive inverse. Moreover, (M1)–(M4) follow for the scalar multiplication on $\bigoplus_{i\in I} M_i$, from the corresponding axioms coordinate-wise on the M_is. The linearity of the embeddings follows since $r.0_{M_i} = 0_{M_i}$ for all $r \in R$ and $0_{M_i} + 0_{M_i} = 0_{M_i}$ for all $i \in I$. □

Remark 11.2. If $I = \emptyset$ then $\bigoplus_{i\in I} M_i$ is the zero module. If $M_1, \dots, M_n$ are modules then we write $M_1 \oplus \cdots \oplus M_n$ for $\bigoplus_{i\in\{1,\dots,n\}} M_i$, and finally M^n for the direct sum of M with itself n-times.

⚠ Although the direct sum $M_1 \oplus \cdots \oplus M_n$ appears to have an order, its definition only depends on the set $\{1, \dots, n\}$, not on an order of the elements of that set.

Remark 11.3. Given a commutative ring R we write $\bigoplus_{i\in I} R$ for the R-module that is the direct sum of I copies of the R-module R (an R-module as in Example 10.14).

Example 11.4 (Baer-Specker group)**.** The set $\mathbb{Z}^{\mathbb{N}_0}$ – that is the set of functions $f : \mathbb{N}_0 \to \mathbb{Z}$ – has the structure of a ring (it is the direct product of $\mathbb{N}_0$ copies of the integers $\mathbb{Z}$ as described in Proposition 4.13). By Proposition 3.7 there is a unique homomorphism $\mathbb{Z} \to \mathbb{Z}^{\mathbb{N}_0}$, and this gives $\mathbb{Z}^{\mathbb{N}_0}$ the structure of a $\mathbb{Z}$-module by Proposition 10.13. Concretely we have $(f + g)(x) = f(x) + g(x)$ and $(\lambda.f)(x) = \lambda f(x)$ for all $x \in \mathbb{N}_0$.

⚠ The module $\mathbb{Z}^{\mathbb{N}_0}$ is uncountable, but $\bigoplus_{i\in\mathbb{N}_0} \mathbb{Z}$ – the direct sum of $\mathbb{N}_0$ copies of $\mathbb{Z}$ – is countable so these modules are *not* isomorphic.

Remark 11.5. There are a few expected relationships between direct sums: If $(M_j)_{j\in J_i}$ is a family of R-modules for each $i \in I$ then the map

$$\bigoplus_{j\in\bigsqcup_{i\in I} J_i} M_j \to \bigoplus_{i\in I}\left(\bigoplus_{j\in J_i} M_j\right); x \mapsto ((x_j)_{j\in J_i})_{i\in I}$$

is a well-defined R-linear isomorphism. $\sqcup$ here denotes disjoint union; we assume the J_is are disjoint for distinct is.

Furthermore, if $(M_i)_{i\in I}$ and $(N_i)_{i\in I}$ are families of R-modules with R-linear maps $\phi_i : M_i \to N_i$. Then the map

$$\phi : \bigoplus_{i\in I} M_i \to \bigoplus_{i\in I} N_i; x \mapsto (\phi_i(x_i))_{i\in I} \tag{11.1}$$

is R-linear; if ϕ_i is an injection for all $i \in I$ then ϕ is an injection; if ϕ_i is a surjection for all $i \in I$ then ϕ is a surjection.

Given elements $(x_i)_{i\in I}$ of an R-module M we write

$$\langle x_i : i \in I\rangle := \bigcap \{N : N \leqslant M \text{ and } x_i \in N \text{ for all } i \in I\}$$

which is a submodule of M by the submodule test *c.f.* (5.1). We call $\langle x_i : i \in I\rangle$ the module **generated** by $(x_i)_{i\in I}$.

Remark 11.6. ⚠Note that if M is also a ring then $\langle x_i : i \in I\rangle$ is ambiguous: it could mean the ideal or module generated by these elements. Although in some important cases these are the same (for example when M is the R-module R of Example 10.14), on other occasions the meaning has to be determined from context.

Example 11.7. Suppose that V is an $\mathbb{F}$-vector space and $(x_i)_{i\in I}$ are elements in V. Then $\langle x_i : i \in I\rangle$ is the span of the vectors in the family $(x_i)_{i\in I}$.

Remark 11.8. Suppose M is an R-module and $x_1, \dots, x_n \in M$. Then

$$\langle x_1, \dots, x_n\rangle := \langle x_i : i \in \{1, \dots, n\}\rangle = \left\{\sum_{i=1}^{n} r_i.x_i : r_1, \dots, r_n \in R^n\right\}$$

since by the submodule test the right hand side is a module and so the middle is contained in it. On the other hand since $\langle x_i : i \in \{1, \dots, n\}\rangle$ contains $x_1, \dots, x_n$ it contains all sums in the set on the right. It may be helpful to compare with Remark 5.8 in view of Example 10.14.

An R-module M is said to be **finitely generated** if it is generated by $(x_i)_{i\in I}$ for some finite set I.

There are many examples of finitely generated modules.

Example 11.9. A finite dimensional vector space has a finite spanning set and so is finitely generated as a module over its field.

Example 11.10. Suppose that V is a vector space over $\mathbb{F}$ and $T : V \to V$ is $\mathbb{F}$-linear with V having the endomorphism module structure as Example 9.12. If V is finite dimensional as a vector space over $\mathbb{F}$, then V is finitely generated as an $\mathbb{F}[X]$-module: $\mathbb{F}$ is a subring of $\mathbb{F}[X]$ in such a way that if $\lambda \in \mathbb{F}$ and $v \in V$ then $\lambda.v$ in the scalar multiplication of the vector space V is the same as $\lambda.v$ in the scalar multiplication in the $\mathbb{F}[X]$-module. It follows that any generating set for V as an $\mathbb{F}$-space is also a generating set for V as an $\mathbb{F}[X]$-module.

⚠The converse does *not* hold: if $V = \mathbb{F}[X]$ as a vector space and $T : V \to V; f(X) \mapsto Xf(X)$ then T is $\mathbb{F}$-linear and V is generated by 1 as an $\mathbb{F}[X]$-module but it is infinite dimensional.

An R-module M is said to be **cyclic** if M is generated by one element.

Example 11.11 (Examples 10.14 & 10.20, continued)**.** Suppose that R is a commutative ring and I is an ideal in R, and R (resp. I) as an R-module (resp. submodule) in the same way as in Examples 10.14 & 10.20. Then R/I is cyclic, generated by $1 + I$.

Suppose that M is an R-module and $x \in M$ and $r \in R$. Then we put

$$\operatorname{Ann}_R(x) := \{r \in R : r.x = 0_M\} \text{ and } r.M := \{r.z : z \in M\}.$$

We call $\operatorname{Ann}_R(x)$ the **annihilator** of x.

Proposition 11.12. *Suppose that R is a commutative ring and M is an R-module. Then*

(i) for $x \in M$, $\operatorname{Ann}_R(x)$ is an ideal in R, and if M is generated by x then there is an R-linear isomorphism $R/\operatorname{Ann}_R(x) \to M$ taking $1 + \operatorname{Ann}_R(x)$ to x;

(ii) for $r \in R$, $r.M$ is a submodule of M, and if $\phi : M \to N$ is an isomorphism then $\tilde{\phi} : r.M \to r.N; r.x \mapsto r.\phi(x)$ is a well-defined isomorphism.

Proof. That $\operatorname{Ann}_R(x)$ is an ideal of R is a short check from the axioms (using that R is commutative). Moreover, by the First Isomorphism Theorem for modules (Theorem 10.21) applied to the R-linear map $R \to M; r \mapsto r.x$ (where we treat R as an R-module as in Example 10.14), the map

$$R/\operatorname{Ann}_R(x) \to \langle x \rangle; r + \operatorname{Ann}_R(x) \mapsto r.x$$

is a well-defined R-linear isomorphism.

The map $M \to M; z \mapsto r.z$ is R-linear (using that R is commutative), and so by Proposition 10.11 its image, $r.M$, is an R-module. Moreover, the given map $\tilde{\phi}$ is well-defined since if $r.z = r.z'$ then $r.\phi(z) = \phi(r.z) = \phi(r.z') = r.\phi(z')$; conversely if $r.\phi(z) = r.\phi(z')$ then $r.z = r.z'$ since ϕ is an injection, so $\tilde{\phi}$ is an injection. The map is a surjection since ϕ is a surjection, and it is linear since

$$\tilde{\phi}(r.z + r.z') = \tilde{\phi}(r.(z + z')) = r.\phi(z + z') = r.\phi(z) + r.\phi(z') = \tilde{\phi}(r.z) + \tilde{\phi}(r.z')$$

and (again using commutativity of R)

$$\begin{aligned}\tilde{\phi}(y.(r.z)) &= \tilde{\phi}((yx).z) = \tilde{\phi}((xy).z) = \tilde{\phi}(r.(y.z))\\ &= r.\phi(y.z) = r.(y.\phi(z)) = (xy).\phi(z) = (yx).\phi(z) = y.(r.\phi(z)) = y.\tilde{\phi}(r.z).\end{aligned}$$

The result is proved. □

Remark 11.13. The above result is one of the places where we use commutativity of our rings of interest. If they are not then the map $M \to M; x \mapsto r.x$ need not be R-linear, and $\mathrm{Ann}_R(x)$ will only be a 'one-sided' ideal, meaning that it will only be closed under multiplication by elements of the ring on one side *c.f.* footnote 24. Many results still follow, but more care is needed, and for some questions attention is restricted to **duo rings**, that is not-necessarily-commutative rings in which every one-sided ideal is two-sided.

In view of Proposition 11.12 and the fact that fields only have two ideals (as shown in Example 5.21), we see that the only cyclic modules over a field are the zero module and the field itself. It follows that for vector spaces the only direct sums of cyclic modules are direct sums of copies of of the field. Moreover, two direct sums of copies of the field are isomorphic if and only if their indexing sets are the same cardinality.[32] In modules there are some less obvious isomorphisms between direct sums of cyclic modules which are captured by the following result.

Theorem 11.14 (Chinese Remainder Theorem). *Suppose that R is a commutative ring and $I_1, \dots, I_k$ are ideals in R such that $I_j + I_i = R$ for all $i \neq j$. Then the map*

$$\phi : R \to (R/I_1) \oplus \cdots \oplus (R/I_k); r \mapsto (r + I_1, \dots, r + I_k)$$

is a surjective R-linear map with kernel $I_1 \cap \cdots \cap I_k$.

Proof. The quotient maps $q_i : R \to R/I_i$ are R-linear and so are the embeddings $\iota_i : R/I_i \to (R/I_1)\oplus\cdots\oplus(R/I_k)$ and hence $\phi = \iota_1 \circ q_1 + \cdots + \iota_k \circ q_k$ is R-linear. The kernel is $I_1 \cap \cdots \cap I_k$; proving the map is surjective is the rub.

[32] This is the so called **Dimension Theorem** for vector spaces which has been seen for finite direct sums in Prelims. The finite case is sufficient for our understanding but a general proof may be found, for example, in [Lan02, Theorem 5.2,Chapter III].

Fix j and note that since $I_j + I_i = R$ for all $i \neq j$ there are elements $z_i \in I_j$ and $w_i \in I_i$ with $z_i + w_i = 1$. It follows that

$$1 = \left(1 - \prod_{i:i\neq j}(1 - z_i)\right) + \left(\prod_{i:i\neq j} w_i\right) = \sum_{\emptyset\neq S\subset\{i:i\neq j\}} (-1)^{|S|+1}\prod_{s\in S} z_s + \left(\prod_{i:i\neq j} w_i\right),$$

and so if we set $y_j := \prod_{i:i\neq j} w_i$ then $y_j \in I_i$ for all $i \neq j$ and $1 - y_j \in I_j$. Thus for $u \in (R/I_1) \oplus \cdots \oplus (R/I_k)$ we have

$$\phi\left(u_1y_1 + \cdots + u_ky_k\right) = (u_1y_1 + I_1, \ldots, u_ky_k + I_k) = (u_1 + I_1, \ldots, u_k + I_k)$$

and the map is surjective as required. □

Remark 11.15. The history of this theorem is involved [She88], but the starting point is work of Sun Zi (孫子) from around 400AD who gave an application of a method for solving given simultaneous congruences.

This has the following immediate and more familiar corollary.

Corollary 11.16. *Suppose that $m_1, \ldots, m_k$ are pairwise coprime natural numbers and $a_1, \ldots, a_k$ are integers. Then there is $x \in \mathbb{Z}$ such that $x \equiv a_i \pmod{m_i}$ for all $1 \leqslant i \leqslant k$.*

Proof. Take $R = \mathbb{Z}$ and $I_i := \langle m_i\rangle$. By Bezout's Lemma (Theorem 4.2) $\langle m_i\rangle + \langle m_j\rangle = \mathbb{Z}$ for $i \neq j$ since m_i and m_j are coprime, and so by Theorem 11.14 there is some $z \in \mathbb{Z}$ such that $z \equiv a_i \pmod{m_i}$ for all $1 \leqslant i \leqslant k$. □

Example 11.17. Since 2 and 3 are coprime in $\mathbb{Z}$ and $\langle 2\rangle \cap \langle 3\rangle = \langle 6\rangle$, the First Isomorphism Theorem for modules (Theorem 10.21) applied to the homomorphism from the Chinese Remainder Theorem (Theorem 11.14) gives a $\mathbb{Z}$-module isomorphism $\mathbb{Z}_6 \cong \mathbb{Z}_2 \oplus \mathbb{Z}_3$.

The above example seems to call into question the possibility of an analogue of the Dimension Theorem for direct sums of cyclic modules, as it produces an isomorphism between different numbers of non-zero cyclic modules. Despite this there is a way to recover a result as follows.

Theorem 11.18 (Uniqueness Theorem)**.** *Suppose that R is a commutative ring, M is an R-module, and $I_1 \subset \cdots \subset I_n$ and $J_1 \subset \cdots \subset J_m$ are proper ideals such that $M \cong (R/I_1) \oplus \cdots \oplus (R/I_n)$ and $M \cong (R/J_1) \oplus \cdots \oplus (R/J_m)$. Then $n = m$ and $J_k = I_k$ for all $1 \leqslant k \leqslant n$.*

Remark 11.19. ⚠Note that we need the ideals to be proper: if $I = R$ then R/I is the zero module as an R-module, and if Z is a zero module then $Z^n \cong Z^m$ for all $n, m \in \mathbb{N}_0$.

We begin with a result which essentially bootstraps the Dimension Theorem for finite dimensional vector spaces.

Lemma 11.20. *Suppose that R is a commutative ring, and $I_1 \subset \cdots \subset I_n$ are proper ideals. Then $(R/I_1) \oplus \cdots \oplus (R/I_n)$ is generated by a set of size n and by no smaller set.*

Proof. Surjective R-linear maps take generating sets to generating sets. The R-module R^n has a generating set of size n and so the R-linear surjection

$$R^n \to (R/I_1) \oplus \cdots \oplus (R/I_n); r \mapsto (r_1 + I_1, \dots, r_n + I_n)$$

ensures the first part of the lemma. For the second, by Theorem 6.9 there is a maximal ideal $J \supset I_n$ and hence $J \supset I_k$ for all $1 \leqslant k \leqslant n$. The R-linear surjection

$$(R/I_1) \oplus \cdots \oplus (R/I_n) \to (R/J)^n; (x_1 + I_1, \dots, x_n + I_n) \mapsto (x_1 + J, \dots, x_n + J)$$

is therefore well-defined, and ensures that if $(R/I_1) \oplus \cdots \oplus (R/I_n)$ has a generating set of size t then so does $(R/J)^n$ as an R-module. Let $x^{(1)}, \dots, x^{(t)}$ be a generating set for $(R/J)^n$ as an R-module, and note that for every $x \in (R/J)^n$ there are elements $r_1, \dots, r_t \in R$ such that

$$\begin{aligned} x &= r_1.x^{(1)} + \cdots + r_t.x^{(t)} \\ &= (r_1.x_1^{(1)} + \cdots + r_t.x_1^{(t)}, \dots, r_1.x_n^{(1)} + \cdots + r_t.x_n^{(t)}) \\ &= ((r_1 + J)x_1^{(1)} + \cdots + (r_t + J)x_1^{(t)}, \dots, (r_1 + J)x_n^{(1)} + \cdots + (r_t + J)x_n^{(t)}) \\ &= (r_1 + J).x^{(1)} + \cdots + (r_t + J).x^{(t)}, \end{aligned}$$

where the scalar multiplication in the last line is that arising from uncurrying the natural map $R/J \to \operatorname{Hom}((R/J)^n, (R/J)^n)$, which is a homomorphism by the First Isomorphism Theorem for rings (Theorem 5.22) applied to the curried scalar multiplication of R on $(R/J)^n$ – the latter is a ring homomorphism $R \to \operatorname{Hom}((R/J)^n, (R/J)^n)$ with kernel J.

Proposition 6.6 ensures that R/J is a field and so $(R/J)^n$ is a vector space over R/J and the above calculation shows that $x^{(1)}, \dots, x^{(t)}$ is a spanning set for $(R/J)^n$ as an (R/J)-module *i.e.* as a vector space over R/J. Since $(R/J)^n$ is an n-dimensional vector space over R/J any spanning set has size at least n *i.e.* $t \geqslant n$. □

Proof of Theorem 11.18. By Lemma 11.20 we have $n = m$. For $x \in R$ (using Proposition 11.12 (ii) so that $x.M$ is a module) we shall show that for $1 \leqslant k \leqslant n$

$$I_k = \{x \in R : x.M \text{ has a generating set with strictly fewer than } k \text{ elements}\},$$

from which the result follows without loss of generality. Write K_k for the set on the right.

Suppose that $x \in R$. R/I_k is an R-module and $x.(R/I_k) = \langle x + I_k \rangle$, and so by Proposition 11.12 (i)

$$x.(R/I_k) \cong R/\operatorname{Ann}_R(x + I_k). \tag{11.2}$$

Now $\operatorname{Ann}_R(x+I_k) = \{r \in R : r(x+I_k) = I_k\} = \{r : rx \in I_k\}$, so $x \notin I_k$ if and only if $\operatorname{Ann}_R(x+I_k)$ is proper[33]; and $\operatorname{Ann}_R(x+I_1) \subset \cdots \subset \operatorname{Ann}_R(x+I_n)$ since the $I_1 \subset \cdots \subset I_k$. Let $0 \leqslant j(x) \leqslant n$ be maximal such that $x \notin I_{j(x)}$ (with $j(x) = 0$ if $x \in I_1$) then by Proposition 11.12 (ii)

$$\begin{aligned} x.M &\cong x.((R/I_1)\oplus\cdots\oplus(R/I_n)) && \textit{by definition}\\ &\cong x.(R/I_1)\oplus\cdots\oplus x.(R/I_n) && \textit{by (11.2) and (11.1)}\\ &\cong (R/\operatorname{Ann}_R(x+I_1))\oplus\cdots\oplus(R/\operatorname{Ann}_R(x+I_n)) && \operatorname{Ann}_R(x+I_k)\ \textit{not proper}\\ &\cong (R/\operatorname{Ann}_R(x+I_1))\oplus\cdots\oplus(R/\operatorname{Ann}_R(x+I_{j(x)})) && \Rightarrow R/\operatorname{Ann}_R(x+I_k)\cong\{0\} \end{aligned}$$

with the convention that this is the zero module if $j(x) = 0$ since then the sum is empty.

By Lemma 11.20 we conclude that if $x \notin I_k$ then $j(x) \geqslant k$ and so $x.M$ is not generated by strictly fewer than $j(x)$ (and hence k) elements and so $x \notin K_k$. On the other hand if $x \in I_k$ then $j(x) < k$ and so $x.M$ *is* generated by at most $j(x)$ (*i.e.* strictly fewer than k) elements and so $x \in K_k$. The result is proved. □

Remark 11.21. There remains the question of whether or not a module has a decomposition of the type described in Theorem 11.18. We shall show that if R is a PID then every finitely generated R-module can be decomposed into cyclic modules in this way. Commutative rings with this property are called **FGCF**-rings and are characterised in [SW74]. However, there remain open questions in this area, if we do not require the nesting of the ideals or if we allow non-commutative rings like duo rings from Remark 11.13 *e.g.* [::006, Problem 2.45].

Theorem 11.22. *Suppose that R is a PID and M is a finitely generated R-module. Then there is $n \in \mathbb{N}_0$ and proper ideals $I_1 \subset \cdots \subset I_n$ such that*

$$M \cong (R/I_1)\oplus\cdots\oplus(R/I_n)$$

with the convention that this is the zero module if the sum is empty i.e. *if $n = 0$.*

To prove this we need the following lemma to let us change variables.

Lemma 11.23. *Suppose that R is a PID with elements $a_1,\dots,a_n,h \in R$, and $\langle a_1,\dots,a_n\rangle = \langle h\rangle$, and M is an R-module with elements $x_1,\dots,x_n \in M$. Then there are elements $y_1,\dots,y_n \in M$ with $\langle y_1,\dots,y_n\rangle = \langle x_1,\dots,x_n\rangle$ such that $h.y_n = a_1.x_1 + \cdots + a_n.x_n$.*

Proof. If $h = 0$ then $a_1,\dots,a_n = 0$ and the result is trivial with $y_i = x_i$ for $1 \leqslant i \leqslant n$, so we may assume $h \in R^*$.

We proceed by induction on n; $n = 1$ is immediate since $a_1 \sim h$ in that case, so there is a $u \in U(R)$ such that $a_1 = hu$ and we can take $y_1 := u.x_1$. For $n > 1$ let h' be a

[33] If $x \in I_k$ then $rx \in I_k$ for all $r \in R$ since I_k is an ideal, and hence $\operatorname{Ann}_R(x+I_k) = R$. Conversely, if $\operatorname{Ann}_R(x+I_k) = R$ then $1(x+I_k) = I_k$ and so $x \in I_k$.

generator of $\langle a_1, \dots, a_{n-1}\rangle$. By the inductive hypothesis we may take $y_1, \dots, y_{n-2}, y^*_{n-1}$ such that $\langle y_1, \dots, y_{n-2}, y^*_{n-1}\rangle = \langle x_1, \dots, x_{n-1}\rangle$ and $h'.y^*_{n-1} = a_1.x_1 + \cdots + a_{n-1}.x_{n-1}$.

Let $\alpha, \beta \in R$ be such that $h' = \alpha h$ and $a_n = \beta h$. Since $\langle h\rangle = \langle h', a_n\rangle$ there are elements $\gamma, \delta \in R$ such that $h = \delta h' + \gamma a_n$ and so $\alpha\delta + \beta\gamma = 1$ by cancellation (since $h \in R^*$). Now put $y_{n-1} := \gamma.y^*_{n-1} - \delta.x_n$ and $y_n := \alpha.y^*_{n-1} + \beta.x_n$. Then $x_n = -\alpha.y_{n-1} + \gamma.y_n$ and $y^*_{n-1} = \beta.y_{n-1} + \delta.y_n$, and so

$$\langle y_1, \dots, y_n\rangle = \langle y_1, \dots, y_{n-2}, y^*_{n-1}, x_n\rangle = \langle x_1, \dots, x_n\rangle.$$

Finally, $h.y_n = h'.y^*_{n-1} + a_n.x_n = a_1.x_1 + \cdots + a_n.x_n$ and the result is proved. □

Proof of Theorem 11.22. We proceed inductively to show that there are elements $z_1, \dots, z_n$ generating M such that

$$M \cong (R/\operatorname{Ann}_R(z_1)) \oplus \cdots \oplus (R/\operatorname{Ann}_R(z_n)) \text{ and } \operatorname{Ann}_R(z_1) \subset \cdots \subset \operatorname{Ann}_R(z_n).$$

Since R is a PID, it is a UFD by Theorem 7.21 and in particular this means for every $x \in R^*$ there is[34] a unique $N(x) \in \mathbb{N}_0$ such that $x \sim x_1 \cdots x_{N(x)}$ for irreducible elements $x_1, \dots, x_{N(x)}$. We declare $N(0_R) = \infty$ and note if $x \mid y$ then $N(x) \leqslant N(y)$ with equality if and only if $x \sim y$.

Since M is finitely generated there is a minimal $n \in \mathbb{N}_0$ such that M is generated by a set of size n. Let $x_1, \dots, x_n$ be a set of generators in which $\operatorname{Ann}_R(x_n)$ is generated by an element r_n (possibly 0_R) with $N(r_n)$ minimal out of all possible sets of generators of size n. Note that $\operatorname{Ann}_R(x_n)$ is proper since otherwise $x_1, \dots, x_{n-1}$ would generate M contradicting the minimality of n.

Let $M' := \langle x_1, \dots, x_{n-1}\rangle$ and consider the map

$$\Psi : M' \oplus \langle x_n\rangle \to M; (x, y) \mapsto x + y.$$

This is an R-linear surjection; the key fact, however, is the following.

Claim. Ψ *is an injection* i.e. $\ker \Psi = \{0\}$.

Proof. Suppose that $x+y = 0$ for some $x \in M'$ and $y \in \langle x_n\rangle$ so that $x = a_1.x_1+\cdots+a_{n-1}.x_{n-1}$ and $y = a_n.x_n$ for some $a_1, \dots, a_n \in R$. Let a^*_n be such that $\langle a^*_n\rangle = \langle a_n, r_n\rangle$; $\alpha, \beta \in R$ be such that $a^*_n = \alpha a_n + \beta r_n$; and h be such that $\langle \alpha a_1, \dots, \alpha a_{n-1}, a^*_n\rangle = \langle h\rangle$. Apply Lemma 11.23 to get $y_1, \dots, y_n \in M$ with $\langle y_1, \dots, y_n\rangle = \langle x_1, \dots, x_n\rangle = M$ and

$$\begin{aligned} h.y_n &= (\alpha a_1).x_1 + \cdots + (\alpha a_{n-1}).x_{n-1} + a^*_n.x_n \\ &= \alpha.(a_1.x_1 + \cdots + a_n.x_n) + (\beta r_n).x_n = \alpha.(x + y) + \beta.(r_n.x_n) = \alpha.0 + \beta.0 = 0. \end{aligned}$$

Now $h \mid a^*_n \mid r_n$ and so by minimality of r_n we have $h \sim r_n$, and hence $a^*_n \sim r_n$. But then $r_n \mid a_n$ and $a_n.x_n = 0$ as required. □

[34] It may be of interest to (re-)visit Remark 7.32.

Finally, by the inductive hypothesis there are elements $z_1, \dots, z_{n-1}$ generating M' such that $M' \cong (R/\operatorname{Ann}_R(z_1)) \oplus \cdots \oplus (R/\operatorname{Ann}_R(z_{n-1}))$ with $\operatorname{Ann}_R(z_1) \subset \cdots \subset \operatorname{Ann}_R(z_{n-1})$. Set $z_n := x_n$ and since $\langle x_n \rangle \cong R/\operatorname{Ann}_R(z_n)$ the result is proved if we can show that $\operatorname{Ann}_R(z_{n-1}) \subset \operatorname{Ann}_R(z_n)$.

To see this last claim, suppose that $r \in \operatorname{Ann}_R(z_{n-1})$ and let h be such that $\langle h \rangle = \langle r, r_n \rangle$. Apply Lemma 11.23 to get $y_1, \dots, y_n$ with $\langle y_1, \dots, y_n \rangle = \langle z_1, \dots, z_n \rangle = M$ and $h.y_n = r.z_{n-1} + r_n.z_n = 0$. But $h \mid r_n$ and so by minimality of the number of irreducible factors of r_n we have $h \sim r_n$ and hence $r_n \mid r$ *i.e.* $r \in \langle r_n \rangle = \operatorname{Ann}_R(z_n)$. □

12 The structure theorem for finitely generated modules over PIDs and applications

With the work of the last section we can now formulate the structure theorem.

Theorem 12.1 (Structure Theorem, Invariant Factor Form). *Suppose that R is a PID and M is a finitely generated R-module. Then there is a (possibly empty) sequence $a_r \mid \cdots \mid a_1$ of elements[35] of R with $a_r \not\sim 1$ such that*

$$M \cong (R/\langle a_1 \rangle) \oplus \cdots \oplus (R/\langle a_r \rangle)$$

and the sequence $(a_i)_{i=1}^r$ is unique up to associates.

Proof. The existence of this isomorphism follows from Theorem 11.22 and the fact that ideals in a PID are generated by a single element. The divisibility relation between the elements is exactly the nesting of the ideals; the fact that $a_r \not\sim 1$ is the fact that all the ideals are proper.

The uniqueness now follows from Theorem 11.18 and the definition of association. □

Theorem 12.2 (Structure Theorem, Primary Form). *Suppose that R is a PID and M is a finitely generated R-module. Then there are some $s, t \in \mathbb{N}_0$, irreducible elements $p_1, \dots, p_t \in R$, and $e_1, \dots, e_t \in \mathbb{N}^*$, such that*

$$M \cong R^s \oplus (R/\langle p_1^{e_1} \rangle) \oplus \cdots \oplus (R/\langle p_t^{e_t} \rangle).$$

Proof. We being with some preliminaries about gathering associates in UFDs.[36] Since R is a PID, it is a UFD (by Theorem 7.21) and if $a \in R^*$ there is some $r \in \mathbb{N}_0$ and irreducible elements $x_1, \dots, x_r$ such that $a \sim x_1 \cdots x_r$. By Proposition 7.2 $\sim$ is an equivalence relation;

[35] As usual $0 \mid 0$ and so this sequence may end in a series of 0s.

[36] The idea here is just to group prime factors together, for example in $\mathbb{Z}$ instead of writing things like $540 = 2 \times 3 \times 5 \times 2 \times 3 \times 3$ we will write $540 = 2^2 \times 3^3 \times 5$.

let $\mathcal{P}$ be the partition of the (multi-)set $\{x_1,\dots,x_r\}$ induced by $\sim$. Let $q_1,\dots,q_l$ be one element from each equivalence class and let $c_1,\dots,c_l \in \mathbb{N}^*$ be the size of the corresponding class. Then for every $1 \leqslant i \leqslant r$ there is a unique j such that $x_i \sim q_j$, and $\prod_{i:x_i\sim q_j} x_i \sim q_j^{c_j}$ since (again part of Proposition 7.2) $\sim$ respects multiplication (meaning $xy \sim x'y'$ if $x \sim x'$ and $y \sim y'$). It follows that $a \sim q_1^{c_1}\cdots q_l^{c_l}$, and $q_i \not\sim q_j$ for $i \neq j$.

The ideal $\langle q_i^{c_i}\rangle + \langle q_j^{c_j}\rangle$ is generated by some $h \in R$ (since R is a PID) and $h \mid q_i^{c_i}$ and $h \mid q_j^{c_j}$. If p is a prime factor of h then $p \sim q_i$ and $p \sim q_j$ since R is a UFD, and by transitivity of $\sim$ we have $q_i \sim q_j$ meaning $i = j$. Thus, if $i \neq j$ then h has no prime factors and hence $h \sim 1$ *i.e.* $\langle q_i^{c_i}\rangle + \langle q_j^{c_j}\rangle = \langle 1\rangle$. It follows from the Chinese Remainder Theorem (Theorem 11.14) that

$$R/\langle a\rangle \cong (R/\langle q_1^{c_1}\rangle) \oplus \cdots \oplus (R/\langle q_l^{c_l}\rangle) \tag{12.1}$$

as R-modules for irreducibles $q_1,\dots,q_l$ and naturals $c_1,\dots,c_l \in \mathbb{N}^*$.

Finally, apply Theorem 12.1 to M to get $a_1,\dots,a_r \in R$ such that $M \cong (R/\langle a_1\rangle) \oplus \cdots \oplus (R/\langle a_r\rangle)$. If $a_i = 0_R$ then $R/\langle a_i\rangle \cong R$ as an R-module (Example 10.22); if $a_i \neq 0_R$ then we have an isomorphism of the form (12.1). Combining these isomorphisms using commutativity of the direct sum (see Remark 11.2) and (11.1) we have the result. □

Remark 12.3. There is a uniqueness statement for the primary form of the structure theorem but we do not pursue that here. What is important about the Primary Form as compare with the Invariant Factor Form is that the building blocks in the former cannot be further decomposed.

We have a couple of important corollaries.

Theorem 12.4 (Structure Theorem for finitely generated commutative groups)**.** *Suppose that G is a finitely generated commutative group. Then there are unique (non-zero) natural numbers $1 \neq d_r \mid d_{r-1} \mid \cdots \mid d_1$ and $s \in \mathbb{N}_0$ such that*

$$G \cong \mathbb{Z}^s \oplus \mathbb{Z}_{d_1} \oplus \cdots \oplus \mathbb{Z}_{d_r}.$$

Proof. G is a $\mathbb{Z}$-module, so we may apply the Invariant Factor Form of Theorem 12.1 to get the desired structure, writing $\mathbb{Z}^s$ for the s copies of $\mathbb{Z}/\langle 0\rangle$ in the given decomposition. Then uniqueness follows from the fact that $U(\mathbb{Z}) = \{-1, 1\}$ since we have restricted the d_is to be naturals. □

Theorem 12.5 (Jordan Normal Form)**.** *Suppose that V is a finite-dimensional vector space over $\mathbb{C}$ and $T : V \to V$ is linear. Then there is a basis for V such that the matrix for T in this basis is*

$$\begin{pmatrix} J(\lambda_1,n_1) & 0_{n_1\times n_2} & \cdots & 0_{n_1\times n_t} \\ 0_{n_2\times n_1} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0_{n_{t-1}\times n_t} \\ 0_{n_t\times n_1} & \cdots & 0_{n_t\times n_{t-1}} & J(\lambda_t,n_t) \end{pmatrix}$$

where $0_{n\times m}$ is the all zeros matrix in $M_{n,m}(\mathbb{C})$, and $J(\lambda,n)$ is the $n\times n$ matrix, called a ***Jordan block,***

$$\begin{pmatrix} \lambda & 0 & \cdots & \cdots & 0 \\ 1 & \lambda & \ddots & & \vdots \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \lambda & 0 \\ 0 & \cdots & 0 & 1 & \lambda \end{pmatrix}.$$

The scalars $\lambda_1,\dots,\lambda_t$ are all the eigenvalues of T.

Proof. We regard V as a $\mathbb{C}[X]$-module in the way described in Example 9.12. Since $\mathbb{C}$ is a subring of $\mathbb{C}[X]$ and V is finite dimensional as a $\mathbb{C}$-vector space, the module V is finitely generated by Example 11.10.

Since $\mathbb{C}[X]$ is a PID we may apply the Structure Theorem (Primary Form, Theorem 12.2) to V. We get $s,t\in\mathbb{N}_0$, irreducible polynomials $p_1,\dots,p_t\in\mathbb{C}[X]$, and natural numbers $n_1,\dots,n_t\in\mathbb{N}^*$ such that

$$\phi : V \to (\mathbb{C}[X])^s \oplus (\mathbb{C}[X]/\langle p_1^{n_1}\rangle) \oplus \cdots \oplus (\mathbb{C}[X]/\langle p_t^{n_t}\rangle)$$

is a $\mathbb{C}[X]$-linear bijection. In particular, ϕ is a $\mathbb{C}$-linear bijection but V is finite-dimensional and $\mathbb{C}[X]$ is infinite dimensional so $s=0$. The irreducible polynomials in $\mathbb{C}[X]$ are all degree 1 (see Remark 8.2) thus there are $\lambda_1,\dots,\lambda_t\in\mathbb{C}$ such that $\langle p_i^{n_i}\rangle = \langle (X-\lambda_i)^{n_i}\rangle$; write $M_i := \mathbb{C}[X]/\langle (X-\lambda_i)^{n_i}\rangle$. For each $1\leqslant i\leqslant t$ let $(e_{i,j})_{j=1}^{n_i}$ be such that

$$\phi(e_{i,j}) = (0_{M_1},\dots,0_{M_{i-1}},(X-\lambda_i)^{j-1}+\langle (X-\lambda_i)^{n_i}\rangle, 0_{M_{i+1}},\dots,0_{M_t}).$$

Then $\phi(e_{1,1}),\dots,\phi(e_{1,n_1}),\phi(e_{2,1}),\dots,\phi(e_{t-1,n_{t-1}}),\phi(e_{t,1}),\dots,\phi(e_{t,n_t})$ is a basis for the $\mathbb{C}$-vector space $M_1\oplus\cdots\oplus M_t$ and since ϕ is a $\mathbb{C}$-linear isomorphism, the sequence of vectors $e_{1,1},\dots,e_{1,n_1},e_{2,1},\dots,e_{t-1,n_{t-1}},e_{t,1},\dots,e_{t,n_t}$ (ordered in this way) is a basis for V as a vector space over $\mathbb{C}$.

The map ϕ is $\mathbb{C}[X]$-linear so

$$\begin{aligned}\phi(Te_{i,j}) = \phi(X.e_{i,j}) = X.\phi(e_{i,j}) &= \begin{cases}\phi(e_{i,j+1})+\lambda_i.\phi(e_{i,j}) & \text{if } j<n_i\\ \lambda_i.\phi(e_{i,j}) & \text{if } j=n_i\end{cases}\\ &= \begin{cases}\phi(e_{i,j+1}+\lambda_i.e_{i,j}) & \text{if } j<n_i\\ \phi(\lambda_i.e_{i,j}) & \text{if } j=n_i\end{cases}.\end{aligned}$$

Since ϕ is a $\mathbb{C}$-linear bijection we conclude that T has the required form.

For the last part, certainly the λ_is are eigenvalues of T since $J(\lambda,n)(0,\dots,0,1)^t = \lambda(0,\dots,0,1)^t$. On the other hand $(J(\lambda,n)-\lambda I)^n=0$ and so the minimal polynomial for T divides $(X-\lambda_1)^{n_1}\cdots(X-\lambda_t)^{n_t}$ and hence all the roots of the minimal polynomial are in the set $\{\lambda_1, \ldots, \lambda_t\}$. Finally, every eigenvalue of T is a root of the minimal polynomial and so the claim is proved. □

Remark 12.6. ⚠The λ_is in the theorem need not be distinct.

The fact that $\mathbb{C}$ is algebraically closed *i.e.* every polynomial with coefficients in $\mathbb{C}$ has a root in $\mathbb{C}$ is vital to the Jordan normal form (and we used this fact when we appealed to Remark 8.2), but there is another simple form available more generally.

Theorem 12.7 (Rational Canonical Form)**.** *Suppose that V is a finite-dimensional vector space over $\mathbb{F}$ and $T : V \to V$ is linear and not identically 0. Then there are monic polynomials $f_1 \mid \cdots \mid f_r$ of degree $n_1, \ldots, n_r$ respectively and with f_1 non-constant, and a basis for V such that the matrix for T in this basis is*

$$\begin{pmatrix} C(f_1) & 0_{n_1 \times n_2} & \cdots & 0_{n_1 \times n_r} \\ 0_{n_2 \times n_1} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0_{n_{r-1} \times n_r} \\ 0_{n_r \times n_1} & \cdots & 0_{n_r \times n_{r-1}} & C(f_r) \end{pmatrix}$$

where $0_{n \times m}$ is the all zeros matrix in $M_{n,m}(\mathbb{F})$, and $C(f)$ is[37] *the $n \times n$ matrix, called the* ***companion matrix****, for the monic $f(X) = X^n + a_{n-1}X^{n-1} + \cdots + a_1X + a_0$,*

$$\begin{pmatrix} 0 & \cdots & \cdots & 0 & -a_0 \\ 1 & \ddots & & \vdots & -a_1 \\ 0 & \ddots & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & 0 & -a_{n-2} \\ 0 & \cdots & 0 & 1 & -a_{n-1} \end{pmatrix}.$$

The minimal polynomial for T is f_r and the characteristic polynomial is $f_1 \cdots f_r$.

Proof. The argument is really the same as that for producing the Jordan Normal Form except we apply the Invariant Factor Form of the Structure Theorem rather than the Primary Form.

As before, we regard V as an $\mathbb{F}[X]$-module in the way described in Example 9.12. Since $\mathbb{F}$ is a subring of $\mathbb{F}[X]$ and V is finite dimensional as an $\mathbb{F}$-vector space, the module V is finitely generated by Example 11.10.

Since $\mathbb{F}[X]$ is a PID we may apply the Structure Theorem (Invariant Factor Form, Theorem 12.1). Then we get polynomials $f_1 \mid \cdots \mid f_r$ with $f_1 \not\sim 1$ and

$$\phi : V \to (\mathbb{F}[X]/\langle f_1 \rangle) \oplus \cdots \oplus (\mathbb{F}[X]/\langle f_r \rangle)$$

[37] If $n = 1$ then $C(f) = (-a_0)$.

an $\mathbb{F}[X]$-linear bijection. In particular, ϕ is an $\mathbb{F}$-linear bijection but V is finite-dimensional and $\mathbb{F}[X]/\langle 0\rangle$ is infinite dimensional so $f_i \in \mathbb{F}[X]^*$ for all $1 \leqslant i \leqslant r$. Thus we may put $n_i := \deg f_i$ and may suppose that each f_i is monic (since multiplying by a unit does not change the ideal).

For $1 \leqslant i \leqslant r$ we write $M_i := \mathbb{F}[X]/\langle f_i\rangle$ and let $(e_{i,j})_{j=1}^{n_i}$ be such that

$$\phi(e_{i,j}) = (0_{M_1}, \dots, 0_{M_{i-1}}, X^{j-1} + \langle f_i\rangle, 0_{M_{i+1}}, \dots, 0_{M_r}).$$

Then $\phi(e_{1,1}), \dots, \phi(e_{1,n_1}), \phi(e_{2,1}), \dots, \phi(e_{r-1,n_{r-1}}), \phi(e_{r,1}), \dots, \phi(e_{r,n_r})$ is a basis for the $\mathbb{F}$-vector space $M_1 \oplus \cdots \oplus M_r$ and since ϕ is an $\mathbb{F}$-linear isomorphism, the sequence of vectors $e_{1,1}, \dots, e_{1,n_1}, e_{2,1}, \dots, e_{r-1,n_{r-1}}, e_{r,1}, \dots, e_{r,n_r}$ (ordered in this way) is a basis for V as a vector space over $\mathbb{F}$.

Write $f_i(X) = X^{n_i} + a^{(i)}_{n_i-1}X^{n_i-1} + \cdots + a^{(i)}_1 X + a^{(i)}_0$ for $1 \leqslant i \leqslant r$. Then since ϕ is $\mathbb{F}[X]$-linear we have

$$\begin{aligned}\phi(T.e_{i,j}) = \phi(X.e_{i,j}) = X.\phi(e_{i,j}) &= \begin{cases}\phi(e_{i,j+1}) & \text{if } j < n_i\\ -a^{(i)}_0.\phi(e_{i,1}) - \cdots - a^{(i)}_{n_i-1}.\phi(e_{i,n_i}) & \text{if } j = n_i\end{cases}\\ &= \begin{cases}\phi(e_{i,j+1}) & \text{if } j < n_i\\ \phi(-a^{(i)}_0.e_{i,1} - \cdots - a^{(i)}_{n_i-1}.e_{i,n_i}) & \text{if } j = n_i\end{cases}.\end{aligned}$$

Since ϕ is an $\mathbb{F}$-linear bijection we conclude that T has the required form.

For the last part we first show that for a monic polynomial f the minimal polynomial of $C(f)$ is f where $f \in \mathbb{F}[X]^*$ has degree n: First, the characteristic polynomial of $C(f)$ can be computed using the Laplace expansion so

$$\begin{aligned}\det(tI - C(f)) &= \det\begin{pmatrix} t & 0 & \cdots & 0 & a_0\\ -1 & \ddots & \ddots & \vdots & a_1\\ 0 & \ddots & \ddots & 0 & \vdots\\ \vdots & \ddots & \ddots & t & a_{n-2}\\ 0 & \cdots & 0 & -1 & t + a_{n-1}\end{pmatrix}\\ &= (-1)^{n+1}a_0 t^0(-1)^{n-1} + \cdots + (-1)^{n+1+i}a_i t^i(-1)^{n-1-i} + \cdots\\ &\qquad + (-1)^{2n-1}a_{n-2}t^{n-2}(-1)^1 + (-1)^{2n}(t + a_{n-1})t^{n-1}(-1)^0 = f(t).\end{aligned}$$

By Cayley-Hamilton, $C(f)$ satisfies f. Moreover, for $0 \leqslant r \leqslant n-1$ the first column of $C(f)^r$ is $(0, \dots, 0, 1, 0, \dots, 0)^t$ where the 1 is in the $(r+1)$st position, thus the matrices $I, C(f), \dots, C(f)^{n-1}$ are linearly independent over $\mathbb{F}$ and hence the degree of the minimal polynomial is at least n; we conclude the minimal polynomial is f.

Since $f_i \mid f_r$ for all $1 \leqslant i \leqslant r$ we see that $f_r(T) = 0$. On the other hand T is conjugate to a matrix containing $C(f_r)$ which we have seen has minimal polynomial f_r and hence f_r is the minimal polynomial of T.

The characteristic polynomial is invariant under change of basis, and hence the characteristic polynomial of T is the product of the characteristic polynomials of the companion matrices in the rational canonical form. It follows that it is $\prod_{i=1}^{r} f_i$ as claimed. □

Remark 12.8. The Rational Canonical Form is also sometimes called the **Frobenius Normal Form**.

13 Bases and matrices; computing with modules

As with vector spaces, generation in modules has an allied concept of linear independence: Suppose that M is an R-module and $(x_i)_{i\in I}$ is a family of elements of M. We say that $(x_i)_{i\in I}$ is **linearly independent** if whenever $S \subset I$ is finite and

$$\sum_{s\in S} \lambda_s.x_s = 0_M \text{ with } \lambda_s \in R \text{ for all } s \in S,$$

then $\lambda_s = 0_R$ for all $s \in S$. We say that $(x_i)_{i\in I}$ is a **basis** for M if it is both linearly independent and generating. An R-module M with a basis is said to be **free**.

Remark 13.1. In vector spaces this terminology coincides with existing terminology.

Remark 13.2. ⚠Unlike vector spaces[38] not every module is free: for example $\mathbb{Z}_N$ does not even have a non-empty independent set as a $\mathbb{Z}$-module, despite being a finitely generated module over a PID, though it *does* have a basis as a $\mathbb{Z}_N$-module.

In particular, while the Structure Theorem (say Theorem 12.1) *does* afford us a nice 'basis-like' set of generators. In particular the module $(R/\langle a_1\rangle)\oplus\cdots\oplus(R/\langle a_r\rangle)$ is generated by

$$(1_{R/\langle a_1\rangle}, 0_{R/\langle a_2\rangle}, \dots, 0_{R/\langle a_r\rangle}), \dots, (0_{R/\langle a_1\rangle}, \dots, 0_{R/\langle a_{r-1}\rangle}, 1_{R/\langle a_r\rangle})$$

but in general this is *not* a basis.

Lemma 13.3. *Suppose that M is an R-module and $(x_i)_{i\in I}$ is a family of elements of M. Then there is a unique R-linear map $\Psi : \bigoplus_{i\in I} R \to M$ such that $\Psi(e_i) = x_i$ for all $i \in I$ where $e_i \in \bigoplus_{i\in I} R$ has 1_R in the position indexed by i and 0_R elsewhere.*[39]

Moreover, $(x_i)_{i\in I}$ is linearly independent if and only if Ψ is injective; it is generating if and only if Ψ is surjective; so it is a basis if and only if Ψ is bijective.

[38] Assuming the Axiom of Choice, every vector space has a basis [Lan02, Theorem 5.1, Chapter III]. (There are different types of basis in different areas of mathematics, for example a Schauder basis is a type of basis suitable for Banach spaces but Schauder bases are not in general bases in the sense we use here. When necessary, the type of basis we are interested in here is disambiguated by calling it a Hammel basis.) It turns out that the use of the Axiom of Choice is unavoidable in the strong sense that if every vector space is assumed to have a basis then (in ZF) the Axiom of Choice follows [Bla84, Theorem 1].

[39] We *are* allowing infinite and unordered indexing sets I, but in the special case $I = \{1, \dots, n\}$ then $e_i = (0_R, \dots, 0_R, 1_R, 0_R, \dots, 0_R)$ where the 1_R is in the ith position.

Proof. The map

$$\Psi : \bigoplus_{i\in I} R \to M; r \mapsto \sum_{i:r_i\neq 0_R} r_i.x_i$$

is well-defined since the sum on the right is finite by definition of the direct sum, and it satisfies $\Psi(e_i) = x_i$ in view of (M1). To see that Ψ is a homomorphism note that that for $u, v \in \bigoplus_{i\in I} R$ we have

$$\begin{aligned}\Psi(u+v) &= \textstyle\sum_{i:u_i+v_i\neq 0_R} (u_i+v_i).x_i\\ &= \textstyle\sum_{i:u_i\neq 0_R \text{ or } v_i\neq 0_R} (u_i+v_i).x_i\\ &= \textstyle\sum_{i:u_i\neq 0_R \text{ or } v_i\neq 0_R} u_i.x_i + \sum_{i:u_i\neq 0_R \text{ or } v_i\neq 0_R} v_i.x_i\\ &= \textstyle\sum_{i:u_i\neq 0_R} u_i.x_i + \sum_{i:v_i\neq 0_R} v_i.x_i = \Psi(u)+\Psi(v).\end{aligned}$$

$$\begin{aligned}&J := \{i : u_i + v_i \neq 0_R\}\\ &\qquad \subset K := \{i : u_i \neq 0_R \textit{ or } v_i \neq 0_R\}\\ &\textit{and } \textstyle\sum_{i\in K\setminus J} (u_i+v_i).x_i = 0_M\\ &\textstyle\sum_{i:u_i=0_R \textit{ and } v_i\neq 0_R} u_i.x_i = 0_M\\ &\textit{and } \textstyle\sum_{i:u_i\neq 0_R \textit{ and } v_i=0_R} v_i.x_i = 0_M\end{aligned}$$

Similarly for $u \in \bigoplus_{i\in I} R$ and $r \in R$ we have

$$\Psi(r.u) = \sum_{i:ru_i\neq 0_R} (ru_i).x_i = \sum_{i:u_i\neq 0_R} (ru_i).x_i = \sum_{i:u_i\neq 0_R} r.(u_i.x_i) = r.\left(\sum_{i:u_i\neq 0_R} u_i.x_i\right) = r.\Psi(u)$$

as required.

Finally, for uniqueness, given two such maps Ψ and Φ the map $\Pi := \Psi - \Phi$ has $\Pi(e_i) = 0_M$ for all $i \in I$. But if $u \in \bigoplus_{i\in I} R$ then $u = \sum_{i:u_i\neq 0_R} u_i.e_i$ and hence by linearity we have $\Pi(u) = \sum_{i:u_i\neq 0_R} u_i.\Pi(e_i) = 0_M$ and so $\Psi = \Phi$ as required.

The 'moreover' part follows by unpacking the definitions. □

Remark 13.4. In view of the above an R-module M is free if and only if $M \cong \bigoplus_{i\in I} R$ for some indexing set I.

Proposition 13.5. *Suppose that R is a non-trivial commutative ring and M is an R-module with a basis of size n. Then any generating set has size at least n.*

Proof. In view of Remark 13.4 $M \cong R^n$. Now apply Lemma 11.20 with all the ideals equal to $\{0\}$ (which are proper since R is non-trivial). □

Remark 13.6. This proposition implies that if R is a non-trivial commutative ring then any two bases of the R-module M have the same size. Thus, in this case if M has a finite basis of size n we say M has **rank** n and this is well-defined.

Remark 13.7. ⚠An independent subset of a rank n module having size n need not be a basis: $\{2\}$ is an independent subset of the rank one module $\mathbb{Z}$ having size 1 but it is not a basis for $\mathbb{Z}$. On the other hand, it *is* true (in our setting of commutative rings) that a generating subset of a rank n module having size n is a basis, though this takes some work (see *e.g.* [Lam99, §1B]). Moreover, it is also true (again, in our setting of commutative rings) that a free submodule of a rank n module must have rank at most n (see *e.g.* [Lam99, §1D]).

Matrices and Smith Normal Form

Finite bases are particularly important because they let us write linear maps as matrices. Given a commutative ring R we write $M_{n,m}(R)$ for the set of $n \times m$ matrices with values in R, and $M_n(R) := M_{n,n}(R)$. This notation generalises the matrix rings of Proposition 4.25.

Given R-modules M and N with bases $\mathcal{X} = (x_i)_{i=1}^m$ and $\mathcal{Y} = (y_i)_{i=1}^n$ respectively there is a bijection $\Phi : L(M,N) \to M_{n,m}(R)$ such that

$$Tx_i = \sum_{j=1}^n \Phi(T)_{j,i}.y_j \text{ for all } 1 \leqslant i \leqslant m. \tag{13.1}$$

We call $\Phi(T)$ the **matrix of T with respect to the bases $\mathcal{X}$ and $\mathcal{Y}$**. The inverse of this map takes $A \in M_{n,m}(R)$ to the R-linear map

$$M \to N; \sum_{i=1}^m \lambda_i.x_i \mapsto \sum_{j=1}^n \left(\sum_{i=1}^m A_{j,i}\lambda_i\right).y_j,$$

which is well-defined since $\mathcal{X}$ is a basis for M.

Remark 13.8. The free module R^n come with the so called **standard bases** $\mathcal{E}_n$, that is the set of elements $e_i = (0_R, \dots, 0_R, 1_R, 0_R, \dots 0_R)$ with 1_R in the ith position. We can give $M_{n,m}(R)$ the structure of an R-module by taking the R-module structure on $L(R^m, R^n)$ (afforded by Proposition 10.2) and using the bijection to above to bring it over to $M_{n,m}(R)$. We shall not do this here, but the next proposition will do this for the ring structure on $\mathrm{End}_R(R^n) = L(R^n, R^n)$.

Proposition 13.9. *Suppose that R is a commutative ring. Then $M_n(R)$ is a ring with*

$$A + B = (A_{i,j} + B_{i,j})_{i,j=1}^n \text{ and } AB = \left(\sum_{k=1}^n A_{i,k}B_{k,j}\right)_{i,j=1}^n \text{ for } A, B \in M_n(R),$$

zero $(0_R)_{i,j=1}^n$, multiplicative identity I where $I_{i,i} = 1_R$ for $1 \leqslant i \leqslant n$ and $I_{i,j} = 0_R$ for $i \neq j$, and $-A = (-A_{i,j})_{i,j=1}^n$ for $A \in M_n(R)$.

Proof. Let $\Phi : \mathrm{End}_R(R^n) \to M_n(R)$ be the bijection taking an R-linear map $R^n \to R^n$ to its matrix w.r.t. to the standard basis $\mathcal{E}_n$ on both the domain and the codomain, as in (13.1). In Remark 10.6 we saw that $\mathrm{End}_R(R^n)$ is a ring and the bijection Φ then makes $M_n(R)$ into a ring by putting $A + B := \Phi(\Phi^{-1}(A) + \Phi^{-1}(B))$, $AB := \Phi(\Phi^{-1}(A) \circ \Phi^{-1}(B))$, $-A := \Phi(-\Phi^{-1}(A))$, $0_{M_n(R)} := \Phi(0_{\mathrm{End}_R(R^n)})$ and $1_{M_n(R)} := \Phi(1_{\mathrm{End}_R(R^n)})$.

The remainder of the proposition is computing what these definitions yield. First, the zero map of $\mathrm{End}_R(R^n)$ maps to the zero matrix and the multiplicative identity maps to the matrix I described in the proposition just by considering (13.1). Secondly, if $A, B \in M_n(R)$

then

$$\begin{aligned}\sum_{j=1}^{n}(A+B)_{j,i}.e_j = \Phi^{-1}(A+B)e_i &= (\Phi^{-1}(A)+\Phi^{-1}(B))(e_i)\\ &= \Phi^{-1}(A)e_i + \Phi^{-1}(B)e_i\\ &= \sum_{j=1}^{n}A_{j,i}.e_j + \sum_{j=1}^{n}B_{j,i}.e_j = \sum_{j=1}^{n}(A_{j,i}+B_{j,i}).e_j,\end{aligned}$$

and since $(e_j)_{j=1}^{n}$ is a basis, matrix addition has the form described. Since additive inverses are unique and we have seen that the zero of $M_n(R)$ is the all zeros matrix, this also gives that $-A$ has the described form. Finally for A and B again,

$$\begin{aligned}\sum_{k=1}^{n}(AB)_{k,i}.e_k = \Phi^{-1}(AB)e_i &= (\Phi^{-1}(A)\circ\Phi^{-1}(B))(e_i)\\ &= \Phi^{-1}(A)(\Phi^{-1}(B)e_i)\\ &= \Phi^{-1}(A)\left(\sum_{j=1}^{n}B_{j,i}.e_j\right)\\ &= \sum_{j=1}^{n}B_{j,i}.\left(\Phi^{-1}(A)e_j\right)\\ &= \sum_{j=1}^{n}B_{j,i}.\left(\sum_{k=1}^{n}A_{k,j}.e_k\right) = \sum_{k=1}^{n}\left(\sum_{j=1}^{n}A_{k,j}B_{j,i}\right).e_k\end{aligned}$$

and the result is proved. □

Remark 13.10. It is perfectly reasonable to define matrix rings $M_n(R)$ when R is not commutative using the identities in the above proposition. However, even for $n = 1$ they do not necessarily arise as R-linear maps $R \to R$ because (as in Remark 11.13) multiplication by scalars need not be linear.

Remark 13.11. In particular the above provides a proof of Proposition 4.25.

Remark 13.12. The group of units of $M_n(R)$ is denoted $\mathrm{GL}_n(R)$. In fact the usual formulae for inverting matrices work with matrices over commutative rings with the modification that rather than having the determinant non-zero we need it to be a unit. The determinant can be defined in all the usual ways it was when considering the case when R is a field, and

$$\det A = \sum_{\sigma\in S_n}\mathrm{sgn}(\sigma)\prod_{i=1}^{n}A_{i,\sigma(i)}.$$

We say that $A, B \in M_{n,m}(R)$ are **equivalent** if there are matrices $S \in \mathrm{GL}_n(R)$ and $T \in \mathrm{GL}_m(R)$ such that $A = SBT$, and that an $n \times m$ matrix A is in **Smith Normal Form** if there are elements $a_1 \mid a_2 \mid \cdots \mid a_{\min\{n,m\}}$ such that $A_{i,i} = a_i$ for $1 \leqslant i \leqslant \min\{n,m\}$ and $A_{i,j} = 0_R$ otherwise. Note the divisibility condition so that, for example,

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -2 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \text{ and } \begin{pmatrix} 5 & 0 & 0 \\ 0 & 25 & 0 \\ 0 & 0 & 100 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

are both in Smith Normal Form over $\mathbb{Z}$, however neither of the matrices

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix} \text{ and } \begin{pmatrix} 3 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

is in Smith Normal Form over $\mathbb{Z}$, although they are both in Smith Normal Form over $\mathbb{Q}$.

Theorem 13.13 (Smith Normal Form)**.** *Suppose that R is a PID and $A \in M_{n,m}(R)$. Then A is equivalent to a matrix in Smith Normal Form. Moreover, the entries of this matrix are unique up to association.*

Remark 13.14. We shall not prove this, but rather we shall give an algorithm for how to find the equivalent Smith Normal Form of a matrix (and the invertible matrices corresponding to the equivalence).

There are particular types of elements of $\mathrm{GL}_n(R)$ whose left and right multiplication correspond to row and column operations respectively. For A an $n \times m$ matrix we write $c_1, \ldots, c_m \in R^n$ for the columns of A so $A = (c_1^t, \ldots, c_m^t)$, and $r_1, \ldots, r_n \in R^m$ for the rows of A so that $A = (r_1, \ldots, r_n)^t$. Write $E_n(i,j)$ for the $n \times n$ matrix with 0s everywhere except for row i and column j where the entry is 1.

(i) *(Transvections)* Given $1 \leqslant i, j \leqslant m$ with $i \neq j$ and $\lambda \in R$ put $P_m(i,j;\lambda) = I_m + \lambda E_m(i,j)$. We write

$$A \xrightarrow{c_j \mapsto c_j + \lambda c_i} AP_m(i,j;\lambda).$$

to mean that the matrix A after the column operation replacing c_j by $c_j + \lambda c_i$ is the matrix A post-multiplied by $P_m(i,j;\lambda)$. This can be checked by direct calculation.

Similarly

$$A \xrightarrow{r_i \mapsto r_i + \lambda r_j} P_n(i,j;\lambda)A$$

means that the matrix A after the row operation replacing r_i by $r_i + \lambda r_j$ is the matrix A pre-multiplied by $P_n(i,j;\lambda)$. Again this can be checked by direct calculation.

(ii) *(Dilations)* Given $1 \leqslant i \leqslant m$ and $u \in U(R)$ let $D_m(i; u) := I_m + (u-1)E_m(i,i)$ so that $D_m(i; u)$ is the matrix with 1s on the diagonal except for the ith element which is u, and 0s elsewhere. As above we write

$$A \xrightarrow{c_i \mapsto uc_i} AD_m(i; u) \text{ and } A \xrightarrow{r_i \mapsto ur_i} D_n(i; u)A$$

to mean the matrix A with column c_i replaced by uc_i *etc.*

(iii) *(Interchanges)* Given $1 \leqslant i, j \leqslant m$ let $S_m(i,j) = I_m + E_m(i,j) + E_m(j,i) - E_m(i,i) - E_m(j,j)$. By

$$A \xrightarrow{c_i \leftrightarrow c_j} AS_m(i,j) \text{ and } A \xrightarrow{r_i \leftrightarrow r_j} S_n(i,j)A$$

we mean the matrix A with c_i and c_j swapped *etc.*

Remark 13.15. These three types of operations are the **elementary column and row operations** respectively. The matrices are all invertible, since their pre- and post- multiplication corresponds to row and column operations respectively, and these operations are easily seen to be invertible. This invertibility is the reason for restricting dilates to elements of the group of units.

In view of the invertibility of these matrices we see that applying these elementary row and column operations to a matrix preserves equivalence of matrices.

Remark 13.16. The subgroup of $\mathrm{GL}_n(R)$ generated by the elementary row operations is denoted $\mathrm{GE}_n(R)$. Of course $\mathrm{GE}_n(R) \leqslant \mathrm{GL}_n(R)$, and for some rings it is a proper subgroup (in fact the ring A in Exercise III.9 is such an example [Gel77], and it is an open problem [SZ14, (3), §7] whether every PID with $\mathrm{GE}_2(R) = \mathrm{GL}_2(R)$ is Euclidean; certainly if R is Euclidean then it is a PID and it happens that $\mathrm{GE}_2(R) = \mathrm{GL}_2(R)$.

Putting a matrix into Smith Normal Form using elementary operations

Suppose that R is a Euclidean Domain with Euclidean function f, and $A \in M_{n,m}(R)$. We shall proceed iteratively either decreasing the quantity $\mu(A) := \min_{i,j:A_{i,j} \neq 0_R} f(A_{i,j})$ or leaving it the same and increasing the number of 0_R entries.

Suppose that $A_{i,j} = \mu(A)$. For $j' \neq j$ the Euclidean function tells us that either

(i) there is $q \in R$ and $r \in R^*$ such that $A_{i,j'} = qA_{i,j} + r$ where $f(r) < A_{i,j}$ and so $\mu(AP_m(j, j'; -q)) < \mu(A)$;

(ii) or there is $q \in R$ such that $A_{i,j'} = qA_{i,j}$ and so $AP_m(j, j'; -q)$ has an extra zero in it unless $A_{i,j'} = 0_R$.

This process eventually terminates with $A_{i,j'} = 0_R$ for all $j' \neq j$. All the operations were column operations adding to every column except the jth, and thus we can proceed similarly to eliminate all the non-zero entries (apart from the ith) in the jth column. The matrix looks like:

$$\begin{pmatrix} * & \cdots & * & 0_R & * & \cdots & * \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ * & \cdots & * & 0_R & * & \cdots & * \\ 0_R & \cdots & 0_R & A_{i,j} & 0_R & \cdots & 0_R \\ * & \cdots & * & 0_R & * & \cdots & * \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ * & \cdots & * & 0_R & * & \cdots & * \end{pmatrix}$$

After this process, if $A_{i,j}$ does not divide every entry of the matrix, then take a row (or column) with an entry it does not divide and add it to the ith row, or jth column. Then an application of (i) above reduces $\mu(A)$. We can repeat this whole process and it must eventually terminate since either $\mu(A)$ decreases, or it stays the same and the number of zeros in the ith row or jth column increases.

A final column and row switch moves the $A_{i,j}$ to the top left of the matrix, and we now repeat the process with the bottom right matrix in $M_{n-1,m-1}(R)$. Any common factor of all the matrix entries remains under the application of the elementary operations, and so when the process terminates we have a matrix in Smith Normal Form as required.

Remark 13.17. ⚠While the above process is sure to work, *any* sequence of operations is allowed so in practice there can be better ways to proceed.

A **finite presentation** of an R-module M is a linear map $T : R^n \to R^m$ and isomorphism $R^m/\operatorname{Im} T \to M$. A module M is said to be **finitely presented** if it has a finite presentation.

Remark 13.18. For comparison, Lemma 13.3 shows that for *every* finitely generated module M there is an R-linear surjection $T : R^n \to M$ and hence by the First Isomorphism Theorem $M \cong R^n/\ker T$.

Remark 13.19. ⚠There are finitely generated modules that are not finitely presented (see *e.g.* Exercise E.16), but the next result shows that this cannot be so for PIDs.

Proposition 13.20. *Suppose that R is a PID and M is a finitely generated R-module. Then M is finitely presented.*

Proof. Theorem 12.1 tells us that there are elements $a_1 \mid \cdots \mid a_r$ such that $M \cong (R/\langle a_1\rangle) \oplus \cdots (R/\langle a_r\rangle)$ so

$$T : R^n \to R^n; x \mapsto (a_1x_1, \dots, a_nx_n)$$

is an R-linear map such that $M \cong R^n/\operatorname{Im} T$, as required. □

As in Remark 13.8, linear maps $T : R^n \to R^m$ are in one-to-one correspondence with matrices $A \in M_{n,m}(R)$ and so we often simply speak of the finitely presented module R^m/AR^n.

Lemma 13.21. *Suppose that $A, B \in M_{n,m}(R)$ are equivalent matrices. Then $R^m/AR^n \cong R^m/BR^n$.*

Proof. Let $PA = BQ$ for $P \in \mathrm{GL}_m(R)$ and $Q \in \mathrm{GL}_n(R)$. Then the map

$$R^m/AR^n \to R^m/BR^n; x + AR^n \mapsto Px + BR^n$$

is a well-defined R-linear isomorphism. First, $x - x' \in AR^n$ if and only if $P(x - x') \in PAR^n = BQR^n$ since P is invertible. But since Q is invertible $QR^n = R^n$ and so $x - x' \in AR^n$ if and only if $Px - Px' \in BR^n$ so that the map is well-defined and injective. Since P is invertible the map is surjective and it is easily seen to be linear. □

Remark 13.22. ⚠The converse of this lemma is *not* true: if $R^m/AR^n \cong R^m/BR^n$ it need not be the case that A and B are equivalent. See Exercise E.15.

Remark 13.23. In view of this lemma we see that putting a matrix in SNF can be used to produce a particularly simple representation of a finitely presented module.

Describing the structure of a commutative group using the SNF

Suppose that G is a commutative group with generators g_1, g_2, g_3, g_4, g_5 and relations

$$2.g_1 + 6.g_2 - 8.g_3 = 0, g_1 + g_2 + g_4 = 0, \text{ and } 5.g_1 + 5.g_4 + 25.g_5 = 0.$$

This group is isomorphic to $\mathbb{Z}_{10} \oplus \mathbb{Z}^2$, and to show this we use the Smith Normal Form. First we put the relation matrix, R, into Smith Normal Form:

$$R := \begin{pmatrix} 2 & 6 & -8 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 5 & 0 & 0 & 5 & 25 \end{pmatrix} \xrightarrow{r_1 \leftrightarrow r_2} \begin{pmatrix} 1 & 1 & 0 & 1 & 0 \\ 2 & 6 & -8 & 0 & 0 \\ 5 & 0 & 0 & 5 & 25 \end{pmatrix} \xrightarrow{\substack{c_2 \mapsto c_2 - c_1 \\ c_4 \mapsto c_4 - c_1}}$$

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 2 & 4 & -8 & -2 & 0 \\ 5 & -5 & 0 & 0 & 25 \end{pmatrix} \xrightarrow{\substack{r_2 \mapsto r_2 - 2r_1 \\ r_3 \mapsto r_3 - 5r_1}} \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 4 & -8 & -2 & 0 \\ 0 & -5 & 0 & 0 & 25 \end{pmatrix} \xrightarrow{r_2 \mapsto r_2 + r_3}$$

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & -1 & -8 & -2 & 25 \\ 0 & -5 & 0 & 0 & 25 \end{pmatrix} \xrightarrow{r_3 \mapsto r_3 - 5r_2} \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & -1 & -8 & -2 & 25 \\ 0 & 0 & 40 & 10 & -100 \end{pmatrix} \xrightarrow{\substack{c_3 \mapsto c_3 - 8c_2 \\ c_4 \mapsto c_4 - 2c_2 \\ c_5 \mapsto c_5 + 25c_2}}$$

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 40 & 10 & -100 \end{pmatrix} \xrightarrow{c_3 \leftrightarrow c_4} \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 10 & 40 & -100 \end{pmatrix} \xrightarrow{\substack{c_4 \mapsto c_4 - 4c_3 \\ c_5 \mapsto c_5 + 10c_3}}$$

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 10 & 0 & 0 \end{pmatrix}.$$

Thus we have $P \in \mathrm{GL}_3(\mathbb{Z})$ and $Q \in \mathrm{GL}_5(\mathbb{Z})$ such that

$$P \begin{pmatrix} 2 & 6 & -8 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 5 & 0 & 0 & 5 & 25 \end{pmatrix} Q = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 10 & 0 & 0 \end{pmatrix}.$$

We can compute the matrix Q by applying the column operations to the identity matrix:

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \xrightarrow{\substack{c_2 \mapsto c_2 - c_1 \\ c_4 \mapsto c_4 - c_1}} \begin{pmatrix} 1 & -1 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \xrightarrow{\substack{c_3 \mapsto c_3 - 8c_2 \\ c_4 \mapsto c_4 - 2c_2 \\ c_5 \mapsto c_5 + 25c_2}}$$

$$\begin{pmatrix} 1 & -1 & 8 & 1 & -25 \\ 0 & 1 & -8 & -2 & 25 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \xrightarrow{c_3 \leftrightarrow c_4} \begin{pmatrix} 1 & -1 & 1 & 8 & -25 \\ 0 & 1 & -2 & -8 & 25 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \xrightarrow{\substack{c_4 \mapsto c_4 - 4c_3 \\ c_5 \mapsto c_5 + 10c_3}}$$

$$\begin{pmatrix} 1 & -1 & 1 & 4 & -15 \\ 0 & 1 & -2 & 0 & 5 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & -4 & 10 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Similarly we can compute P:

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \xrightarrow{r_1 \leftrightarrow r_2} \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \xrightarrow{\substack{r_2 \mapsto r_2 - 2r_1 \\ r_3 \mapsto r_3 - 5r_1}} \begin{pmatrix} 0 & 1 & 0 \\ 1 & -2 & 0 \\ 0 & -5 & 1 \end{pmatrix}$$

$$\xrightarrow{r_2 \mapsto r_2 + r_3} \begin{pmatrix} 0 & 1 & 0 \\ 1 & -7 & 1 \\ 0 & -5 & 1 \end{pmatrix} \xrightarrow{r_3 \mapsto r_3 - 5r_2} \begin{pmatrix} 0 & 1 & 0 \\ 1 & -7 & 1 \\ -5 & 30 & -4 \end{pmatrix}.$$

This gives us a well-defined isomorphism

$$\begin{aligned} \phi : G &\to \mathbb{Z}_{10} \oplus \mathbb{Z}^2 \\ z_1.g_1 + \cdots + z_5.g_5 &\mapsto (z_1 - 2z_2 + z_4, 4z_1 + z_3 - 4z_4, -15z_1 + 5z_2 + 10z_4 + z_5). \end{aligned}$$

For a matrix A we write $\mathrm{RowSpan}(A)$ for the $\mathbb{Z}$-module generated by the rows of A. To see that ϕ is a well-defined injection note:

$$
\begin{aligned}
& z_1.g_1 + \cdots + z_5.g_5 = z_1'.g_1 + \cdots + z_5'.g_5 && \Big\rangle\ \textit{Definition of } G \\
\Leftrightarrow & (z_1 - z_1', \ldots, z_5 - z_5') \in \mathrm{RowSpan}(R) && \Big\rangle\ \textit{Since } P \in \mathrm{GL}_3(\mathbb{Z}) \\
\Leftrightarrow & (z_1 - z_1', \ldots, z_5 - z_5') \in \mathrm{RowSpan}(PR) && \Big\rangle\ \textit{Since } Q \in \mathrm{GL}_5(\mathbb{Z}) \\
\Leftrightarrow & (z_1 - z_1', \ldots, z_5 - z_5')Q \in \mathrm{RowSpan}(PRQ) && \Big\rangle\ \textit{Design of } PRQ \\
\Leftrightarrow & (z_1 - z_1', \ldots, z_5 - z_5')Q \in \{(u, -v, 10w, 0, 0) : u, v, w \in \mathbb{Z}\} && \Big\rangle\ \textit{Definition of } \phi \\
\Leftrightarrow & \phi((z_1 - z_1').g_1 + \cdots + (z_5 - z_5').g_5) = 0 \\
\Leftrightarrow & \phi(z_1.g_1 + \cdots + z_5.g_5) = \phi(z_1'.g_1 + \cdots + z_5'.g_5).
\end{aligned}
$$

The map ϕ is also certainly $\mathbb{Z}$-linear (in fact we have already used this to some extent above). Moreover, since ϕ is well-defined and $\phi(g_5) = (0,0,1)$, $\phi(g_3) = (0,1,0)$, and $\phi(g_1 - 4.g_3 + 15.g_5) = (1,0,0)$ we see that the image of ϕ contains a generating set for the codomain and hence ϕ is a surjection. The claim that ϕ is an isomorphism is complete.

Computing the rational canonical form using the SNF

Suppose we wish to compute the rational canonical form of the matrix

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

We begin by putting the matrix $XI - A$ in Smith Normal Form over the Euclidean domain $\mathbb{Q}[X]$:

$$
\begin{pmatrix} X-1 & 1 & -1 \\ 0 & X & -1 \\ 0 & -1 & X \end{pmatrix}
\xrightarrow{c_1 \leftrightarrow c_2}
\begin{pmatrix} 1 & X-1 & -1 \\ X & 0 & -1 \\ -1 & 0 & X \end{pmatrix}
\xrightarrow{\substack{c_2 \mapsto c_2 - (X-1)c_1 \\ c_3 \mapsto c_3 + c_1}}
$$

$$
\begin{pmatrix} 1 & 0 & 0 \\ X & X - X^2 & X-1 \\ -1 & X-1 & X-1 \end{pmatrix}
\xrightarrow{\substack{r_2 \mapsto r_2 - X r_1 \\ r_3 \mapsto r_3 + r_1}}
\begin{pmatrix} 1 & 0 & 0 \\ 0 & X - X^2 & X-1 \\ 0 & X-1 & X-1 \end{pmatrix}
\xrightarrow{c_2 \leftrightarrow c_3}
$$

$$
\begin{pmatrix} 1 & 0 & 0 \\ 0 & X-1 & X - X^2 \\ 0 & X-1 & X-1 \end{pmatrix}
\xrightarrow{c_3 \mapsto c_3 + X c_2}
\begin{pmatrix} 1 & 0 & 0 \\ 0 & X-1 & 0 \\ 0 & X-1 & X^2-1 \end{pmatrix}
$$

$$
\xrightarrow{r_3 \mapsto r_3 - r_2}
\begin{pmatrix} 1 & 0 & 0 \\ 0 & X-1 & 0 \\ 0 & 0 & X^2-1 \end{pmatrix}.
$$

As above we can identify the matrices $P, Q \in \mathrm{GL}_3(\mathbb{Q}[X])$ such that

$$\begin{pmatrix} 1 & 0 & 0 \\ -X & 1 & 0 \\ X+1 & -1 & 1 \end{pmatrix}\begin{pmatrix} X-1 & 1 & -1 \\ 0 & X & -1 \\ 0 & -1 & X \end{pmatrix}\begin{pmatrix} 0 & 0 & 1 \\ 1 & 1 & 1 \\ 0 & 1 & X \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & X-1 & 0 \\ 0 & 0 & X^2-1 \end{pmatrix}.$$

This form can be used to identify the rational canonical form of A: the invariant polynomials are read off the diagonal as $X-1$ and X^2-1 and A is similar to

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}.$$

References

[::006] Dniester notebook: unsolved problems in the theory of rings and modules. In M. V. Kochetov, V. T. Filippov, V. K. Kharchenko, and I. P. Shestakov, editors, *Non-associative algebra and its applications*, volume 246 of *Lect. Notes Pure Appl. Math.*, pages 461–516. Chapman & Hall/CRC, Boca Raton, FL, 2006. doi:10.1201/9781420003451.axb. Translated from the 1993 Russian edition by Murray R. Bremner and Mikhail V. Kochetov and edited by V. T. Filippov, V. K. Kharchenko and I. P. Shestakov.

[Ber14] D. Berlyne. Ideal theory in rings (Translation of “Idealtheorie in Ringbereichen” by Emmy Noether). 2014, arXiv:1401.2577.

[Bla84] A. Blass. Existence of bases implies the axiom of choice. In *Axiomatic set theory (Boulder, Colo., 1983)*, volume 31 of *Contemp. Math.*, pages 31–33. Amer. Math. Soc., Providence, RI, 1984. doi:10.1090/conm/031/763890.

[Cla10] P. L. Clark. Factorization in integral domains. 2010. URL `http://alpha.math.uga.edu/~pete/factorization2010.pdf`.

[CNT19] C. J. Conidis, P. P. Nielsen, and V. Tombs. Transfinitely valued euclidean domains have arbitrary indecomposable order type. *Communications in Algebra*, 47(3):1105–1113, 2019. doi:10.1080/00927872.2018.1501569.

[Coh00] P. M. Cohn. *Introduction to ring theory.* Springer Undergraduate Mathematics Series. Springer-Verlag London, Ltd., London, 2000. doi:10.1007/978-1-4471-0475-9.

[Ear19] R. Earl. Rings and modules. Lecture notes, Oxford Part A, 2019. URL `https://courses.maths.ox.ac.uk/node/5393/materials`.

[Fuc58] L. Fuchs. *Abelian groups.* Publishing House of the Hungarian Academy of Sciences, Budapest, 1958.

[Gal13] J. A. Gallian. *Contemporary Abstract Algebra.* Cengage Learning, 8th edition, 2013.

[Gel77] S. C. Geller. On the GE_n of a ring. *Illinois J. Math.*, 21(1):109–112, 1977. URL `http://projecteuclid.org/euclid.ijm/1256049506`.

[Gra74] A. Grams. Atomic rings and the ascending chain condition for principal ideals. *Proc. Cambridge Philos. Soc.*, 75:321–329, 1974. doi:10.1017/s0305004100048532.

[Gra08] J. Gray. *Plato's Ghost: The Modernist Transformation of Mathematics.* Princeton University Press, 2008. doi:10.2307/j.ctt7rq1t.

[Hod79] W. Hodges. Krull implies Zorn. *Journal of the London Mathematical Society*, s2-19(2):285–287, 1979. doi:10.1112/jlms/s2-19.2.285.

[Hun80] T. W. Hungerford. *Algebra*, volume 73 of *Graduate Texts in Mathematics.* Springer-Verlag, New York-Berlin, 1980. doi:10.1007/978-1-4612-6101-8. Reprint of the 1974 original.

[Kea98] M. E. Keating. *A First Course in Module Theory.* Imperial College Press, 1998. doi:https://doi.org/10.1142/p082.

[Lam99] T. Y. Lam. *Lectures on modules and rings*, volume 189 of *Graduate Texts in Mathematics.* Springer-Verlag, New York, 1999. doi:10.1007/978-1-4612-0525-8.

[Lam07] T. Y. Lam. *Exercises in modules and rings.* Problem Books in Mathematics. Springer, New York, 2007. doi:10.1007/978-0-387-48899-8.

[Lan02] S. Lang. *Algebra*, volume 211 of *Graduate Texts in Mathematics.* Springer-Verlag, New York, third edition, 2002. doi:10.1007/978-1-4613-0041-0.

[Noe21] E. Noether. Idealtheorie in Ringbereichen. *Math. Ann.*, 83(1-2):24–66, 1921. doi:10.1007/BF01464225.

[Öhm19] L.-D. Öhman. Are induction and well-ordering equivalent? *The Mathematical Intelligencer*, 41(3):33–40, 2019. doi:10.1007/s00283-019-09898-4.

[SCL99] K. Shen, J. N. Crossley, and A. W.-C. Lun. *The nine chapters on the mathematical art.* Oxford University Press, New York; Science Press Beijing, Beijing, 1999. Companion and commentary, With forewords by Wentsün Wu and Ho Peng Yoke.

[She88] K. Shen. The historical development of the Chinese remainder theorem. *J. Hangzhou Univ. Natur. Sci. Ed.*, 15(3):270–282, 1988.

[SW74] T. S. Shores and R. Wiegand. Rings whose finitely generated modules are direct sums of cyclics. *J. Algebra*, 32:152–172, 1974. doi:10.1016/0021-8693(74)90178-1.

[SZ14] L. Salce and P. Zanardo. Products of elementary and idempotent matrices over integral domains. *Linear Algebra Appl.*, 452:130–152, 2014. doi:10.1016/j.laa.2014.03.042.

[Web92] H. Weber. Leopold Kronecker. *Jahresber. Dtsch. Math.-Ver.*, 2:5–31, 1891–92. URL `https://gdz.sub.uni-goettingen.de/id/PPN37721857X_0002`.

[ZD16] K. Zhao and D. E. Dobbs. On the commutative rings with at most two proper subrings. *International Journal of Mathematics and Mathematical Sciences*, 2016:6912360, 2016. doi:10.1155/2016/6912360.

[Zor35] M. Zorn. A remark on method in transfinite algebra. *Bull. Amer. Math. Soc.*, 41(10):667–670, 1935. doi:10.1090/S0002-9904-1935-06166-X.