A3: Rings and Modules, 2022–2023

Tom Sanders

We begin with the course overview as described on `https://courses.maths.ox.ac.uk/course/view.php?id=1042`.

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 characteristic of a ring. Discussion of fields of fractions and their characterisation (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]

Part I

Rings

1 Rings: a recap

A set R containing two (possibly equal) elements 0 and 1, and supporting two binary operations + and × is a **ring** if

- R equipped with + is a commutative group with identity 0;
- × is an associative binary operation on R with identity 1;
- × is distributive over +.

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, $+_R$ or 1_R instead of + and 1 above.

The operation + is the **addition** of the ring, 0 is the **zero** of the ring, and the set R with the operation + is the **additive group** of the ring. For each $x \in R$ we write $-x$ for the unique inverse of x w.r.t. addition, and the map $R \to R; x \mapsto -x$ is the **negation** of the ring; we write $x - y$ for $x + (-y)$.

We call a map $\phi : R \to S$ between rings **additive** if it is a homomorphism of the additive groups.

Observation 1.1. Identities are self-inverse so $-0 = 0$; inverses are symmetric so $-(-x) = x$ for all $x \in R$; and inversion is a homomorphism of the additive group since a group operation is commutative (if and) only if inversion is a homomorphism of the group.

Group homomorphisms map identities to identities and inverses to inverses, so if $\phi : R \to S$ is additive then $\phi(0_R) = 0_S$ and $\phi(-x) = -\phi(x)$ for all $x \in R$.

The operation × is the **multiplication** of the ring, and we write xy in place of $x \times y$, and in the absence of parentheses multiplication precedes addition in the usual way. We say R is a **commutative** ring if the multiplication is commutative.

Remark 1.2. The modern notion of commutative ring can be traced back to Emmy Noether [Noe21, §1] (translated into English in [Ber14]), though unlike us her definition does not assume the multiplication has an identity; Poonen [Poo19] defends our position.

We call a map $\phi : R \to S$ between rings **multiplicative** if $\phi(xy) = \phi(x)\phi(y)$ for all $x, y \in R$, and we call it a **ring homomorphism** if ϕ is additive, multiplicative, and $\phi(1_R) = 1_S$.

Observation 1.3. The composition of additive (resp. multiplicative) maps is additive (resp. multiplicative), and hence the composition of ring homomorphisms is a ring homomorphism.

Definition 1.4. For a set $A \subset X$ and a function $f : X \to Y$ we write $f(A) := \{f(x) : x \in A\}$.

For sets $A \subset X$, $B \subset Y$, and a function $X \times Y \to Z$ denoted by infixing the symbol $*$ between the two arguments, we write $A * B := \{a * b : a \in A, b \in B\}$; and denoted by juxtaposing the two arguments, we write $AB := \{ab : a \in A, b \in B\}$.

For $x \in X$ and $y \in Y$, in the case of infix notation we put $x*A := \{x\}*A$ and $A*y := A*\{y\}$; and in the case of juxtaposition we put $xA := \{x\}A$ and $Ay := A\{y\}$.

Units and the trivial ring

An element $x \in R$ is a **unit** if it is invertible w.r.t. multiplication *i.e.* if there is some $y \in R$ such that $xy = yx = 1$. We write* $U(R)$ for the set of units of R, and R^* for the set of non-zero elements of R. Inverses w.r.t. associative binary operations are unique when they exist, so for $x \in U(R)$ we can unambiguously write x^{-1} for the inverse of x.

Observation 1.5. Identities are self-inverse and so 1 is a unit and $1^{-1} = 1$, and inverses a symmetric so if $x \in U(R)$ then $x^{-1} \in U(R)$ and $(x^{-1})^{-1} = x$.

For $x, y \in U(R)$ we have $(y^{-1}x^{-1})(xy) = 1 = (xy)(y^{-1}x^{-1})$, and so $xy \in U(R)$, and the multiplication on R restricts to a well-defined binary operation on $U(R)$. This operation is *a fortiori* associative; it has an identity – 1; and if $x \in U(R)$ then x^{-1} is an inverse of x w.r.t. this binary operation. In particular, $U(R)$ is a group called the **group of units** with identity the same as the multiplicative identity of R, such that the inverse of x in the group $U(R)$ is the same as the inverse of x as an element of the ring R.

Remark 1.6. 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].

Given $y \in R$, the map $R \to R; x \mapsto yx$ (resp. $R \to R; x \mapsto xy$) is called **left** (resp. **right**) **multiplication by** y.

Observation 1.7. The fact that multiplication is distributive over addition in R is exactly to say that all the left and right multiplication maps are group homomorphisms of the additive group of R.

Group homomorphisms map identities to identities and inverses to inverses, so $x0 = 0x = 0$ for all $x \in R$ – we say **zero annihilates**; and $x(-y) = -(xy) = (-x)y$ for all $x, y \in R$ – we say that **negation distributes**. In particular $(-1)x = -x$ for all $x \in R$.

Example 1.8. The set $\{0\}$, with $1 = 0$, and addition and multiplication given by $0 + 0 = 0 \times 0 = 0$, is a ring called the **trivial** or **zero** ring. A ring in which $1 \neq 0$ is called a **non-trivial** ring.

If R is not non-trivial then it is trivial: Indeed, since $0 = 1$, for all $x \in R$ we have $x = 1x = 0x = 0$ since zero annihilates and so $R = \{0\}$. There is only one function into a set of size one, and so the addition and multiplication on R are uniquely determined and must be that of the trivial ring.

* ⚠ Some authors (*e.g.* [Lan02, p84] and [Lam07, xiv]) write R^* for the group of units of R.

Example 1.9. The **zero map** $z_R : R \to \{0\}; x \mapsto 0$ from a ring R to the trivial ring is a ring homomorphism.

The integers and characteristic

We write $\mathbb{Z}$ for the integers; $\mathbb{N}^*$ for the positive integers, that is $\{1, 2, 3, \dots\}$; and $\mathbb{N}_0$ for the non-negative integers, that is $\{0, 1, 2, \dots\}$.

Example 1.10. $\mathbb{Z}$ with their usual addition, multiplication, zero and 1 form a non-trivial commutative ring for which $U(\mathbb{Z}) = \{-1, 1\}$.

Theorem 1.11 (The One Ring)**.** [†] *Suppose that R is a ring. Then there is a unique ring homomorphism $\chi_R : \mathbb{Z} \to R$, and we have*

$$\chi_R(n-m) = \overbrace{(1_R + \cdots + 1_R)}^{n \text{ times}} - \overbrace{(1_R + \cdots + 1_R)}^{m \text{ times}}$$

Remark 1.12. The proof is omitted. It is a series of inductions and to do it properly we would need to be careful about what our definitions of $\mathbb{N}^*$ and $\mathbb{Z}$ are.

If there is $n \in \mathbb{N}^*$ such that $\chi_R(n) = 0_R$ then there is a smallest such n and we call this the **characteristic** of the ring; if there is no such n then the characteristic is said to be 0.

Example 1.13. For $N \in \mathbb{N}^*$, we write $\mathbb{Z}_N$ for the integers modulo N. This is a commutative ring whose zero is $0 \pmod N$, and with multiplicative identity $1 \pmod N$. If $N = 1$ then $0 \equiv 1 \pmod N$ and so the ring is trivial; otherwise it is non-trivial.

The characteristic of $\mathbb{Z}_N$ is N since $n \in \mathbb{N}^*$ has $\chi_{\mathbb{Z}_N}(n) = 0_{\mathbb{Z}_N}$ if and only if $n \equiv 0 \pmod N$, so $n \geqslant N$ and $\chi_{\mathbb{Z}_N}(N) = 0_{\mathbb{Z}_N}$.

Isomorphisms and subrings

A **ring isomorphism** is a map $\phi : R \to S$ that is a ring homomorphism with an inverse that is a ring homomorphism.

Example 1.14. The identity map $\iota_R : R \to R; x \mapsto x$ is a ring isomorphism.

A ring S is a **subring** of a ring R if the inclusion map $j : S \to R; s \mapsto s$ is a well-defined – all this does is ensure that $S \subset R$ – ring homomorphism called the **inclusion homomorphism**; S is **proper** if $S \neq R$.

Example 1.15. $\mathbb{C}$ with its usual addition, multiplication, zero and 1 is a non-trivial commutative ring and $\mathbb{Z}$ is a proper subring of $\mathbb{C}$.

[†]Following [Tol04, Book I, Chapter 2, p66] one might describe the integes as the one ring (up to unique isomorphism) ruling (uniquely embedding in) all others.

Observation 1.16. The 0 and 1 of a subring are the same as for the containing ring and so a subring of a non-trivial ring is non-trivial, and the characteristic of a subring is the same as the characteristic of the ring it is contained in.

⚠ In particular, the trivial ring is *not* a subring of any non-trivial ring R despite the fact that the inclusion map taking 0 to 0_R is both additive and multiplicative. It follows that the requirement that ring homomorphisms send 1 to 1 cannot be dropped from the definition.

Proposition 1.17 (Subring test)**.** *Suppose that R is a ring and $S \subset R$ has $1 \in S$ and $x - y, xy \in S$ for all $x, y \in S$. 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 $x-y \in S$ whenever $x, y \in S$ 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, multiplication on R restricts to a well-defined binary operation on S that is *a fortiori* associative and distributive, and since $1 \in S$ and 1 is *a fortiori* an identity for S, we have that S with these restricted operations is a ring. The map $S \to R; s \mapsto s$ is then well-defined since S is a subset of R, and a ring homomorphism as required. □

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 restricted operations described by the lemma.

Example 1.18. For $d \in \mathbb{N}^*$ we write $\mathbb{Z}[\sqrt{-d}]$ for the set $\{z + w\sqrt{-d} : z, w \in \mathbb{Z}\}$, which is a subring of $\mathbb{C}$ by the subring test. $\mathbb{Z}[i]$ – the case $d = -1$ – is called the set of **Gaussian integers**.

We have $U(\mathbb{Z}[i]) = \{1, -1, i, -i\}$: Certainly all the elements of $\{1, -1, i, -i\}$ are units. In the other direction, suppose $(z+wi)(x+yi) = 1$ for some $x, y \in \mathbb{Z}$. Taking absolute values we have $(z^2 + w^2)(x^2 + y^2) = 1$, so $z^2 + w^2 = 1$, and hence $(z, w) \in \{(1, 0), (-1, 0), (0, 1), (0, -1)\}$ as required.

For $d > 1$ we have $U(\mathbb{Z}[\sqrt{-d}]) = \{-1, 1\}$ since certainly 1 and -1 are units, and if $z+w\sqrt{-d}$ is a unit then taking absolute values as above we get $x, y \in \mathbb{Z}$ such that $(z^2+dw^2)(x^2+dy^2) = 1$; since $d > 1$ we get $w = 0$ and $z \in \{-1, 1\}$ as required.

Example 1.19. Given a ring R we write $Z(R)$ for the **centre** of R, that is the set of $x \in R$ that commute with all other elements of R *i.e.* such that $xy = yx$ for all $y \in R$.

The centre is a subring by the subring test: $1x = x = x1$ for all $x \in R$, so $1 \in Z(R)$. Secondly, for $x, x' \in Z(R)$, and $y \in R$ we have $(x - x')y = xy + (-x')y = xy + x'(-y) = yx + (-y)x' = yx + y(-x') = y(x - x')$ and $(xx')y = x(x'y) = x(yx') = (xy)x' = (yx)x' = y(xx')$.

Example 1.20. The ring of integers has no proper subrings, since any such subring must contain 1 and so by induction $\mathbb{N}^*$ and hence $\mathbb{N}^* - \mathbb{N}^* = \mathbb{Z}$.

⚠ The set $\mathbb{N}^*$ contains 1 and if $x, y \in \mathbb{N}^*$ then $x + y, xy \in \mathbb{N}^*$, but $\mathbb{N}^*$ is *not* a subring of $\mathbb{Z}$ because it does not contain 0. It follows that $x - y$ may not be replaced by $x + y$ in the hypotheses of the subring test.

Observation 1.21. For $\phi : R \to S$ a ring homomorphism, $\operatorname{Im}\phi$ is a subring of S by the subring test: $1_S = \phi(1_R) \in \operatorname{Im}\phi$; and if $x, y \in \operatorname{Im}\phi$ then there are $z, w \in R$ such that $x = \phi(z)$ and $y = \phi(w)$ so $xy = \phi(zw) \in \operatorname{Im}\phi$ and $x - y = \phi(x) - \phi(y) = \phi(x - y) \in \operatorname{Im}\phi$.

Fields

We say that a commutative ring R is a **field** if $U(R) = R^*$. A subring that is also a field is called a **subfield**. Throughout these notes $\mathbb{F}$ always denotes a field.

Example 1.22. The complex numbers $\mathbb{C}$ are a field with $\mathbb{R}$ as a subfield.

Proposition 1.23. *Suppose that $\phi : \mathbb{F} \to R$ is a ring homomorphism and R is non-trivial. Then ϕ is an injection and $\operatorname{Im}\phi$ is a subfield of R.*

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.

The image of ϕ is a subring of R which is non-trivial since R is non-trivial, and it is commutative since $\phi(x)\phi(y) = \phi(xy) = \phi(yx) = \phi(y)\phi(x)$. If $\phi(y) \neq 0$ then since ϕ is an injection, $y \neq 0$ and so y^{-1} exists and $\phi(y)\phi(y^{-1}) = \phi(1) = 1$, whence $\phi(y)$ is a unit in $\operatorname{Im}\phi$, and $\operatorname{Im}\phi$ is a subfield. □

Proposition 1.24. *Suppose that $\phi : \mathbb{F} \to R$ is a ring homomorphism. Then the map $\mathbb{F} \times R \to R; (\lambda, r) \mapsto \lambda.r := \phi(\lambda)r$ is a scalar multiplication of the field $\mathbb{F}$ on the additive group of R giving an $\mathbb{F}$-vector space such that the right multiplication maps on R are linear, and if ϕ maps $\mathbb{F}$ into the centre of R then so are the left multiplication maps. In particular if R is commutative then the left and right multiplication maps are linear.*

Conversely, if R is a ring which is also an $\mathbb{F}$-vector space in such a way that all the right multiplication maps are linear then the map $\mathbb{F} \to R; \lambda \mapsto \lambda.1_R$ is a ring homomorphism and if all the left multiplication maps are also linear then its image is in the centre of R.

Proof. The additive group of R is a commutative group by definition. We also have $(\lambda\mu).v = \phi(\lambda\mu)v = (\phi(\lambda)\phi(\mu))v = \phi(\lambda)(\phi(\mu)v) = \lambda.(\mu.v)$; $1_{\mathbb{F}}.v = \phi(1_{\mathbb{F}})v = 1_R v = v$; $(\lambda + \mu).v = \phi(\lambda + \mu)v = (\phi(\lambda) + \phi(\mu))v = \phi(\lambda)v + \phi(\mu)v = \lambda.v + \mu.v$; and $\lambda.(v + w) = \phi(\lambda)(v + w) = \phi(\lambda)v + \phi(\lambda)w = \lambda.v + \lambda.w$. It follows R is an $\mathbb{F}$-vector space as claimed.

Right multiplication by r is linear since it is a group homomorphism and $\lambda.(vr) = \phi(\lambda)(vr) = (\phi(\lambda)v)r = (\lambda.v)r$. Finally, left multiplication by r is a group homomorphism, and if it commutes with all elements of the image of ϕ (which is certainly true if ϕ maps to the centre of R), then $\lambda.(rv) = \phi(\lambda)(rv) = (\phi(\lambda)r)v = (r\phi(\lambda))v = r(\phi(\lambda)v) = r(\lambda.v)$, and so left multiplication by r is linear.

Conversely, write ϕ for the given map then $\phi(1_{\mathbb{F}}) = 1_{\mathbb{F}}.1_R = 1_R$; $\phi(x+y) = (x+y).1_R = x.1_R + y.1_R = \phi(x) + \phi(y)$; and $\phi(xy) = (xy).1_R = x.(y.1_R) = x.\phi(y) = x.(1_R\phi(y)) = (x.1_R)\phi(y) = \phi(x)\phi(y)$ since the map $R \to R; z \mapsto z\phi(y)$ is linear. It follows that ϕ is a ring homomorphism as claimed. If all left multiplication maps are linear then for $r \in R$ we have $r\phi(\lambda) = r(\lambda.1) = \lambda.(r1_R) = \lambda.(1_R r) = (\lambda.1)r = \phi(\lambda)r$ and so $\phi(\lambda) \in Z(R)$. □

We call the vector space structure of the proposition the **$\mathbb{F}$-(vector) space structure on R induced by ϕ**.

Example 1.25. The inclusion map $\mathbb{R} \to \mathbb{C}$ induces the usual $\mathbb{R}$-vector space structure on the additive group of $\mathbb{C}$. $\{1, i\}$ is a basis for this vector space, which is another way of saying that every complex number can be written uniquely in the form $a + bi$ for reals a and b.

Example 1.26. Complex conjugation, $\mathbb{C} \to \mathbb{C}; z \mapsto \overline{z}$ is a ring homomorphism that is different from the identity. The identity map induces a $\mathbb{C}$-vector space structure with $\lambda.z := \lambda z$, and complex conjugation a different structure with $\lambda.z := \lambda z$.

Zero divisors and integral domains

In a ring R we call an element $y \in R$ a **left** (resp. **right**) **zero-divisor** if the left (resp. right) multiplication-by-y map has a non-trivial kernel *i.e.* if there is some $x \neq 0$ such that $yx = 0$ (resp. $xy = 0$). A non-trivial commutative ring R is an **integral domain** if it has no non-zero zero-divisors.

Example 1.27. $\mathbb{Z}$ is an integral domain – it is our prototypical example.

Observation 1.28. If $x \in U(R)$ then x is not a left (resp. right) zero-divisor since if $xy = 0$ (resp. $yx = 0$) then $0 = x^{-1}0 = x^{-1}(xy) = 1y = y$ (resp. $0 = 0x^{-1} = (yx)x^{-1} = y1 = y$).

Example 1.29. Every field $\mathbb{F}$ is an integral domain since it is certainly a non-trivial commutative ring and every non-zero element is a unit and so not a zero-divisor.

Example 1.30 (Example 1.13, contd.)**.** By Bezout's Lemma if $\gcd(a, N) = 1$ then there are $\alpha, \beta \in \mathbb{Z}$ such that $\alpha a + \beta N = 1$ and so $\alpha a \equiv 1 \pmod{N}$. Since $\mathbb{Z}_N$ is commutative it follows that $a\alpha \equiv 1 \pmod{N}$ and so a is a unit in $\mathbb{Z}_N$. On the other hand, if $\gcd(a, N) > 1$ then $a(N/\gcd(a, N)) \equiv 0 \pmod{N}$ and $N/\gcd(a, N) \not\equiv 0 \pmod{N}$, so a is a zero-divisor and hence not a unit. It follows that $U(\mathbb{Z}_N) = \{a \pmod{N} : \gcd(a, N) = 1\}$.

If $p > 1$ is prime then for all a, either $p \mid a$ or $\gcd(a, p) = 1$. Hence $U(\mathbb{Z}_p) = \mathbb{Z}_p^*$ and so $\mathbb{Z}_p$ is a field; we denote it $\mathbb{F}_p$ to emphasise this fact.

If $N > 1$ is composite, say $N = ab$ for $a, b > 1$ then $ab \equiv 0 \pmod{N}$ but $a, b \not\equiv 0 \pmod{N}$ and so $\mathbb{Z}_N$ is not even an integral domain.

If $N = 1$ then $\mathbb{Z}_N$ is trivial, and so not even non-trivial!

Observation 1.31. If R is an integral domain and S is a subring of R then S is an integral domain: S is certainly non-trivial and commutative since R is, and for $y \in S$, the multiplication-by-y map on S is the restriction of the multiplication-by-y map on R, and so if the kernel of the latter is trivial then so is the kernel of the former.

Example 1.32. For $d \in \mathbb{N}^*$, the ring $\mathbb{Z}[\sqrt{-d}]$, and in particular the Gaussian integers, is a subring of $\mathbb{C}$ and so an integral domain.

Example 1.33. The algebraic integers, denoted $\overline{\mathbb{Z}}$, are the complex numbers α for which there is $d \in \mathbb{N}^*$ and $a_{d-1}, \ldots, a_0 \in \mathbb{Z}$ such that $\alpha^d + a_{d-1}\alpha^{d-1} + \cdots + a_1\alpha + a_0 = 0$. We shall make use of arguments from the modules part of the course to show that $\overline{\mathbb{Z}}$ is a subring of $\mathbb{C}$, and given this we conclude $\overline{\mathbb{Z}}$ is an integral domain.

$\overline{\mathbb{Z}}$ is not a field since $1/2 \notin \overline{\mathbb{Z}}$, because if it were then there would be $a_{d-1}, \ldots, a_0 \in \mathbb{Z}$ such that $1 + 2(a_{d-1} + \cdots + a_0 2^{d-1}) = 0$; a contradiction.

Proposition 1.34. *Suppose that R is a ring with no non-zero zero divisors that is also a finite dimensional vector space in such a way that left and right multiplication maps are linear. Then $U(R) = R^*$, and in particular if R is an integral domain then R is a field.*

Proof. For $a \in R$ the map $R \to R; x \mapsto xa$ is linear, and since R is an integral domain it is injective if $a \in R^*$. Since R is finite dimensional the Rank-Nullity theorem tells us that the map is surjective, and hence there is $x \in R$ such that $xa = 1$. Similarly there is y such that $ay = 1$, and finally $x = x1 = x(ay) = (xa)y = 1y = y$ so $a \in U(R)$ as required. □

Products of rings

For rings $R_1, \ldots, R_n$ the product group $R_1 \times \cdots \times R_n$ of the additive groups of the rings R_i may be equipped with a binary operation defined by $(xy)_i := x_i y_i$ for $1 \leqslant i \leqslant n$ making it into a ring with identity $1 = (1_{R_1}, \ldots, 1_{R_n})$. This ring is called the **direct product** of the R_is.

Observation 1.35. The group of units of a product ring is equal to the product group of the groups of units of the rings *i.e.* $U(R_1 \times \cdots \times R_n) = U(R_1) \times \cdots \times U(R_n)$.

Example 1.36. The maps $\pi_i : R_1 \times \cdots \times R_n \to R_i; x \mapsto x_i$ are ring homomorphisms called **projection homomorphisms**.

Example 1.37. The map $R \to R^n; x \mapsto (x, \dots, x)$ is a ring homomorphism called the **diagonal homomorphism (into R^n)**.

The diagonal homomorphism $\mathbb{F} \to \mathbb{F}^n$ induces an $\mathbb{F}$-vector space structure on $\mathbb{F}^n$ which is the usual $\mathbb{F}$-vector space structure on $\mathbb{F}^n$ *i.e.* having scalar multiplication $\lambda.v = (\lambda v_1, \dots, \lambda v_n)$ for $\lambda \in \mathbb{F}$ and $v \in \mathbb{F}^n$. ⚠ The ring $\mathbb{F}^n$ has more structure than the vector space $\mathbb{F}^n$ because the former comes with a multiplication.

Example 1.38. For R a ring, R^2 is never an integral domain: if R is trivial then $1_{R^2} = (1_R, 1_R) = (0_R, 0_R) = 0_{R^2}$, so R^2 is trivial. Otherwise $(0_R, 1_R)(1_R, 0_R) = (0_R, 0_R) = 0_{R^2}$ $(0_R, 1_R), (1_R, 0_R) \in (R^2)^*$ and so these are non-zero zero-divisors.

Prototypical rings

Groups of symmetries are the prototypes for abstract groups and rings have a similar prototype in which the underlying set is replaced by a commutative group.

Proposition 1.39. *Suppose that M and N are commutative groups with binary operations $+_M$ and $+_N$, and identities 0_M and 0_N respectively. Then* $\mathrm{Hom}(M, N)$*, the set of group homomorphisms $M \to N$, is itself a commutative group under $+$ defined pointwise on* $\mathrm{Hom}(M, N)$ *by*

$$(\phi + \psi)(x) := \phi(x) +_N \psi(x) \text{ for all } x \in M,$$

with identity $z : M \to N; x \mapsto 0_N$, and the inverse of ϕ is the pointwise negation, meaning for all $x \in M$, $(-\phi)(x)$ is the inverse of $\phi(x)$ in N.

Proof. Suppose that $\phi, \psi \in \mathrm{Hom}(M, N)$. Then for all $x, y \in M$ we have

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

It follows that $\phi + \psi \in \mathrm{Hom}(M, N)$. Pointwise addition is commutative and associative because addition on N is commutative and associative. The map z is a homomorphism because $z(x) +_N z(y) = 0_N +_N 0_N = 0_N = z(x +_M y)$, and it is an identity for pointwise addition because 0_N is an identity for N. Finally, if $\phi \in \mathrm{Hom}(M, N)$ then $-\phi \in \mathrm{Hom}(M, N)$ because it is the composition of the homomorphism ϕ and negation which is a homomorphism on N since $+_N$ is commutative, and this map is an inverse for $\phi(x)$ under pointwise addition by design. □

Remark 1.40. To show that $\mathrm{Hom}(M, N)$ is a closed under pointwise addition and negation it is essential that N be commutative.

Proposition 1.41. *Suppose that M, N, and P are commutative groups, and $+_N$ and $+_P$ are the group operations on N and $\mathrm{Hom}(M,N)$, and P and $\mathrm{Hom}(N,P)$ respectively. If $\phi \in \mathrm{Hom}(M,N)$ and $\psi \in \mathrm{Hom}(N,P)$, then $\psi \circ \phi \in \mathrm{Hom}(M,P)$; if $\pi \in \mathrm{Hom}(M,N)$ then $\psi \circ (\phi +_N \pi) = (\psi \circ \phi) +_P (\psi \circ \pi)$; and if $\pi \in \mathrm{Hom}(N,P)$ then $(\psi +_P \pi) \circ \phi = (\psi \circ \phi) +_P (\pi \circ \phi)$.*

Proof. The composition of homomorphisms is a homomorphism which 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

$$\psi \circ (\phi +_N \pi)(x) = \psi(\phi(x) +_N \pi(x)) = \psi(\phi(x)) +_P \psi(\pi(x)) = ((\psi \circ \phi) +_P (\psi \circ \pi))(x)$$

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

$$(\psi +_P \pi) \circ \phi(x) = \psi(\phi(x)) +_P \pi(\phi(x)) = ((\psi \circ \phi) +_P (\pi \circ \phi))(x)$$

by definition. The result is proved. □

Remark 1.42. For the identity $\psi \circ (\phi +_N \pi) = (\psi \circ \phi) +_P (\psi \circ \pi)$ we used the homomorphism property of ψ, while the identity $(\psi +_P \pi) \circ \phi = (\psi \circ \phi) +_P (\pi \circ \phi)$ followed simply from the definition; *c.f.* Exercise I.1.

Theorem 1.43. *Suppose that M is a commutative group. Then the set $\mathrm{Hom}(M,M)$ equipped with pointwise addition as its addition and functional composition as its multiplication is a ring whose zero is the map $M \to M; x \mapsto 0_M$ and whose multiplicative identity is the map $M \to M; x \mapsto x$.*

Proof. $\mathrm{Hom}(M,M)$ is a commutative group with the given identity under this addition, and by the second part the proposed multiplication distributes over this addition. It remains to note 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. □

Matrix rings

Given a ring R, we write $M_{n,m}(R)$ for the set of $n \times m$ matrices with entries in R, and $M_n(R) \coloneqq M_{n,n}(R)$. For $A, A' \in M_{n,m}(R)$ and $B \in M_{m,p}(R)$ we define matrices $A + A' \in M_{n,m}(R)$ and $AB \in M_{n,p}(R)$ by

$$(A + A')_{i,j} \coloneqq A_{i,j} + A'_{i,j} \text{ and } (AB)_{i,k} \coloneqq \sum_{j=1}^{m} A_{i,j} B_{j,k}. \tag{1.1}$$

We write $0_{n\times m}$ for the matrix in $M_{n,m}(R)$ with 0_R in every entry, and I_n for the $n \times n$ matrix with 1_Rs on the diagonal and 0_Rs elsewhere.

Proposition 1.44 (Algebra of matrices)**.** *Suppose that R is a ring. Then $M_{n,m}(R)$ is a commutative group under $+$ with identity $0_{n\times m}$ and for which the inverse of $A \in M_{n,m}(R)$ is the matrix $-A$ with $(-A)_{i,j} = -A_{i,j}$. Furthermore, if $A \in M_{n,m}(R)$, $B, B' \in M_{m,l}(R)$, and $C, C' \in M_{p,n}(R)$ then $C(AB) = (CA)B$, $A(B + B') = (AB) + (AB')$, $(C + C')A = (CA) + (C'A)$, $AI_m = A$ and $I_nA = A$.*

Remark 1.45. We omit the proof. One can proceed directly using a change of variables and distributivity, or using some of the language of modules.

The commutative group $M_n(R)$ with multiplication $M_n(R)\times M_n(R) \to M_n(R); (A, B) \mapsto AB$ is a ring with multiplicative identity I_n as a result of the algebra of matrices. A **matrix ring** is a subring of $M_n(R)$.

Remark 1.46. For $A \in M_n(R)$ the **determinant** of A is defined to be

$$\det A := \sum_{\sigma\in S_n} \operatorname{sign}(\sigma) A_{1,\sigma(1)}\cdots A_{n,\sigma(n)},$$

where S_n is the symmetry group of permutations of $\{1, \ldots, n\}$, and $\operatorname{sign}(\sigma)$ is the sign of the permutation σ.

We shall see in the second half of the course that for R commutative, $A \in U(M_n(R))$ if and only if $\det A \in U(R)$, generalising what we already know for fields since $\det A \in U(\mathbb{F})$ if and only if $\det A \neq 0_{\mathbb{F}}$. For non-commutative rings Exercise I.5 gives an example to show that this equivalence can fail.

Example 1.47. For R non-trivial the ring $M_2(R)$ is not commutative:

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

Example 1.48. Given a ring R, the map

$$\Delta : R \to M_n(R); \lambda \mapsto \begin{pmatrix} \lambda & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \lambda \end{pmatrix}$$

is a ring homomorphism called the **diagonal homomorphism (into $M_n(R)$)**.

The diagonal homomorphism into $M_n(\mathbb{F})$ induces the usual $\mathbb{F}$-vector space structure on $M_n(\mathbb{F})$ with scalar multiplication $(\lambda.A)_{i,j} = \lambda A_{i,j}$. Writing $E^{(i,j)}$ for the matrix with $E^{(i,j)}_{i,j} = 1$ and $E^{(i,j)}_{k,l} = 0$ for $(k, l) \neq (i, j)$, the set $\{E^{(i,j)} : 1 \leqslant i, j \leqslant n\}$ is a basis for this vector space.

Example 1.49. The **quaternions** are the set

$$\mathbb{H} := \left\{ \begin{pmatrix} z & w \\ -\overline{w} & \overline{z} \end{pmatrix} : z, w \in \mathbb{C} \right\}.$$

They form a subring of $M_2(\mathbb{C})$ by the subring test, and in particular $\mathbb{H}$ has zero $0_{2\times 2}$ and multiplicative identity I_2. Now,

$$A := \begin{pmatrix} z & w \\ -\overline{w} & \overline{z} \end{pmatrix} \neq 0_{2\times 2} \text{ if and only if } \det A = |z|^2 + |w|^2 \neq 0,$$

and hence if $A \in \mathbb{H}^*$ then the inverse of A in $M_2(\mathbb{C})$ exists and it is also in $\mathbb{H}$. Hence $A \in U(\mathbb{H})$ and since $\mathbb{H}$ is non-trivial, $U(\mathbb{H}) = \mathbb{H}^*$. The quaternions are not, however, commutative and so this is not a field. A not-necessarily commutative ring in which $U(R) = R^*$ is called a **division ring** or **skew field**.

Frobenius showed that any real division ring that is also a vector space over $\mathbb{R}$ in such a way that left and right multiplication is linear, is isomorphic (via a map that is both a ring isomorphism and a linear isomorphism) to either $\mathbb{R}$, $\mathbb{C}$, or $\mathbb{H}$.

The ring homomorphism

$$\mathbb{R} \to \mathbb{H}; \lambda \mapsto \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix}$$

has image equal to the centre of $\mathbb{H}$, and so induces a real vector space structure on $\mathbb{H}$ in which left and right multiplication maps. The vector space if 4-dimensional and

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \text{ and } \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix}$$

form a basis. As element of the group $U(\mathbb{H})$, these generate an 8 element subgroup called the **quaternion group** and denoted Q_8.

There is another natural ring homomorphism: the map

$$\mathbb{C} \to \mathbb{H}; \lambda \mapsto \begin{pmatrix} \lambda & 0 \\ 0 & \overline{\lambda} \end{pmatrix},$$

which induces a 2-dimensional $\mathbb{C}$-vector space structure on $\mathbb{H}$ in which right multiplication maps are linear, but left multiplication maps are not (in general).

In fact there is no $\mathbb{C}$-vector space structure on $\mathbb{H}$ such that all left and right multiplication maps are linear: If there were it would give rise to a ring homomorphism $\mathbb{C} \to \mathbb{H}$ mapping into the centre of $\mathbb{H}$. The centre of $\mathbb{H}$ is isomorphic to $\mathbb{R}$, and hence we would have a ring homomorphism $\mathbb{C} \to \mathbb{R}$ which we see in Exercise I.3 is not possible. ⚠ In particular, $\mathbb{H}$ is *not* a subspace of the usual $\mathbb{C}$-vector space $M_2(\mathbb{C})$ as defined in Example 1.48 because in that structure the left and right multiplication maps *are* linear, and since $\mathbb{H}$ is a subring if it were also subspace they would restrict to be linear on $\mathbb{H}$.

Polynomial rings

Proposition 1.50 (Algebra of polynomials)**.** *Suppose that R is a subring of S, $\lambda \in S$ commutes with all elements of R, and $a_0, a_1, \ldots, b_0, b_1, \cdots \in R$ have $a_i = 0$ for all $i > n$ and $b_j = 0$ for all $j > m$. Then*

$$\left(\sum_{i=0}^{n} a_i\lambda^i\right) + \left(\sum_{j=0}^{m} b_j\lambda^j\right) = \sum_{i=0}^{\max\{n,m\}} (a_i + b_i)\lambda^i \text{ and } -\left(\sum_{i=0}^{n} a_i\lambda^i\right) = \sum_{i=0}^{n} (-a_i)\lambda^i,$$

and

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

Remark 1.51. We omit the proof though it is not difficult: it makes essential use of distributivity and changes of variables.

For a non-trivial ring R there is a non-trivial ring $R[X]$ called the **polynomial ring over R with variable X** with R as a subring, and a distinguished element $X \in R[X]$ which commutes with all elements of $R[X]$, *i.e.* $pX = Xp$ for all $p \in R[X]$, such that

$$R[X] = \{a_0 + a_1X + \cdots + a_nX^n : n \in \mathbb{N}_0, a_0, \ldots, a_n \in R\}, \tag{1.2}$$

and

$$a_0 + a_1X + \cdots + a_nX^n = 0_R \Rightarrow a_0, \ldots, a_n = 0_R. \tag{1.3}$$

Remark 1.52. We omit the proof that such a ring exists, but the idea is to take the additive group of functions $\mathbb{N}_0 \to R$ with a finite number of non-zero entries and group operation coordinate-wise addition, and identify X^n with the function taking m to 0_R if $m \neq n$ and 1_R if $m = n$.

For more variables we define $R[X_1, \ldots, X_n] \coloneqq R[X_1, \ldots, X_{n-1}][X_n]$ and call it the **polynomial ring over R in the variables $X_1, \ldots, X_n$**.

The algebra of polynomials and (1.3) allows the **equating of coefficients**, meaning that if $a_0 + a_1X + \cdots + a_nX^n = b_0 + b_1X + \cdots + b_mX^m$ for $a_0, a_1, \ldots, b_0, b_1, \cdots \in R$ with $a_i = 0$ for $i > n$ and $b_j = 0$ for $j > m$, then $a_i = b_i$ for all i.

If $p \in R[X]^*$ then there is a minimal $d \in \mathbb{N}_0$ and unique elements $a_0, a_1, \ldots, a_d \in R$ with $a_d \neq 0_R$ such that $p(X) = a_0 + a_1X + \cdots + a_dX^d$. We call this minimal d the **degree** of p and denote it $\deg p$; we call a_i the **coefficient** of X^i; a_d the **lead coefficient** and a_0 the **constant coefficient**.

A polynomial is **monic** if its lead coefficient is 1, and the **constant polynomials** are those for which the constant coefficient is the only coefficient that may be non-zero.

Example 1.53. The inclusion homomorphism $\mathbb{F} \to \mathbb{F}[X]$ induces an $\mathbb{F}$-vector space structure on $\mathbb{F}[X]$ in such a way that all multiplication maps are linear. In this space, (1.2) says exactly that $\{1, X, X^2, \ldots\}$ is a spanning set, while (1.3) tells us it is linearly independent.

Proposition 1.54. *Suppose that $\phi : R \to S$ is a ring homomorphism from a non-trivial ring, and $\lambda \in S$ commutes with all elements of the image of ϕ. Then there is a unique ring homomorphism $R[X] \to S$ extending ϕ and mapping X to λ, and we have*

$$a_0 + a_1 X + \cdots + a_d X^d \mapsto \phi(a_0) + \phi(a_1)\lambda + \cdots + \phi(a_d)\lambda^d.$$

Proof. The proposed map is well-defined because we can equate coefficients. It extends ϕ since the constant polynomial r is mapped to $\phi(r)$, and it certainly maps X to λ. Finally, it is additive and multiplicative by the algebra of polynomials, and certainly maps 1 to 1 since it extends ϕ, and ϕ maps 1 to 1. It follows that the given map is a ring homomorphism.

Any other ring homomorphism ψ with $\psi(r) = r$ for all $r \in R$, and $\psi(X) = \lambda$ must agree with the given map on $R[X]$ by the homomorphism property of ψ, and hence uniqueness follows. □

We call the homomorphism of this proposition the **evaluation homomorphism at λ extending ϕ** and write $p(\lambda)$ for the image of p under this map. ⚠ The notation $p(\lambda)$ does not make explicit reference to ϕ.

For R a subring of S and $\lambda \in S$ commuting with all elements of R, the image of the evaluation homomorphism at λ extending the inclusion homomorphism $R \to S$ is denoted $R[\lambda]$ and is a subring of S.

Remark 1.55. This proposition for polynomial rings should be compared with Theorem 1.11 for the integers.

We say that α is a **root** of p if $p(\alpha) = 0$.

Theorem 1.56 (Factor theorem)**.** *Suppose R is a non-trivial ring and α is a root of p. Then there is $q \in R[X]$ such that $p(X) = q(X)(X - \alpha)$.*

Proof. Write $p(X) = a_0 + a_1 X + \cdots + a_n X^n$ and note that

$$p(X) = p(X) - p(\alpha) = \sum_{i=0}^{n} a_i (X^i - \alpha^i) = \left(\sum_{i=0}^{n} a_i (X^{i-1} + X^{i-2}\alpha + \cdots + X\alpha^{i-2} + \alpha^{i-1}) \right) (X - \alpha).$$

□

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

Proposition 1.57. *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 $R[X]$ is a non-trivial commutative ring, and so the fact it is an integral domain follows by forgetting the degree equation in (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 by the algebra of polynomials we see that $\deg pq \leqslant n + m$ and the coefficient of X^{n+m} is $a_n b_m$. The coefficient of X^{n+m} is non-zero since R is an integral domain and $a_n, b_m \in R^*$. We conclude that $pq \in R[X]^*$ and $\deg pq = n + m = \deg p + \deg q$ as required. □

Example 1.58. $\mathbb{Z}[X]$ is an integral domain since $\mathbb{Z}$ is an integral domain.

Example 1.59. $\mathbb{F}[X_1, \ldots, X_n]$ is an integral domain by induction on n: for the base case every field is an integral domain, and for the inductive step we have Proposition 1.57.

Example 1.60. When R is an integral domain we have $U(R[X]) = U(R)$. To see this, suppose that $p \in U(R[X])$. Then there is some $q \in U(R[X])$ such that $pq = 1$, and so $0 = \deg p + \deg q$, whence $\deg p = 0$ and $\deg q = 0$. Thus $p(X) = a_0$ and $q(X) = b_0$ for some $a_0, b_0 \in R^*$. Since $a_0 b_0 = 1$ and R is commutative we have $b_0 a_0 = a_0 b_0 = 1$, so $p(X) = a_0 \in U(R)$ as required. Conversely, if $p \in U(R)$ then $p \in U(R[X])$ and we are done.

2 Ideals and quotients

Subrings are an important substructure of rings, but just as groups have subgroups and normal subgroups, rings have subrings and a further type of structure called an ideal. Normal subgroups are connected to quotient groups, and ideals are connected to quotient rings in the same way.

Given an ring R, a **left (resp. right) ideal** in R is a subgroup I of the additive group of R that is closed under multiplication on the left (resp. right) by all elements of R *i.e.* I is a subgroup with $rx \in I$ (resp. $xr \in I$) for all $r \in R$ and $x \in I$. An **ideal** in R – also called a **two-sided ideal** – is a left ideal *and* right ideal.

Remark 2.1. Left and right ideals are connected with the module structure of rings which we will examine more closely in the second part of the course. For now, two-sided ideals are our focus.

Observation 2.2. If R is commutative then every left ideal (resp. right) ideal is a (two-sided) ideal and hence a right (resp. left) ideal.

Example 2.3. In any ring R the sets $\{0\}$ and R are ideals called the **zero ideal** and **unit ideal** respectively.

Observation 2.4. If I is a left (resp. right) ideal containing a unit x then for all $r \in R$, $rx^{-1}x \in I$ (resp. $xx^{-1}r \in I$) so $I = R$. In particular, any left, right, or two-sided ideal containing a unit is the unit ideal.

Example 2.5. Every non-zero element of a field is a unit, and so any non-zero ideal is the unit ideal. In other words, fields have only two ideals.

Example 2.6. Since every non-zero element of the quaternions $\mathbb{H}$ is a unit, the only ideals in $\mathbb{H}$ are the zero ideal and the unit ideal.

For $x \in R$ the set Rx is a left ideal, and xR is a right ideal but neither, in general, is an ideal. The set

$$\langle x \rangle := \{r_1 x r_1' + \cdots + r_n x r_n' : n \in \mathbb{N}_0, r_1, \ldots, r_n, r_1', \ldots, r_n' \in R\}$$

is a subgroup by the subgroup test and is closed under multiplication on the left and right by elements of R and so is an ideal. ⚠ In general $\langle x \rangle \neq RxR$.

Example 2.7. In the ring $M_2(\mathbb{F})$ put

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

Then $I_2 = A + PAP \in \langle A \rangle$, but A is not invertible so none of the matrices in $M_2(\mathbb{F})AM_2(\mathbb{F})$ is invertible, and hence $\langle A \rangle \neq M_2(\mathbb{F})AM_2(\mathbb{F})$.

If there is $x \in R$ such that $I = \langle x \rangle$ then we say I is **principal** and is **generated by** x.

Example 2.8. For $N \in \mathbb{N}^*$, the ideal $\langle N \rangle$ in $\mathbb{Z}$ is the set of multiples of N. Moreover, if I is a non-zero ideal in $\mathbb{Z}$ then it has a minimal positive element N. If $z \in I$, then by the division algorithm we can write $z = Nw + r$ for some $q \in \mathbb{Z}$ and $0 \leqslant r < N$. But $r = z - Nw \in I$ and hence $r = 0$ by minimality of N, and so $I = \langle N \rangle$. In particular, *every* ideal in $\mathbb{Z}$ is principal.

Observation 2.9. Given (left, right, resp. two-sided) ideals $I_1, \ldots, I_n$ in a ring R, $I_1 + \cdots + I_n$ and $\bigcap_{j=1}^n I_j$ are both (left, right, resp. two-sided) ideals.

For $x_1, \ldots, x_n \in R$ we define $\langle x_1, \ldots, x_n \rangle := \langle x_1 \rangle + \cdots + \langle x_n \rangle$, and call it the **ideal generated by** $x_1, \ldots, x_n$. We say that an ideal is **finitely generated** if $I = \langle x_1, \ldots, x_n \rangle$ for some $x_1, \ldots, x_n$.

Remark 2.10. Rings in which every ideal is finitely generated are called **Noetherian** rings and these, and their close cousins for left and right ideals, are very important but will not be our focus in this course.

Example 2.11. The algebraic integers contain an ideal that is not finitely generated. Exercise II.4 develops a proof of this.

Example 2.12. 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 $2q(X) + Xp(X)$ for some $p \in \mathbb{Z}[X]$ and constant polynomial $q \in \mathbb{Z}[X]$.

The ideal $\langle 2, X\rangle$ is not principal. To see this, suppose that $p \in \langle 2, X\rangle$ were such that $\langle 2, X\rangle = \langle p\rangle$. Since $2 \in \langle p\rangle = p(X)\mathbb{Z}[X]$ there is $r \in \mathbb{Z}[X]$ such that $2 = pr$. But $0 = \deg 2 = \deg p + \deg r$, so $\deg p = 0$; say $p(X) = a$ for $a \in \mathbb{Z}^*$. Since $X \in \langle p\rangle = p(X)\mathbb{Z}[X]$ there is $q \in \mathbb{Z}[X]$ such that $X = p(X)q(X)$, and hence $1 = p(1)q(1) = aq(1)$. Hence $p(X) = \pm 1$ and $\langle p\rangle = \mathbb{Z}[X]$ contradicting the fact that $\langle 2, X\rangle \neq \mathbb{Z}[X]$.

Quotient rings

Ideals are particularly important because they let us generalise the construction of the rings $\mathbb{Z}_N$ from $\mathbb{Z}$.

Theorem 2.13. *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 $q : R \to R/I; x \mapsto x + I$ is a surjective ring homomorphism with kernel[‡] I. If R is commutative then so is this multiplication.*

Proof. I is a subgroup of a commutative group and so normal, and so by the quotient group construction R/I is a commutative group and q is a surjective group homomorphism with kernel I. The key is now 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 $u\widehat{\times}v := q(xy)$; this is well-defined by the previous.

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 group by definition of the quotient group; and q is multiplicative by definition. Thus q is

[‡]A ring homomorphism is, in particular, a group homomorphism and so has a kernel.

a ring homomorphism. Moreover, $\widehat{\times}$ is visibly commutative if the multiplication on R is commutative. The result is proved. □

Since the map q above is a surjective ring homomorphism the multiplication on R/I is determined by q: $1_{R/I} = 1 + I$; $(x + I) \times_{R/I} (y + I) = (xy) + I$ for all $x, y \in R$; and if $x \in U(R)$ then $x + I \in U(R/I)$ and $(x + I)^{-1} = x^{-1} + I$, where the first $(\cdot)^{-1}$ is multiplicative inversion in R/I, and the second is in R.

By *the* ring R/I we mean this ring structure and we call this the **quotient ring of** R **by the ideal** I.

Example 2.14. The ring of integers $\mathbb{Z}$ has $\langle N \rangle$ as an ideal, and the quotient ring $\mathbb{Z}/\langle N \rangle$ is none other than the ring $\mathbb{Z}_N$.

Formally $\mathbb{Z}_N$ is realised as a set of cosets, but this can lead to burdensome notation so in practice we just do arithmetic with the integers as usual, but with a coarser notion of equality: that of equivalence $\pmod N$. The fact that we can do this is exactly the fact that the quotient map q is a ring homomorphism.

The same notational convenience is useful in polynomial rings. If $f \in \mathbb{F}[X]^*$ we write $p \equiv q \pmod f$ to mean that $p + \langle f \rangle = q + \langle f \rangle$ or, equivalently, that $p - q$ is a multiple of f. We can do arithmetic in $\mathbb{F}[X]/\langle f \rangle$ by doing it first in $\mathbb{F}[X]$ and then declaring two results to be equivalent if they differ by a multiple of f.

Proposition 2.15. *Suppose that $f \in \mathbb{F}[X]^*$ has degree d. The map $\mathbb{F} \times \mathbb{F}[X]/\langle f \rangle \to \mathbb{F}[X]/\langle f \rangle; (\lambda, p \pmod f) \mapsto \lambda p \pmod f$ is a scalar multiplication of $\mathbb{F}$ on the additive group of $\mathbb{F}[X]/\langle f \rangle$ such that the ring multiplication maps are linear and $1, X, \ldots, X^{d-1}$ is a basis.*

Proof. For the first part it is enough to note that the inclusion map $\mathbb{F} \to \mathbb{F}[X]$ composed with the quotient map $\mathbb{F}[X] \to \mathbb{F}[X]/\langle f \rangle$ induces an $\mathbb{F}$-vector space structure with the given scalar multiplication such that the ring multiplication maps are linear.

To see that $1, X, \ldots, X^d$ is spanning, note that by the division algorithm for polynomials, for every $g \in \mathbb{F}[X]$ there is $q, r \in \mathbb{F}[X]$ with $g(X) = f(X)q(X) + r(X)$ and $r(X) = a_0 + \cdots + a_{d-1}X^{d-1}$, whence $g(X) \equiv a_0 + \cdots + a_{d-1}X^{d-1} \pmod f$.

To see that $1, X, \ldots, X^d$ is linearly independent, suppose that $a_0, \ldots, a_{d-1} \in \mathbb{F}$ have $a_0 + a_1X + \cdots + a_{d-1}X^{d-1} \equiv 0 \pmod f$. If the a_is are not all 0 then the polynomial $r(X) = a_0 + a_1X + \cdots + a_{d-1}X^{d-1}$ has a degree, and its degree is at most $d - 1$. This contradicts the fact it is divisible by f. □

Example 2.16. The ring $\mathbb{R}[X]/\langle X^2 \rangle$ is called the ring of **dual numbers**, and in this ring we have $(1 + X)^n \equiv 1 + nX \pmod{X^2}$. So for a polynomial f we have $f(1 + X) \equiv f(1) + f'(1)X \pmod{X^2}$ where f' denotes the usual derivative of f.

Example 2.17. In the ring $\mathbb{R}[X]/\langle X^2+1\rangle$, we have

$$(a+bX)+(c+dX) \equiv (a+c)+(b+d)X \pmod{X^2+1}$$

and

$$(a+bX)(c+dX) \equiv (ac-bd)+(bc+ad)X \pmod{X^2+1}.$$

These are the same rules for arithmetic as those on the complex numbers with X replaced by i. Put formally, the map

$$\phi : \mathbb{C} \to \mathbb{R}[X]/\langle X^2+1\rangle; a+bi \mapsto a+bX \pmod{X^2+1}$$

is a ring homomorphism. Moreover, ϕ is a surjection because $\{1, X\}$ is a basis for the codomain, so every $f \in \mathbb{R}[X]/\langle X^2+1\rangle$ can be written as $f \equiv a+bX \pmod{X^2+1}$ for some $a, b \in \mathbb{R}$, and hence $f \equiv \phi(a+bi) \pmod{X^2+1}$; and ϕ is an injection because it is a group homomorphism, and if $\phi(a+bi) \equiv 0 \pmod{X^2+1}$ then $a+bi = 0$ since $\{1, X\}$ is a basis. Thus $\mathbb{C}$ and $\mathbb{R}[X]/\langle X^2+1\rangle$ are isomorphic.

The Chinese remainder theorem

Theorem 2.18. *Suppose that R is a ring and $I_1, \ldots, I_n$ are ideals with $I_i + I_j = R$ for all $i \neq j$. Then the map*

$$\psi : R \to (R/I_1) \times \cdots \times (R/I_n); x \mapsto (x+I_1, \ldots, x+I_n)$$

is a surjective ring homomorphism.

Proof. The given map is certainly a ring homomorphism; the content of this proof is surjectivity: For $j \neq i$ let $y_{i,j} \in I_j$ have $1 - y_{i,j} \in I_i$, and put $w_i := y_{i,1} \cdots y_{i,i-1} \cdot y_{i,i+1} \cdots y_{i,n}$. Then $w_i + I_j = I_j$ for all $j \neq i$; and $w_i + I_i = 1 + I_i$. In particular for all $1 \leqslant i \leqslant n$ we have $\psi(w_i) = (0_{R/I_1}, \ldots, 0_{R/I_{i-1}}, 1_{R/I_i}, 0_{R/I_{i+1}}, \ldots, 0_{R/I_n})$ and so if $z \in R^n$ then $\psi(z_1 w_1 + \cdots + z_n w_n) = (z_1 + I_1, \ldots, z_n + I_n)$ and the map is surjective as claimed. □

Remark 2.19. For G a group and $H_1, H_2 \leqslant G$ with $H_1 H_2 = G$ the map $G \to (G/H_1) \times (G/H_2); x \mapsto (xH_1, xH_2)$ is surjective though the codomain need not even be a group; the substance of Theorem 2.18 is in the fact it applies for $n > 2$.

Remark 2.20. The history of this theorem is involved – see [She88] – but the starting point is work of Sun Zi (孫子) from around 400AD with the particular problem of finding an integer z such that $z \equiv 2 \pmod 3$, $z \equiv 3 \pmod 5$, and $z \equiv 2 \pmod 7$. To connect this to Theorem 2.18 note that 3, 5, and 7 are coprime in pairs, so Bezout's Lemma tells us that

$$\langle 3\rangle + \langle 5\rangle = \mathbb{Z}, \langle 3\rangle + \langle 7\rangle = \mathbb{Z}, \text{ and } \langle 5\rangle + \langle 7\rangle = \mathbb{Z},$$

and hence the map

$$\mathbb{Z} \to \mathbb{Z}_3 \times \mathbb{Z}_5 \times \mathbb{Z}_7; z \mapsto (z \pmod 3, z \pmod 5, z \pmod 7)$$

is surjective from which we can conclude that an integer satisfying the desired congruences exists.

The first isomorphism theorem and consequences

Theorem 2.21 (First isomorphism theorem)**.** *Suppose that $\phi : R \to S$ is a ring homomorphism. Then* $\ker \phi$ *is an ideal in R, and the map*

$$\widetilde{\phi} : R/\ker \phi \to S; x + \ker \phi \mapsto \phi(x)$$

is a well-defined injective ring homomorphism. In particular, $R/\ker \phi$ is isomorphic to $\operatorname{Im} \phi$*.*

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 $\phi(xr) = \phi(x)\phi(r) = 0\phi(r) = 0$ since zero annihilates, and similarly $\phi(rx) = 0$. It follows that $xr, rx \in \ker \phi$ so that $\ker \phi$ is an ideal.

The map $\widetilde{\phi}$ is a well-defined injective group homomorphism by the first isomorphism theorem for groups. In addition,

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

and $\widetilde{\phi}(1_R + \ker \phi) = \phi(1_R) = 1_S$. The result is proved. □

Example 2.22. For R a subring of S and $\lambda \in S$ commuting with all elements of R, the kernel of the evaluation homomorphism at λ extending the inclusion homomorphism $R \to S$, that is the set $\{p \in R[X] : p(\lambda) = 0\}$ of polynomials of which λ is a root, is an ideal.

The first isomorphism theorem is often used to show that a given ring homomorphism is well-defined by showing that it arises by factoring a homomorphism that is more easily seen to be well-defined.

Example 2.23. The map ψ from Theorem 2.18 is a surjective ring homomorphism with $\ker \psi = \{x \in R : x \in I_i \text{ for all } i\} = I_1 \cap \cdots \cap I_n$, and hence by the first isomorphism theorem we have an isomorphism between $R/(I_1 \cap \cdots \cap I_n)$ and $(R/I_1) \times \cdots \times (R/I_n)$ when $I_1, \ldots, I_n$ are ideals in R with $I_i + I_j = R$ for all $i \neq j$.

Example 2.24. Given a ring homomorphism $\phi : R \to S$ and an ideal J contained in $\ker \phi$, the map $R/J \to S; x + J \mapsto \phi(x)$ is a well-defined ring homomorphism: Apply the first isomorphism theorem to the map $R \to (R/J) \times S; x \mapsto (x + J, \phi(x))$. The kernel of this map is J since $J \subset \ker \phi$ and hence the map $R/J \to (R/J) \times S; x + J \mapsto (x + J, \phi(x))$ is a well-defined ring homomorphism and the result follows by composition with projection onto the second factor.

Relationship between ideals in R and R/I

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

Theorem 2.25. *Suppose that R is a ring and I is an ideal in R. Then the map*

$$\phi : \text{Ideals}_I(R) \to \text{Ideals}(R/I); I' \mapsto q(I')$$

is a well-defined inclusion-preserving bijection.

Proof. Since q is a surjective ring homomorphism, if I' is an ideal in R then $q(I')$ is an ideal and the map is well-defined. It is visibly inclusion-preserving. If J is an ideal in R/I then $q^{-1}(J)$ is an ideal in R since q is a ring homomorphism. Since $I = 0_{R/I} \in J$ we have $I \subset q^{-1}(J)$, and hence $q^{-1}(J) \in \text{Ideals}_I(R)$. Since q is surjective $q(q^{-1}(J)) = J$, and so ϕ is surjective. Finally, if $I' \neq I''$ are two ideals containing I then $I' + I = I' \neq I'' = I'' + I$ and so, without loss of generality, there is $x \in I''$ such that $(x + I) \cap I' = \varnothing$. It follows that $q(x) \notin q(I')$, and hence $q(I') \neq q(I'')$. In particular, ϕ is injective. □

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

Example 2.26 (Example 1.13, contd.)**.** $\mathbb{Z}_N$ is a ring in which every ideal is principal. To see this, let $\phi : \text{Ideals}_{\langle N \rangle}(\mathbb{Z}) \to \text{Ideals}(\mathbb{Z}_N)$ be the map from the Correspondence theorem and suppose J is an ideal in $\mathbb{Z}_N$. Since every ideal in $\mathbb{Z}$ is principal, $\phi^{-1}(J) = \langle M \rangle$ for some $M \in \mathbb{N}^*$, and furthermore $\langle M \rangle \supset \langle N \rangle$. Since ϕ is a bijection, $J = \phi(\langle M \rangle) = \{Mz \pmod N : z \in \mathbb{Z}\} = \langle M \pmod N \rangle$ is principal.

Proper, prime, and maximal ideals

Some of the properties of ideals are reflected in properties of quotient rings, and we will look at three important ones now. An ideal I is **proper** if $1 \notin I$ or, equivalently, $I \neq R$.

Observation 2.27. An ideal I is proper if and only if R/I is non-trivial since $1_{R/I} \neq 0_{R/I}$ if and only if $1 + I \neq I$, if and only if $1 \notin I$.

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 2.28. *Suppose that R is a commutative ring and I is an ideal in R. Then I is 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. □

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 $J = R$.

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

Proposition 2.29. *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.*

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 $x \in J \setminus I$ 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.) □

Discussion of fields of fractions and their characterisation

A subring of an integral domain is an integral domain and so, in particular, a subring of a field is an integral domain. Conversely we have the following:

Theorem 2.30. *Suppose that R is an integral domain. Then there is a field $\mathbb{F}$ containing R as a subring.*

Remark 2.31. The proof of this is omitted, but such a field can be constructed in a similar way to the way to which one constructs the integers from the naturals by 'adding in' the negative numbers.

For R an integral domain and $\mathbb{F}$ a field containing R the **field of fractions of R in $\mathbb{F}$** is the field $\mathrm{Frac}_{\mathbb{F}}(R) := \{ab^{-1} : a \in R, b \in R^*\}$. This is a subring of $\mathbb{F}$ containing R by the subring test since it contains $1 = 1.1^{-1}$, and is closed under subtraction and multiplication since

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

Now, if $ab^{-1} \neq 0$ then $a \in R^*$ so $ba^{-1} \in \mathrm{Frac}_{\mathbb{F}}(R)$, and hence $\mathrm{Frac}_{\mathbb{F}}(R)$ is closed under multiplicative inverses and so a field.

For the most part the containing field $\mathbb{F}$ will be clear – indeed it will very often be $\mathbb{C}$ – in which case we write $\mathrm{Frac}(R)$ for $\mathrm{Frac}_{\mathbb{F}}(R)$.

Example 2.32. The field of fractions of $\mathbb{Z}$ in $\mathbb{C}$ is $\mathbb{Q}$, and this is the prototype.

Observation 2.33. If $R \subset \mathbb{K} \subset \mathbb{F}$ for fields $\mathbb{K}$ and $\mathbb{F}$, then $\mathrm{Frac}_{\mathbb{F}}(R) \subset \mathbb{K}$. In particular, $\mathrm{Frac}_{\mathbb{F}}(\mathbb{K}) = \mathbb{K}$.

Our definition of field of fractions characterises it in the following sense:

Theorem 2.34. *Suppose that $\mathbb{F}$ and $\mathbb{K}$ are fields containing R as a subring. Then there is a unique isomorphism $\phi : \mathrm{Frac}_{\mathbb{F}}(R) \to \mathrm{Frac}_{\mathbb{K}}(R)$ such that $\phi(r) = r$ for all $r \in R$.*

Remark 2.35. Again we omit the proof, but the idea is to define ϕ by $\phi(rs^{-1}) := rs^{-1}$ for $r \in R$ and $s \in R^*$. ⚠ s^{-1} on the left is the inverse of s in $\mathbb{F}$, and on the right in $\mathbb{K}$.

Example 2.36. We write $\mathbb{Q}(i)$ for their field of fractions of the Gaussian integers inside $\mathbb{C}$. Since $\mathbb{Z} \subset \mathbb{Z}[i]$ we must have $\mathbb{Q} \subset \mathbb{Q}(i)$, and since $i \in \mathbb{Z}[i]$ we must have $\mathbb{Q} + i\mathbb{Q} \subset \mathbb{Q}(i)$. On the other hand by the subring test $\mathbb{Q} + i\mathbb{Q}$ is a subring of $\mathbb{C}$, and if $0 \neq a + bi \in \mathbb{Q}[i]$ then

$$(a + bi)^{-1} = \frac{a}{a^2 + b^2} + i\frac{-b}{a^2 + b^2} \in \mathbb{Q} + i\mathbb{Q},$$

so $\mathbb{Q} + i\mathbb{Q}$ is a field and hence $\mathbb{Q}(i) = \mathbb{Q} + i\mathbb{Q}$.

⚠ Complex conjugation $\mathbb{Q}(i) \to \mathbb{Q}(i); z \mapsto \overline{z}$ is an isomorphism that is different from the identity map $\mathbb{Q}(i) \to \mathbb{Q}(i); z \mapsto z$ isomorphism, but complex conjugation is not the identity on $\mathbb{Z}[i]$, and hence this does not violate the uniqueness of the isomorphism in Theorem 2.34.

Field extensions

We say that $\mathbb{K}$ is a **field extension** of $\mathbb{F}$ if $\mathbb{K}$ is a field and $\mathbb{F}$ is a subfield of $\mathbb{K}$. Given a field extension $\mathbb{K}$ of $\mathbb{F}$, the inclusion map $\mathbb{F} \to \mathbb{K}$ induces an $\mathbb{F}$-vector space structure on $\mathbb{K}$ (such that the multiplication maps on $\mathbb{K}$ are $\mathbb{F}$-linear) and we call the dimension of this the **degree** of the field extension, denoted $|\mathbb{K} : \mathbb{F}|$.

Given a field extension $\mathbb{K}$ of $\mathbb{F}$, we say $\alpha \in \mathbb{K}$ is **$\mathbb{F}$-algebraic** if there is some $p \in \mathbb{F}[X]^*$ such that $p(\alpha) = 0$, and it is **$\mathbb{F}$-transcendental** if there is no such polynomial.

Example 2.37. $\mathbb{C}$ is a field extension of $\mathbb{R}$ of degree 2, and *any* $z \in \mathbb{C}$ is $\mathbb{R}$-algebraic since $p(X) := X^2 - 2\,\mathrm{Re}\,zX + |z|^2$ has $p \in \mathbb{R}[X]^*$ and $p(z) = 0$.

Example 2.38. $\mathbb{R}$ is an infinite degree field extension of $\mathbb{Q}$, and α in $\mathbb{R}$ is $\mathbb{Q}$-algebraic (resp. $\mathbb{Q}$-transcendental) if and only if it is algebraic (resp. transcendental) in the usual sense.

For $\mathbb{F}$ a subfield of $\mathbb{K}$, and $\alpha \in \mathbb{K}$, the set $\mathbb{F}[\alpha]$ (recall the definition from Example 2.22) is an integral domain since it is a subring of a field, but in general $\mathbb{F}[\alpha]$ is not a field. We write $\mathbb{F}(\alpha)$ for $\mathrm{Frac}_{\mathbb{K}}(\mathbb{F}[\alpha])$, the field of fractions of $\mathbb{F}[\alpha]$, and call it the **field $\mathbb{F}$ adjoined by α** – we 'construct $\mathbb{F}(\alpha)$ by adjoining α to $\mathbb{F}$'.

Example 2.39. The ring $\mathbb{Q}[\sqrt{2}]$ contains $\mathbb{Q}$ as a subfield and is a ring by the subring test, and so an integral domain. It follows by Proposition 1.34 that it is in fact a field and so $\mathbb{Q}(\sqrt{2}) = \mathbb{Q}[\sqrt{2}]$.

The ring $\mathbb{Q}[\sqrt{2}+\sqrt{3}]$ is certainly contained in $\mathbb{Q}+\sqrt{2}\mathbb{Q}+\sqrt{3}\mathbb{Q}+\sqrt{6}\mathbb{Q}$ which itself is a ring by the subring test and hence an integral domain. This ring contains $\mathbb{Q}$ as a subfield and so has an induced $\mathbb{Q}$-vector space structure in which it is (at most) 4-dimensional – in particular it is finite dimensional – and so by Proposition 1.34 in fact it is a field, and hence it contains the field $\mathbb{Q}(\sqrt{2}+\sqrt{3})$, and we conclude the latter is a field extension of $\mathbb{Q}$ of degree at most 4.

Theorem 2.40 (Tower Law)**.** *Suppose that $\mathbb{L}$ is a field extension of $\mathbb{K}$ and $\mathbb{K}$ is a field extension of $\mathbb{F}$. Then $\mathbb{L}$ is a field extension of $\mathbb{F}$, 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. The first part is immediate because the relation 'is a subfield of' is transitive, and certainly if $|\mathbb{L}:\mathbb{F}| < \infty$ then $|\mathbb{L}:\mathbb{K}|, |\mathbb{K}:\mathbb{F}| < \infty$. Let $e_1, \ldots, e_n$ be a basis for $\mathbb{L}$ as a vector space over $\mathbb{K}$, and let $f_1, \ldots, f_m$ be a basis for $\mathbb{K}$ as a vector space over $\mathbb{F}$. Now, for $x \in \mathbb{L}$ there are scalars $\lambda_1, \ldots, \lambda_n \in \mathbb{K}$ such that $x = \lambda_1 e_1 + \cdots + \lambda_n e_n$, and since $f_1, \ldots, f_m$ is spanning, for each $1 \leqslant j \leqslant n$ there are scalars $\mu_{1,j}, \ldots, \mu_{m,j} \in \mathbb{F}$ such that $\lambda_j = \mu_{1,j} f_1 + \cdots + \mu_{m,j} f_m$. Hence $x = \sum_{j=1}^{n} \sum_{i=1}^{m} \mu_{i,j} f_i e_j$, so we have that $(f_i e_j)_{i=1,j=1}^{m,n}$ is an $\mathbb{F}$-spanning subset of $\mathbb{L}$. Now suppose $\mu_{1,1}, \ldots, \mu_{m,n} \in \mathbb{F}$ are such that $\sum_{j=1}^{n} \sum_{i=1}^{m} \mu_{i,j} f_i e_j = 0_{\mathbb{L}}$. 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 $e_1, \ldots, e_n$ 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 $f_1, \ldots, f_m$ 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 $(f_i e_j)_{i=1,j=1}^{m,n}$ is a basis for $\mathbb{L}$ as an $\mathbb{F}$-vector space as required. □

Remark 2.41. If $\mathbb{F}$ is a finite field, and $|\mathbb{K}:\mathbb{F}| = n$, $|\mathbb{L}:\mathbb{K}| = m$, and $|\mathbb{L}:\mathbb{F}| = k$ then $|\mathbb{K}| = |\mathbb{F}|^n$, $|\mathbb{L}| = |\mathbb{K}|^m$, and $|\mathbb{L}| = |\mathbb{F}|^k$ from which it follows that $k = nm$. The proof above is really just the observation that we only need to use the 'relative size of $\mathbb{F}$ in $\mathbb{K}$'.

Example 2.42 (Example 2.39, contd.)**.** The field $\mathbb{Q}(\sqrt{2}+\sqrt{3})$ contains $\sqrt{2} = \frac{1}{2}((\sqrt{2}+\sqrt{3})^3 - 9(\sqrt{2}+\sqrt{3}))$, and hence also contains $\sqrt{3}$. Now, $\sqrt{3} \notin \mathbb{Q}(\sqrt{2})$. Indeed, suppose for a contradiction that there were $a, b \in \mathbb{Q}$ with $\sqrt{3} = a + b\sqrt{2}$ (which would have to be the case since $\mathbb{Q}[\sqrt{2}] = \mathbb{Q}(\sqrt{2})$). Then squaring both sides and using the irrationality of $\sqrt{2}$ (which exactly says that 1 and $\sqrt{2}$ are rationally independent), we have $2ab = 0$. But $b \neq 0$ since $\sqrt{3}$ is irrational; and $a \neq 0$ since $\sqrt{3}/\sqrt{2}$ is irrational. We have a contradiction.

By the Tower Law $|\mathbb{Q}(\sqrt{2}+\sqrt{3}):\mathbb{Q}(\sqrt{2})||\mathbb{Q}(\sqrt{2}):\mathbb{Q}| = |\mathbb{Q}(\sqrt{2}+\sqrt{3}):\mathbb{Q}| \leqslant 4$. However, $|\mathbb{Q}(\sqrt{2}):\mathbb{Q}| \geqslant 2$, since $\sqrt{2} \notin \mathbb{Q}$; and $|\mathbb{Q}(\sqrt{2}+\sqrt{3}):\mathbb{Q}(\sqrt{2})| \geqslant 2$ since $\sqrt{3} \notin \mathbb{Q}(\sqrt{2})$. Hence both of these extensions are of degree exactly 2, and $|\mathbb{Q}(\sqrt{2}+\sqrt{3}):\mathbb{Q}|$ is 4.

3 Divisibility

Divisibility in $\mathbb{Z}$ is a mysterious relation of intrinsic mathematical interest as well as wider importance. It is similar to divisibility in rings of the form $\mathbb{F}[X]$, and in this section we look to understand the source of these similarities.

In a commutative ring R we say a is a **divisor** of b, or a **divides** b, or b is a **multiple** of a, and write $a \mid b$, if there is $x \in R$ such that $b = ax (= xa)$; or, equivalently, if $b \in \langle a \rangle$; or, equivalently, if $\langle b \rangle \subset \langle a \rangle$.

Observation 3.1. If $a \mid b_1, \ldots, b_n$, and $x_1, \ldots, x_n, y_1, \ldots, y_n \in R$ then $a \mid x_1 b_1 y_1 + \cdots + x_n b_n y_n$.

The relation $\mid$ is reflexive and transitive – relations that are reflexive and transitive are sometimes called preorders, and we shall think of divisibility with the language of order in mind. When $a \mid b$ *and* $b \mid a$ we say that a and b are **associates** and write $a \sim b$; $\sim$ is an equivalence relation.

Example 3.2. Divisibility in fields is very simple: all elements divide zero, and every non-zero element divides every other non-zero element, and so all non-zero elements are associates.

Lemma 3.3. *Suppose that R is an integral domain. Then*

(i) for all $x \in R^*$, $xa \mid xb$ if and only if $a \mid b$;

(ii) $a \sim b$ if and only if there is $u \in U(R)$ such that $a = bu$.

Proof. For (i) the 'if' is immediate. To prove the 'only if' suppose $xa \mid xb$. Then there is $z \in R$ such that $x(az) = (xa)z = xb$. $x \neq 0$ and so left multiplication by x is injective, and $az = b$ *i.e.* $a \mid b$.

For (ii), again the 'if' part is immediate. To prove the 'only if' suppose $a \sim b$. Then $a \mid b$ and $b \mid a$, so there are $v, w \in R$ such that $av = b$ and $bw = a$, and hence $b(wv) = (bw)v = av = b = b1$. If $b \neq 0$ then left multiplication by b is injective and $1 = wv (= vw)$ so $w \in U(R)$ and we may take $u = w$; if $b = 0$ then $a = 0$, and we may take u to be any unit. □

Remark 3.4. The commutative rings where (i) holds are exactly the integral domains, since if R is a commutative ring that is not an integral domain then there are $x, a \in R^*$ with $xa = 0$, and so $xa \mid x0$, but $a \nmid 0$.

Commutative rings where (ii) holds are sometimes called associator rings. Exercise I.6 asks for a proofs that $C(\mathbb{R})$ with the operations of pointwise addition and multiplication,

which is a commutative ring, is not an associator ring; and that $\mathbb{Z}_N$ is an associator ring, though $\mathbb{Z}_N$ is not an integral domain when N is not prime.

Irreducibles, primes, and uniqueness of factorisation

We say that $x \in R$ is **irreducible** if $x \nmid 1$ and whenever $a \mid x$ we have $a \sim x$ or $a \sim 1$; or, equivalently, if $\langle x \rangle$ is maximal amongst proper *principal* ideals. In particular, if $y \sim x$ and x is irreducible then y is also irreducible.

Remark 3.5. ⚠ 0 is sometimes explicitly excluded from being irreducible. If 0 is irreducible in the sense above, then in fact R is a field: For $x \in R^*$ we have $\langle x \rangle \supsetneq \langle 0 \rangle$, and so by the maximality of $\langle 0 \rangle$ amongst proper *principal* ideals, we conclude that $\langle x \rangle$ is not proper *i.e.* $\langle x \rangle = R$. Hence there is $y \in R$ with $1 = xy (= yx)$, meaning $x \in U(R)$.

Example 3.6. ⚠ Irreducible elements can have unexpected behaviours: $2 \equiv 2 \times 2 \times 2 \pmod 6$ but 2 is irreducible in $\mathbb{Z}_6$ (the ideal does not contain 3, and so is proper, and has index 2, so by Lagrange's theorem is maximal.

Example 3.7. The irreducible positive integers in $\mathbb{Z}$ are exactly the prime numbers, and hence the irreducible integers are those of the form $\pm p$ for p a prime number.

Example 3.8. The algebraic integers are a non-trivial commutative ring containing no irreducible elements. (Exercise II.4 asks for a proof.)

We say that an element $x \in R$ is **prime** if $x \nmid 1$, and $x \mid ab$ implies $x \mid a$ or $x \mid b$. In the language of ideals $\langle x \rangle$ is a prime ideal.

Observation 3.9. By induction if x is prime and $x \mid \prod_{i \in I} b_i$ then there is $i \in I$ such that $x \mid b_i$.

Example 3.10. ⚠ In the ring $\mathbb{Z}$ this replaces any previous definition of prime, though we shall see later that a positive integer is prime in the old sense if and only if it is prime in the new sense.

The integer 2 is prime because i) it is *not* either 1 or -1; and ii) if $2 \mid ab$ – in words, if ab is even – then $2 \mid a$ or $2 \mid b$ – in words, at least one of a or b is even.

The integer 0 is prime because i) it is *not* either 1 or -1; and ii) if $0 \mid ab$ then in fact $0 = ab$ and so either $0 = a$, which can be rewritten as $0 \mid a$, or $0 = b$, which can be rewritten as $0 \mid b$. This is the special case in the integers of the fact in Proposition 2.28 that a ring is an integral domain if and only if 0_R is prime.

Example 3.11. For R an integral domain and $\alpha \in R$, if $X - \alpha \mid f(X)g(X)$ then $f(\alpha)g(\alpha) = 0$ and hence either $f(\alpha) = 0$ and $X - \alpha \mid f(X)$ by the Factor theorem, or $g(\alpha) = 0$ and similarly $X - \alpha \mid g(X)$. Since $X - \alpha \nmid 1$ we have that it is prime.

Proposition 3.12. *Suppose that R is an integral domain. Then $r \in R$ is prime as an element of R, if and only if r is prime as an element of $R[X]$.*

Proof. First $U(R) = U(R[X])$ and so $r \not\sim 1$ in R if and only if $r \not\sim 1$ in $U(R[X])$.

Suppose r is prime in $R[X]$, and that $r \mid ab$ in R. If either a or b is 0 then without loss of generality $a = 0$, and $r \mid a$. Thus we may restrict attention to the case when $a, b \in R^*$. By primality of r in $R[X]$, without loss of generality $r \mid a$ in $R[X]$. Hence there is $p(X) \in R[X]$ such that $rp(X) = a$. Since $a \in R^*$ we have $\deg p = \deg r + \deg p = \deg a = 0$, and hence $r \mid a$ in R as required.

Now suppose that r is prime in R, and $r \mid pq$ in $R[X]$ with $p(X) = a_0 + a_1X + \cdots + a_nX^n$ and $q(X) = b_0 + b_1X + \cdots + b_mX^m$ with $r \nmid p$ 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

$$\sum_{j=0}^{k+l} a_j b_{k+l-j} = \sum_{j=0}^{k-1} a_j b_{k+l-j} + a_k b_l + \sum_{j=0}^{l-1} a_{k+l-j} b_j.$$

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 3.13. ⚠ Neither direction follows simply because R is a subring of $R[X]$: For example $\mathbb{Z}$ is a subring of $\mathbb{Q}$ and 2 is prime in $\mathbb{Z}$; and $\mathbb{Z}[X]$ is a subring of $\mathbb{Q}[X]$, but $2X$ is prime in $\mathbb{Q}[X]$, since $2X \sim X$ and X is prime in $\mathbb{Q}[X]$, but $2X$ is not prime in $\mathbb{Z}[X]$ since $2X \mid 2 \times X$ but $2X \nmid 2$ and $2X \nmid X$.

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** of the factorisation, 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 a 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 $1 \leqslant i \leqslant r$. ⚠ In particular, every unit has a unique factorisation into irreducibles with the convention that the empty product is 1_R.

Proposition 3.14. *Suppose that R is an integral domain and $x \in R^*$ has a (possibly empty) factorisation in which every factor is prime. Then any factorisation of x into irreducibles is unique.*

Proof. Let $(x_1, \ldots, x_r)$ be a factorisation of x in which every factor is prime. We shall prove that if $(y_i)_{i \in I}$ are non-zero irreducible elements indexed by a finite set I such that $x \sim \prod_{i \in I} y_i$

then there is a bijection $\pi : \{1, \dots, r\} \to I$ such that $x_i \sim y_{\pi(i)}$ for all $1 \leqslant i \leqslant r$, and by transitivity of association the result follows.

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 so $y_j \in U(R)$. 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 $x_r \mid \prod_{i \in I} y_i$. By primality there is some $j \in I$ such that $x_r \mid y_j$. But y_j is irreducible and $x_r \nmid 1$ and so $x_r \sim y_j$. Cancelling y_j we get $x_1 \cdots x_{r-1} \sim \prod_{i \in I \setminus \{j\}} y_i$ and by the inductive hypothesis there is a bijection $\tilde{\pi} : \{1, \dots, r-1\} \to I \setminus \{j\}$ such that $x_i \sim y_{\tilde{\pi}(i)}$ for all $1 \leqslant i \leqslant r-1$. Extend this to a bijection $\pi : \{1, \dots, r\} \to I$ by setting $\pi(r) = j$ and the result is proved. □

Proposition 3.15. *Suppose that R is an integral domain and $x \in R^*$ is prime. Then x is irreducible.*

Proof. First, $x \nmid 1$. Now suppose that $a \mid x$. Then there is $b \in R$ such that $x = ab$, and $b \neq 0$ since $x \neq 0$. By primality of x either $x \mid a$ and so $x \sim a$; or $ab = x \mid b$, but $b \neq 0$ and so $a \mid 1$, and hence $a \sim 1$ (since certainly $1 \mid a$). □

Example 3.16 (Example 3.11, contd.)**.** For R an integral domain we saw that the polynomials $X - \alpha$ are prime in $R[X]$, but then $R[X]$ is an integral domain and so $X - \alpha$ is irreducible by the above.

Example 3.17. 2 is an irreducible element of $\mathbb{Z}[\sqrt{-5}]$ that is not prime; Exercise II.2 asks for a proof.

⚠ $X \pmod{2X}$ is a non-zero prime in the commutative ring $R = \mathbb{Z}[X]/\langle 2X \rangle$, but it is not irreducible. Of course R is *not* an integral domain! For primality, the evaluation homomorphism $\mathbb{Z}[X] \to \mathbb{Z}; p(X) \mapsto p(0)$ has kernel $\langle X \rangle$, which contains $\langle 2X \rangle$ and hence by Example 2.24 $R \to \mathbb{Z}; p(X) \pmod{2X} \mapsto p(0)$ is a well-defined surjective ring homomorphism. Its kernel is $\langle X \pmod{2X} \rangle$, and $\mathbb{Z}$ is an integral domain so $X \pmod{2X}$ is prime by the first isomorphism theorem and Proposition 2.28. To see $X \pmod{2X}$ is not irreducible, $\langle 3 \pmod{2X} \rangle$ is a proper principal ideal in R which properly contains $\langle X \pmod{2X} \rangle$.

In the integers (as we shall see shortly) the converse of Proposition 3.15 holds as a consequence of Bezout's Lemma, and we make a definition which captures rings in which Bezout's Lemma holds: we say that an integral domain R is a **Bezout domain** if every finitely generated ideal is principal.

Example 3.18. The ring of integers, $\mathbb{Z}$, is a Bezout domain since every ideal is principal, so in particular every finitely generated ideal in principal. However, the connection with Bezout's Lemma is closer: in the language of ideals this states that any ideal in $\mathbb{Z}$ that is

generated by two elements can also be generated by one element *i.e.* is principal, and by induction that any finitely generated ideal in $\mathbb{Z}$ is principal.

Example 3.19. The algebraic integers $\overline{\mathbb{Z}}$ is a Bezout domain. A proof may be found in [Kap70, Theorem 102] though the prerequisites are considerable.

Example 3.20. $\mathbb{Z}[X]$ is an example of an integral domain that is not a Bezout domain because (as we saw in Example 2.12) $\langle 2, X\rangle$ is finitely generated but not principal.

Proposition 3.21. *Suppose that R is a Bezout domain and $x \in R$ is irreducible. Then x is prime.*

Proof. Suppose $x \mid ab$ and let d be a generator of the ideal $\langle x, b\rangle$. Then $d \mid x$, and since x is irreducible either $d \sim x$ or $d \sim 1$. Since we also have $d \mid b$, if $d \sim x$ then $x \mid d \mid b$. On the other hand, if $d \sim 1$ then there are elements $u, v \in R$ such that $1 = ux + bv$. Multiplying by a we have $aux + abv = a$, but $x \mid aux$ and $x \mid abv$, and so $x \mid a$ as required. □

Proposition 3.22. *Suppose that R is a Bezout domain. Then for every pair $a, b \in R$ there is d and l with $ab = ld$, and $\langle a\rangle + \langle b\rangle = \langle d\rangle$, and $\langle a\rangle \cap \langle b\rangle = \langle l\rangle$.*

Proof. Since every finitely generated ideal in R is principal there is some $d \in R$ such that $\langle a\rangle + \langle b\rangle = \langle a, b\rangle = \langle d\rangle$. Let $x, y \in R$ be such that $d = xa + by$, and $z, w \in R$ be such that $b = zd$ and $a = dw$; put $l := zdw$ and note $ab = ld$.

Now, $l = bw \in \langle b\rangle$ and $l = za \in \langle a\rangle$, so $l \in \langle a\rangle \cap \langle b\rangle$. On the other hand if $m \in \langle a\rangle \cap \langle b\rangle$ then $a, b \mid m$ so $ab \mid am$, and $ab \mid mb$. Hence $ld = ab \mid xam + mby = md$. If $d \neq 0$ then by cancellation we have $l \mid m$ which is to say $m \in \langle l\rangle$ as required. If $d = 0$ then $a = b = 0$, and so $l = 0$ and we are done. □

Remark 3.23. The set $\langle a\rangle$ is the set of multiples of a, and the set $\langle b\rangle$ is the set of multiples of b, hence $\langle a\rangle \cap \langle b\rangle$ is the set of common multiples of a and b, and to say that it is generated by l is exactly to say that there is a common multiple of a and b that divides all other common multiples – such a common multiple is called a **least common multiple (lcm)**.

On the other hand if $\langle a\rangle + \langle b\rangle = \langle d\rangle$, then $a \in \langle d\rangle$ and $b \in \langle d\rangle$ so that $d \mid a$ and $d \mid b$ *i.e.* d is a common divisor of a and b. Moreover there are $z, w \in R$ with $d = xa + by$, so if c is another common divisor of a and b then $c \mid xa + by = d$ – which is to say that every common divisor of a and b divides d. A common divisor such that every other common divisor is a factor is called a **greatest common divisor (gcd)**.

Euclidean domains and division algorithms

The process of dividing integers (or polynomials) is captured by the division algorithm, and rings where we have such an algorithm will be particularly good to work with. A **Euclidean function** on an integral domain R is a function $f : R^* \to \mathbb{N}_0$ such that

- $f(a) \leqslant f(b)$ whenever $a \mid b$ (both non-zero);
- and 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 3.24. ⚠ Keating [Kea98, p17] 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].

Example 3.25. Let $f : \mathbb{F}^* \to \mathbb{N}_0$ be the constant function 1. Since $f(a) = f(b)$ for all a and b, and 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.

Example 3.26. 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 $\mathbb{Z}$ and $\mathbb{Z}$ is a Euclidean domain. (It certainly has $|a| \leqslant |b|$ whenever $a \mid b$.)

Example 3.27. If $a, b \in \mathbb{F}[X]^*$ and $b \nmid a$ then $a - bq$ is not the zero polynomial for any $q \in \mathbb{F}[X]$, and we can pick bq such that $a - bq$ has smallest possible degree. Then $r := a + bq$ has $\deg r < \deg b$, since otherwise writing λ for the ratio between the lead coefficient of r and that of b we have $r(X) - \lambda X^{\deg r - \deg b} b(X)$ of the form $a - bq'$ and of strictly smaller degree than r. Finally, $\deg p \leqslant \deg q$ whenever $p \mid q$, and so deg is a Euclidean function and $\mathbb{F}[X]$ is a Euclidean domain.

An integral domain in which every ideal is principal is called a **principal ideal domain (PID)**. In particular, every PID is *a fortiori* a Bezout domain so all the work of the previous section applies to PIDs.

Proposition 3.28. *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 3.29. The ring $\mathbb{Z}[\theta]$, where $\theta^2 - \theta + 5 = 0$, is an example of a PID that is not a Euclidean domain, though in view of Exercise II.9 we shall not treat them very differently; a proof may be found in [Con, Theorem 5.13].

The ACCP and unique factorisation domains

Other than Bezout's lemma, the integers enjoy another important property: we cannot 'keep dividing indefinitely', and this is what ensures the existence of factorisations into primes.

An integral domain R has the **ascending chain condition on principal ideals** or **ACCP** if for every sequence $(d_n)_{n=0}^{\infty}$ of elements 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 $n \geqslant N$.

Proposition 3.30. *Suppose that R is a Bezout domain. Then R has the ACCP if and only if R is a PID.*

Proof. For 'only if', suppose that I is an ideal in R that is not principal and generate a chain of elements of I iteratively: Let $d_0 \in I$, and suppose we have $d_0, \dots, d_n \in I$. Since I is not principal there is $d' \in I \setminus \langle d_n \rangle$ and since R is Bezout there is d_{n+1} such that $\langle d_{n+1} \rangle = \langle d', d_n \rangle \subset I$, so in particular $d_{n+1} \in I$. Since $d' \notin \langle d_n \rangle$ we have $d_{n+1} \mid d_n$ and $d_{n+1} \nmid d_n$. The chain $(d_n)_{n=0}^{\infty}$ violates the ACCP, and this contradiction proves the result.

For the 'if' direction, suppose that $(d_n)_{n=0}^{\infty}$ has $d_{n+1} \mid d_n$ for all $n \in \mathbb{N}_0$, so that $\langle d_0 \rangle \subset \langle d_1 \rangle \subset \cdots$ and let $I = \bigcup_{n \in \mathbb{N}_0} \langle d_n \rangle$. I is an ideal: If $s, t \in I$ then there are $n, m \in \mathbb{N}_0$ such that $s \in \langle d_n \rangle$ and $t \in \langle d_m \rangle$ and so $s, t \in \langle d_{\max\{n,m\}} \rangle$ by nesting, and hence $s - t \in \langle d_{\max\{n,m\}} \rangle \subset I$. Since $0 \in I$, it is a subgroup by the subgroup test, and finally if $r \in R$ then $rs, sr \in \langle d_n \rangle \subset I$ as required.

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 $n \geqslant N$. □

Example 3.31. The ring of algebraic integers $\overline{\mathbb{Z}}$ does not satisfy the ACCP giving an example of a Bezout domain that is not a PID. Exercise II.4 develops a proof of this.

Proposition 3.32. *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 a 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 $x_0 \in R^* \setminus \mathcal{F}$. Now create a chain iteratively: at step i suppose we have $x_i \in R^* \setminus \mathcal{F}$. Since x_i is not irreducible and not a unit there is $y_i \mid x_i$ with $y_i \nmid 1$ and $y_i \nmid x_i$; 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 $y_i, z_i \nmid x_i$.

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 $x_{i+1} \nmid x_i$. This process produces a sequence $\cdots \mid x_2 \mid x_1 \mid x_0$ in which $x_i \nmid x_{i+1}$ for all $i \in \mathbb{N}_0$ contradicting the ACCP. □

Remark 3.33. 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 Anne 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.

Theorem 3.34. *Suppose that R is a PID. Then R is a UFD.*

Proof. Since every PID has the ACCP, Proposition 3.32 tells us that every $x \in R^*$ has a factorisation into irreducibles. But every PID is a Bezout domain, and every irreducible in a Bezout domain is prime, and the result follows from Proposition 3.14. □

Example 3.35. $\mathbb{Z}[X]$ is an example of a UFD that is not a PID; see Exercise II.8 for details.

Finding irreducibles

Irreducible elements of a ring are of interest in the same way that the elements (in the sense of the periodic table) are of interest in chemistry: they are the building blocks of the non-zero elements (in the sense of elements of a set) of the ring.

In PIDs irreducibles are of even more interest because they generate maximal ideals: not just maximal amongst principal ideals, but maximal amongst *all* ideals, because all ideals are principal in a PID. This means that the quotient of a PID by the ideal generated by an irreducible element produces a field. We have already seen this with the primes in $\mathbb{Z}$ producing the fields $\mathbb{F}_p$, but there are many more fields arising from quotient rings.

We begin with a short technical lemma which can help in finding irreducible polynomials of degree 2 and 3.

Lemma 3.36. *Suppose that R is an integral domain and $f \in R[X]$. Then if f has a root and degree at least 2, it is not irreducible; and if f is monic of degree at most 3 and is not irreducible then it has a root.*

Proof. If f has a root α then by the Factor theorem $X - \alpha$ divides f. Since $\deg(X - \alpha) = 1$ we have $X - \alpha \nmid 1$, and since additionally $\deg f \geqslant 2$ we have $X - \alpha \nmid f$. We conclude that f is not irreducible.

If f has degree at most 3, and $g \mid f$ has $g \nmid 1$ and $g \nmid f$ then let $h \in R[X]^*$ be such that $f = gh$. Since $g, h \mid f$, and f is monic the lead coefficients of g and h are both units. Since $g \nmid 1$ we have $\deg g > 0$; since $g \nmid f$ we have $\deg g < \deg f$. But then since $3 \geqslant \deg f = \deg g + \deg h$ we have either $\deg g = 1$ or $\deg h = 1$. In the first case, since the lead coefficient of g is a unit, g has a root in R; in the second case similarly h has a root in R. It follows that f has a root in R. □

Example 3.37. $X^2 + X + 1 \in \mathbb{F}_2[X]$ has no root in $\mathbb{F}_2$ and is monic, so is irreducible in $\mathbb{F}_2[X]$. Hence $(X^2 + X + 1)^2 = X^4 + X^2 + 1$ is not irreducible but it is also monic and has no root.

Example 3.38. The polynomials X^3+X^2+1 and X^3+X+1 are the only degree 3 irreducible polynomials in $\mathbb{F}_2[X]$: There are only eight degree 3 polynomials in $\mathbb{F}_2[X]$ and the constant term may not be 0, or else 0 is a root. Hence there are only four polynomials to consider: $X^3 + X^2 + X + 1$, $X^3 + X + 1$, $X^3 + X^2 + 1$, and $X^3 + 1$. The first and last of these have 1 as a root, and the other two do not.

Every finite field has size a power of a prime (Exercise I.7 asks for a proof of this), and we can produce a field of order p^n for p a prime if we can find $f \in \mathbb{F}_p[X]$ irreducible of degree n. A proof that we can find such irreducibles, modelled on the proof of Bertrand's postulate, may be found in [Sou20]; for now we content ourselves for finding a large class of fields of order p^2:

Example 3.39. We call $a \in \mathbb{F}_p$ a **quadratic non-residue** if there is no $x \in \mathbb{F}_p$ such that $x^2 \equiv a \pmod{p}$. For example, -1 is a quadratic non-residue if p is a prime with $p \equiv 3 \pmod 4$ because if $x \in \mathbb{F}_p$ had $x^2 \equiv -1 \pmod{p}$ then x would generate a subgroup of order 4 in $U(\mathbb{F}_p)$. However, $U(\mathbb{F}_p)$ has order $p-1$, which is not divisible by 4 violating Lagrange's theorem.

Thus when $p \equiv 3 \pmod 4$, $X^2 + 1$ is irreducible, and hence $\mathbb{F}_p[X]/\langle X^2 + 1\rangle$ is a field and it is 2-dimensional in the $\mathbb{F}_p$-vector space structure induced by the quotient map (composed with the inclusion of $\mathbb{F}_p$). In particular, it has size p^2 and so is not isomorphic to $\mathbb{F}_q$ for any prime q – these are new fields – and it is not isomorphic to $\mathbb{Z}_{p^2}$ since this is not even an integral domain.

The rationals are an infinite field and so checking a polynomial for rational roots does not yield to the same brute force approaches that can work in finite fields. However, there is a result of Gauss which lets us connect irreducibility of polynomials in $\mathbb{Z}[X]$, where we only have to check for integer roots, with irreducibility in $\mathbb{Q}[X]$.

Example 3.40. ⚠ $2X \in \mathbb{Z}[X]$ is not irreducible in $\mathbb{Z}[X]$ because $2 \mid 2X$ and $2 \nmid 1$ and $2 \nmid X$. On the other hand $2X \sim X$ in $\mathbb{Q}[X]$, and so it *is* irreducible in $\mathbb{Q}[X]$.

We say that $f \in \mathbb{Z}[X]$ is **primitive** if 1 is a greatest common divisor of the coefficients in f. In particular, if f is primitive and of degree 0 then f is a unit in $\mathbb{Z}[X]$.

Theorem 3.41 (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 it were not there would be $n \nmid 1$ such that $n \mid f$ in $\mathbb{Z}[X]$. Since $n \nmid 1$ we conclude that $n \sim f$ (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. □

Example 3.42. The polynomial $p(X) = X^3 + X + 1$ is non-constant and irreducible in $\mathbb{Z}[X]$ because it has degree at most 3 and no root in $\mathbb{Z}$. Hence it is irreducible in $\mathbb{Q}[X]$.

Proposition 3.43 (Eisenstein's Criterion)**.** *Suppose that $f(X) = a_n X^n + \cdots + a_1 X + a_0$ is a primitive polynomial in $\mathbb{Z}[X]$ and p is a prime in $\mathbb{Z}$ such that $p \mid a_i$ in $\mathbb{Z}$ for all $0 \leqslant i < n$; $p \nmid a_n$; and $p^2 \nmid a_0$ in $\mathbb{Z}$. Then f is irreducible in $\mathbb{Z}[X]$.*

Proof. Suppose that $f = gh$ for $g, h \in \mathbb{Z}[X]$. Write $\phi : \mathbb{Z}[X] \to \mathbb{F}_p[X]$ for the evaluation homomorphism at X (*i.e.* mapping X to X) extending the quotient map $\mathbb{Z} \to \mathbb{F}_p$. Then

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

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

Since $\phi(g)$ and $\phi(h)$ can be factorised into irreducibles, and $X \in \mathbb{F}_p[X]$ is prime it follows that $\phi(g) \sim X^i$ and $\phi(h) \sim X^{n-i}$ by Proposition 3.14. 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, so $i = n$. 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 leads to $g \sim 1$ □

Example 3.44. For $n \in \mathbb{N}^*$, the polynomial $X^n - 2$ 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).

Part II
Modules

4 Modules: an introduction

Suppose that R is a ring and M is a commutative group with operation $+$. A map $. : R \times M \to M; (r, x) \mapsto r.x$ is called a **scalar multiplication of R on M** if

(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 $x, y \in M$.

An **R-module** is a commutative group M, called the **additive group** of the module and whose operation is called **addition**, equipped with a scalar multiplication of R on M. We often speak of simply the module M if all other data is clear, and in this case R is the **ring of scalars of** M.

The elements of M are called **vectors** and the elements of R are called **scalars**. The identity of M is called the **zero** of the module and denoted 0, and for each $x \in M$ we write $-x$ for the unique inverse of x; the map $M \to M; x \mapsto -x$ is the **negation** of the module.

Another way of capturing the axioms (M1)–(M4) is to say that the map

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

is a well-defined ring homomorphism, where $\mathrm{Hom}(M, M)$ is the ring of group homomorphisms of the additive group of M. Indeed, to say that Ψ is well-defined, meaning it really does map into $\mathrm{Hom}(M, M)$, is exactly to say (M4); and to say that Ψ is a ring homomorphism is exactly (M1)–(M3).

Remark 4.1. Since M is a commutative group, $-0 = 0$, $-(-x) = x$ for all $x \in M$, and negation is a homomorphism.

(M4) says exactly that for $r \in R$ the map $M \to M; x \mapsto r.x$ is a group homomorphism of the additive group of M, so $r.0_M = 0_M$ and $r.(-x) = -(r.x)$ for all $x \in M$.

(M3) says exactly that for $x \in M$ the map $R \to M; r \mapsto r.x$ is a group homomorphism from the additive group of R to the additive group of M, so $0_R.x = 0_M$ and $(-r).x = -(r.x)$ for all $r \in R$.

Example 4.2 (Vector spaces as modules)**.** Given a field $\mathbb{F}$, a vector space V is exactly an $\mathbb{F}$-module, with the two notions of scalar multiplication coinciding.

Example 4.3 (The zero R-module)**.** For a 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**.

⚠ If R is trivial this is the *only* R-module, since $x = 1.x = (1+1).x = 1.x + 1.x = x + x$, so $x = 0$ for all $x \in M$.

Example 4.4 (The R-module R)**.** For a ring R, the multiplication map on R is also a scalar multiplication of the ring R on the additive group of R making R into an R-module which we call **the R-module R**.

(M1) is exactly the statement that 1_R is a left identity of ring multiplication; (M2) is exactly associativity of ring multiplication; (M3) is exactly that all right multiplication maps on a ring are homomorphisms of the additive group; and (M4) is exactly that all left multiplication maps on a ring are homomorphisms.

⚠ There may be more than one scalar multiplication of the ring R on the additive group of R: we saw in Example 1.26 that $\lambda.z := \lambda z$ and $\lambda.z := \overline{\lambda}.z$ are two different scalar multiplications of $\mathbb{C}$ on $\mathbb{C}$.

Example 4.5 (Direct sums)**.** Given R-modules $M_1, \ldots, M_n$, the product group $M_1 \times \cdots \times M_n$ equipped with the map $(r, x) \mapsto (r_1.x_1, \ldots, r_n.x_n)$ where the ith instance of $.$ is the scalar multiplication in M_i, is a scalar multiplication of R on $M_1 \times \cdots \times M_n$. This module is denoted $M_1 \oplus \cdots \oplus M_n$ and is called the **direct sum** of the R-modules M_1, …, M_n.

In particular the direct sum of n copies of the R-module R is called **the R-module R^n** and the scalar multiplication is given by $r.x = (rx_1, \ldots, rx_n)$.

The $\mathbb{F}$-module $\mathbb{F}^n$ is the usual vector space $\mathbb{F}^n$.

Example 4.6 (The $M_n(R)$-module R^n_{COL})**.** For a ring R, we write R^n_{COL} for $M_{n,1}(R)$. By Proposition 1.44, this is a commutative group and the map $M_n(R) \times R^n_{\text{COL}} \to R^n_{\text{COL}}; (A, v) \mapsto Av$ is a scalar multiplication of the ring $M_n(R)$ on R^n_{COL}. We call this **the $M_n(R)$-module R^n_{COL}**.

Example 4.7 (The R-module R^n_{COL})**.** For a ring R, the additive group R^n_{COL} has the structure of an R-module called **the R-module R^n_{COL}** with scalar multiplication

$$r.\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} r & & 0 \\ & \ddots & \\ 0 & & r \end{pmatrix}\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} rx_1 \\ \vdots \\ rx_n \end{pmatrix} \text{ for } r \in R, x \in R^n_{\text{COL}}.$$

The $\mathbb{F}$-module $\mathbb{F}^n_{\text{COL}}$ is the usual $\mathbb{F}$-vector space $\mathbb{F}^n_{\text{COL}}$.

Example 4.8 (The R-module $M_n(R)$)**.** For a ring R, the additive group $M_n(R)$ has the structure of an R-module called **the R-module $M_n(R)$** with scalar multiplication $(r.A)_{i,j} := rA_{i,j}$ for all $1 \leqslant i, j \leqslant n$.

For fields this is the vector space structure described in Example 1.48.

Example 4.9 (The $\mathbb{Z}$-module of a commutative group)**.** Given a commutative group M, the map $\mathbb{Z} \times M \to M$ defined by

$$(n-m).x := \overbrace{(x+\cdots+x)}^{n \text{ times}} - \overbrace{(x+\cdots+x)}^{m \text{ times}}$$

is a scalar multiplication of $\mathbb{Z}$ on M giving it the structure of a $\mathbb{Z}$-module called **the $\mathbb{Z}$-module** M.

Example 4.10 (Polynomial rings as R-modules)**.** The additive group of the ring $R[X]$ can be made into an $R[X]$-module – for example the $R[X]$-module $R[X]$ – but it can also be made into an R-module with scalar multiplication $r.(a_0+a_1X+\cdots+a_nX^n) = (ra_0)+(ra_1)X+\cdots+(ra_n)X^n$.

Example 4.11 (Modules over matrix rings as vector spaces)**.** An $M_n(\mathbb{F})$-module M is also vector space over $\mathbb{F}$ with scalar multiplication

$$\lambda.v := \begin{pmatrix} \lambda & & 0 \\ & \ddots & \\ 0 & & \lambda \end{pmatrix}.v \text{ for } \lambda \in \mathbb{F}, v \in M_n(\mathbb{F}).$$

Example 4.12 (Vector spaces with an endomorphism)**.** For $T : V \to V$ an $\mathbb{F}$-linear map we can define a scalar multiplication of $\mathbb{F}[X]$ on the additive group of V by

$$(a_0 + a_1X + \cdots + a_dX^d).v := a_0.v + a_1.T(v) + \cdots a_d.T^d(v) \text{ for all } p \in \mathbb{F}[X] \text{ and } v \in V$$

where the . on the right is the scalar multiplication of $\mathbb{F}$ on V resulting from the given vector space structure.

Linear maps

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 $\phi(r.x) = r.\phi(x)$ for all $x \in M$ and $r \in R$.

Remark 4.13. Since an R-linear map $\phi : M \to N$ is a group homomorphism, $\phi(0_M) = 0_N$ and $\phi(-x) = -\phi(x)$ for all $x \in M$.

Example 4.14 (Example 4.2, contd.)**.** For vector spaces V and W over a field $\mathbb{F}$, the linear maps $V \to W$ in the usual sense are exactly the $\mathbb{F}$-linear maps in the sense defined here.

Example 4.15. For an R-module M and elements $x_1, \ldots, x_n \in M$, the map $\Phi_x : R^n \to M; r \mapsto r_1.x_1 + \cdots + r_n.x_n$ is R-linear by (M2) and (M3).

Example 4.16. For $A \in M_{n,m}(R)$ the map $R^n \to R^m; v \mapsto vA$ between the R-modules R^n and R^m is an R-linear map between the R-modules R^n and R^m since $r.(vA) = r(vA) = (rv)A = (r.v)A$ for all $r \in R$ and $v \in R^n$, and $(v+w)A = vA + wA$ for all $v, w \in R^n$ by Proposition 1.44.

Example 4.17. For $A \in M_{n,m}(R)$ the map $R^m_{\text{COL}} \to R^n_{\text{COL}}; v \mapsto Av$ between the R-modules R^m_{COL} and R^n_{COL} is additive since $A(v+w) = Av + Aw$ by Proposition 1.44.

If R is commutative then (writing $\Delta(r)$ for the matrix with rs on the diagonal and 0 elsewhere as in Example 1.48) we have $\Delta(r)$ is in the centre of the ring $M_n(R)$ and so $A(r.v) = A(\Delta(r)v) = (A\Delta(r))v = (\Delta(r)A)v = \Delta(r)(Av) = r.(Av)$. Hence the map $R^m_{\text{COL}} \to R^n_{\text{COL}}; v \mapsto Av$ is R-linear.

If R is a non-commutative ring then there are elements $r, s \in R$ with $rs \neq sr$ and the map $R^1_{\text{COL}} \to R^1_{\text{COL}}; x \mapsto rx$ is *not* linear since $r(s.1) = rs \neq sr = s.(r1)$.

Example 4.18. For M an R-module and $x \in M$, the map $\phi : R \to M; r \mapsto r.x$ from the R-module R to M is R-linear since $(r+s).x = (r.x) + (s.x)$ by (M3) for M, and $(s.r).x = (sr).x = s.(r.x)$ for all $r, s \in R$ by definition of scalar multiplication in the R-module R and (M2) for M.

Example 4.19. If R is commutative then the scalar multiplication map $M \to M; x \mapsto s.x$ is R-linear since $s.(r.x) = (sr).x = (rs).x = r.(s.x)$ for all $r \in R$ and $x \in M$.

On the other hand, for any ring R if M is the zero-module then the map $M \to M; x \mapsto s.x$ is R-linear.

Proposition 4.20 (Algebra of linear maps)**.** *Suppose that M and N are R-modules. Then $L(M,N)$, the set of R-linear maps $M \to N$, is a subgroup of* $\text{Hom}(M,N)$ *(under pointwise addition). Furthermore, if $\phi \in L(M,N)$ and $\psi \in L(N,P)$ then $\psi \circ \phi \in L(M,P)$, and $L(M,M)$ is a subring of* $\text{Hom}(M,M)$*.*

Proof. Certainly $L(M,N)$ is a subset of $\text{Hom}(M,N)$, and the zero map $z : M \to N; x \mapsto 0_N$ is a homomorphism, and linear since $z(r.x) = 0_N = r.0_N = r.z(x)$ and so $L(M,N)$ is non-empty. If $\phi, \psi \in L(M,N)$ then $\phi - \psi$ is a homomorphism since $\text{Hom}(M,N)$ is a group, and $(\phi-\psi)(r.x) = \phi(r.x) - \psi(r.x) = r.\phi(x) - r.\phi(x) = r.(\phi(x) - \psi(x)) = r.((\phi-\psi)(x))$ so $\phi - \psi \in L(M,N)$ and hence $L(M,N)$ is a subgroup by the subgroup test.

For the second part, $\psi \circ \phi$ is a group homomorphism, and it is R-linear since $(\psi\circ\phi)(r.x) = \psi(\phi(r.x)) = \psi(r.\phi(x)) = r.\psi(\phi(x)) = r.(\psi\circ\phi)(x)$ for all $r \in R$ and $x \in M$ *i.e.* $\psi \circ \phi \in L(M,P)$ as claimed. Finally, the identity map $\iota : M \to M$ is R-linear since $\iota(r.x) = r.x = r.\iota(x)$ for all $r \in R$ and $x \in M$, so $1_{\text{Hom}(M,M)} \in L(M,M)$. $L(M,M)$ is closed under differences by the first part of the proposition, and is closed under products by what we just showed. By the subring test $L(M,M)$ is a subring of $\text{Hom}(M,M)$ as claimed. □

Remark 4.21. If R is a commutative ring then we can define a scalar multiplication on $L(M,N)$ by $(r.\phi)(x) \coloneqq \phi(r.x)$ giving it the structure of an R-module.

⚠ In Exercise III.10 there is an example of a ring R and R-module M such that the commutative group $L(M,M)$ cannot be given the structure of an R-module.

Isomorphisms and submodules

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

Observation 4.22. If $\phi : M \to N$ is an R-linear bijection then its inverse map is a group homomorphism, and $\phi^{-1}(\lambda.x) = \phi^{-1}(\lambda.\phi(\phi^{-1}(x))) = \phi^{-1}(\phi(\lambda.\phi^{-1}(x))) = \lambda.\phi^{-1}(x)$ so that ϕ is an R-linear isomorphism.

Example 4.23. The map

$$R^n \to R^n_{\mathrm{COL}}; r \mapsto \begin{pmatrix} r_1 \\ \vdots \\ r_n \end{pmatrix}$$

is an R-linear bijection between the R-module R^n and the R-module R^n_{COL}, and hence an R-linear isomorphism.

Example 4.24. The map $\mathbb{Q} \to \mathbb{Q}; x \mapsto 2x$ is a $\mathbb{Z}$-linear bijection from the $\mathbb{Z}$-module $\mathbb{Q}$ to itself arising via scalar multiplication as in Example 4.19. ⚠ The inverse map, while also $\mathbb{Z}$-linear does *not* arise via scalar multiplication when $\mathbb{Q}$ is considered as a $\mathbb{Z}$-module since there is no integer $z \in \mathbb{Z}$ such that $2zx = x$ for all $x \in \mathbb{Q}$.

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 $N \leqslant M$ and also say that N is **proper** if $M \neq N$.

Example 4.25 (Example 4.2, contd.)**.** When V is a vector space, a submodule of V is exactly a subspace of V.

Example 4.26 (Left ideals are submodules)**.** I is a left ideal in a ring R if and only if I is a submodule of the R-module R.

Example 4.27. The ideal $\langle 2 \rangle$ in the ring $\mathbb{Z}$ is a proper submodule of the $\mathbb{Z}$-module $\mathbb{Z}$ and it is $\mathbb{Z}$-linearly isomorphic to the $\mathbb{Z}$-module $\mathbb{Z}$ via $\mathbb{Z} \to \langle 2 \rangle; z \mapsto 2z$.

⚠ This is quite different from the situation with vector spaces: the only subspaces of the $\mathbb{F}$-vector space $\mathbb{F}$ are $\{0\}$ and $\mathbb{F}$.

Lemma 4.28 (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 addition on M restricted to N, is a subgroup of M by the subgroup test. Since $r.x \in N$ whenever $r \in R$ and $x \in N$, 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. □

As with rings, 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.

Quotients and the first isomorphism theorem

Theorem 4.29 (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 $q : M \to M/N; x \mapsto x + N$ is an R-linear surjection with kernel N.*

Proof. Since N is a commutative subgroup of M we have that M/N is a commutative group and the map q is a surjective homomorphism with kernel N by definition of the quotient group. Define a scalar multiplication of R on M/N by $r.(x + N) \coloneqq r.x + N$. This is well-defined: if $x + N = y + N$ then $x + n = y + n'$ for some $n, n' \in N$, so $r.x + r.n = r.y + r.n'$, but since N is a submodule $r.n, r.n' \in N$ and hence $r.x + N = r.y + N$ as required.

(M1) follows since $1.(x + N) = (1.x) + N = x + N$ for all $x \in M$ by (M1) for the scalar multiplication on M. (M2) follows since $r.(s.(x+N)) = r.(s.x+N) = (r.(s.x)) + N = (rs).x + N = (rs).(x+N)$ for all $r, s \in R$ and $x \in M$ by (M2) for the scalar multiplication on M. (M3) follows by (M3) for the scalar multiplication on M and the fact that q is a homomorphism so $(r+s).(x+N) = (r+s).x+N = ((r.x)+(s.x))+N = (r.x+N)+(s.x+N) = r.(x+N)+s.(x+N)$ for all $r, s \in R$ and $x \in M$. Finally, (M4) follows by (M4) for the scalar multiplication on M and the fact that q is a homomorphism so $r.((x + N) + (y + N)) = r.((x + y) + N) = r.(x + y) + N = ((r.x) + (r.y)) + N = (r.x + N) + (r.y + N)$ for all $r \in R$ and $x, y \in M$.

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

Remark 4.30. Since the map q above is a surjective R-linear map the scalar multiplication on M/N is determined by q: $r.(x + N) = r.x + N$ for all $x \in M$ and $r \in R$, where the first . is scalar multiplication in M/N, and the second in M.

By *the* R-module M/N we mean the module structure of this theorem.

Theorem 4.31 (first isomorphism theorem for modules)**.** *Suppose that $\phi : M \to N$ is R-linear. Then $\ker \phi$ is a submodule of M; $\operatorname{Im} \phi$ is a submodule of N; and the map*

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

is an injective R-linear map with image $\operatorname{Im}\phi$. *In particular,* $\operatorname{Im}\phi$ *is R-linearly isomorphic to* $M/\ker\phi$.

Proof. Both $\ker\phi$ and $\operatorname{Im}\phi$ are subgroups of the additive groups of M and N respectively by the first isomorphism theorem for groups since ϕ is, in particular, a group homomorphism. Therefore by the submodule test $\ker\phi$ and $\operatorname{Im}\phi$ are submodules since 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 $r.x = r.\phi(y) = \phi(r.y) \in \operatorname{Im}\phi$.

By Theorem 4.29 $M/\ker\phi$ is an R-module. $\widetilde{\phi}$ is an injective well-defined group homomorphism by the first isomorphism theorem for groups. It remains to check that it is linear which follows since $\widetilde{\phi}(r.(x+\ker\phi)) = \widetilde{\phi}((r.x)+\ker\phi) = \phi(r.x) = r.\phi(x) = r.\widetilde{\phi}(x+\ker\phi)$ for all $r \in R$ and $x \in M$. □

5 Free modules

Generation

For an R-module M and $\Lambda \subset M$ we write

$$\langle\Lambda\rangle := \{r_1.x_1 + \cdots + r_n.x_n : n \in \mathbb{N}_0, x_1, \ldots, x_n \in \Lambda, r_1, \ldots, r_n \in R\}.$$

This is a submodule of M by the submodule test, and we call this submodule the **module generated by** Λ. For $x_1, \ldots, x_n \in M$ we write

$$\langle x_1, \ldots, x_n\rangle := \{r_1.x_1 + \cdots + r_n.x_n : r_1, \ldots, r_n \in R\},$$

and since $0_R.x_i = 0_M$ we have that $\langle x_1, \ldots, x_n\rangle = \langle\{x_1, \ldots, x_n\}\rangle$.

Example 5.1. An R-module M is generated by the set M itself, and M is generated by the empty set if and only if it is the zero R-module.

Example 5.2 (Vector spaces, contd.)**.** For $\mathbb{F}$ a field, V an $\mathbb{F}$-module, and $\Lambda \subset V$ the submodule generated by Λ is the same as subspace spanned by Λ.

Example 5.3. Write $e_i := (0, \ldots, 0, 1, 0, \ldots, 0)$ for the elements of R^n with 1_R in the ith position and 0_R elsewhere. Similarly writer e_i^t for the column vector in R^n_{COL} with 1_R in the ith row and 0_Rs elsewhere.

$\{e_1, \ldots, e_n\}$ generates the R-module R^n since if $r \in R^n$ then $r = r_1.e_1 + \cdots + r_n.e_n$, and similarly $\{e_1^t, \ldots, e_n^t\}$ generates R^n_{COL}.

If there is a finite set Λ such that M is generated by Λ then we say that M is **finitely generated**. If M is generated by a set of size 1 we say that M is **cyclic**.

Example 5.4 (Commutative groups, contd.)**.** A commutative group M is cyclic if and only if the $\mathbb{Z}$-module M is cyclic.

For M a *finite* commutative group, the $\mathbb{Z}$-module M is finitely generated since it is generated by the finite set M.

Example 5.5. The $\mathbb{Z}$-module $\mathbb{Q}$ is *not* cyclic. Indeed, for any $q \in \mathbb{Q}^*$ there is no $z \in \mathbb{Z}$ such that $zq = q/2$, and since $\mathbb{Q} \neq \langle 0 \rangle$ the claim follows. ⚠ The $\mathbb{Q}$-module $\mathbb{Q}$ *is* cyclic and it is generated by any set $\{q\}$ with $q \in \mathbb{Q}^*$.

Example 5.6. For R a ring, the R-module R is cyclic – it is generated by 1 – and if K a submodule of the R-module R (equivalently K is a left ideal in the ring R), the quotient module R/K is cyclic – it is generated by $1 + K$.

In fact *every* cyclic R-module is isomorphic to a module of this form: if M is a cyclic R-module then the map $\phi : R \to M; r \mapsto r.x$ is surjective and R-linear, and so by the first isomorphism theorem there is a submodule K of the R-module R – in this case $\ker \phi$ – such that R/K is R-linearly isomorphic to M.

Observation 5.7. The R-linear image of an R-module generated by a set of size n is generated by a set of size n.

Proposition 5.8. *Suppose that $\phi : M \to N$ is an R-linear map and $\operatorname{Im}\phi$ and $\ker\phi$ are generated by sets of sizes n and m respectively. Then M is generated by a set of size $n+m$.*

Proof. Let $x_1, \ldots, x_n \in M$ be such that $\phi(x_1), \ldots, \phi(x_n)$ generate $\operatorname{Im}\phi$, and let $x_{n+1}, \ldots, x_{n+m}$ generate $\ker\phi$. Then if $x \in M$, there are elements $r_1, \ldots, r_n \in R$ such that $\phi(x) = r_1.\phi(x_1) + \cdots + r_n.\phi(x_n)$, and hence $\phi(x - r_1.x_1 - \cdots - r_n.x_n) = 0$ and so there are elements $r_{n+1}, \ldots, r_{n+m} \in R$ with $x - r_1.x_1 - \cdots - r_n.x_n = r_{n+1}.x_{n+1} + \cdots + r_{n+m}.x_{n+m}$. Rearranging the result is proved. □

Example 5.9 (Vector spaces, contd.)**.** The proof above is modelled on a proof of the Rank-Nullity theorem, and in fact since a basis for a vector space is certainly a spanning and so generating set, it tells us that if V is a vector space and $T : V \to W$ is linear with finite rank and nullity then $\dim V \leqslant \mathrm{rk}(T) + \mathrm{n}(T)$. The Rank-Nullity theorem is the stronger claim that we have equality here.

In an R-module M, we say $\mathcal{E} \subset M$ is a **minimal generating set** if $\mathcal{E}$ generates M and no proper subset of $\mathcal{E}$ generates M.

Observation 5.10. A *finite* generating set for an R-module M contains a minimal generating set by induction.

Example 5.11. ⚠ Minimal generating sets need not exist: Exercise III.1 asks for a proof that the $\mathbb{Z}$-module $\mathbb{Q}$ does not have a minimal generating set. In particular, in view of the preceding observation, the $\mathbb{Z}$-module $\mathbb{Q}$ is *not* finitely generated.

Example 5.12. The set $\{2,3\}$ is a generating set for the $\mathbb{Z}$-module $\mathbb{Z}$, and no proper subset is generating so it is a minimal generating set. ⚠ There are smaller generating sets of $\mathbb{Z}$ – $\{1\}$ and $\{-1\}$.

Proposition 5.13. *Suppose that M is a finitely generated R-module. Then every generating set for M contains a finite subset that is also a generating set. In particular, every minimal generating set is finite.*

Proof. Let $\{x_1,\dots,x_n\}$ generate M and suppose that $\mathcal{E}$ is a generating set for M. For each $1 \leqslant i \leqslant n$ there is a finite subset $S_i \subset \mathcal{E}$ such that $x_i \in \langle S_i\rangle$, and hence $x_1,\dots,x_n \in \langle \bigcup_{i=1}^n S_i\rangle$. Since $\{x_1,\dots,x_n\}$ generates M we have $M = \langle x_1,\dots,x_n\rangle \subset \langle\langle \bigcup_{i=1}^n S_i\rangle\rangle = \langle \bigcup_{i=1}^n S_i\rangle$. However, $\bigcup_{i=1}^n S_i \subset \mathcal{E}$, and a finite union of finite sets is finite as required. □

Linear independence

For an R-module M we say that a finite sequence $x_1,\dots,x_n \in M$ is **(R-)linearly independent** if whenever $r_1,\dots,r_n \in R$ have $r_1.x_1 + \cdots + r_n.x_n = 0_M$ we have $r_1,\dots,r_n = 0_R$. A set Λ is **(R-)linearly independent** if for every $n \in \mathbb{N}_0$, $x_1,\dots,x_n$ is R-linearly independent for every sequence of *distinct* $x_1,\dots,x_n \in \Lambda$.

Sets and sequences are **(R-)linearly dependent** if they are not R-linearly independent.

Example 5.14. In an R-module M the empty set or empty sequence is R-linearly independent.

Example 5.15 (Vector spaces, contd.)**.** A subset of a vector space is linearly independent in the usual sense if and only if it is linearly independent in the sense here.

Example 5.16 (Example 5.3, cont.)**.** $e_1,\dots,e_n$ are R-linearly independent in R^n: if $r_1.e_1 + \cdots + r_n.e_n = 0$ for $r_1,\dots,r_n \in R$ then $r_1,\dots,r_n = 0$, and similarly for $e_1^t,\dots,e_n^t$ in R^n_{COL}.

Example 5.17 (Commutative groups, contd.)**.** If M is a finite commutative group then by Lagrange's theorem $|M|.x = 0$ for all x in the $\mathbb{Z}$-module M, and hence there are no non-empty $\mathbb{Z}$-linearly independent subsets of M.

Example 5.18. The $\mathbb{Z}$-module $\mathbb{Q}$ has no $\mathbb{Z}$-linearly independent subset of size 2. Indeed, suppose that $e_1, e_2 \in \mathbb{Q}$ were a $\mathbb{Z}$-linearly independent sequence with $e_2 \neq 0$. There is $z \in \mathbb{Z}^*$ such that $ze_1, ze_2 \in \mathbb{Z}$, and hence $(ze_2).e_1 + (-ze_1).e_2 = 0$ but $ze_2 \neq 0$ so e_1, e_2 is $\mathbb{Z}$-linearly dependent – a contradiction.

Bases

For an R-module M we say that $\mathcal{E}$ is a **basis** for M if it is a linearly independent generating set for M. A module with a basis is called **free**.

Example 5.19. The zero R-module has the empty set as a basis and so is free.

⚠ If R is trivial then $\{0\}$ is also a basis since it is R-linearly independent.

Example 5.20 (Commutative groups, contd.)**.** If M is a non-trivial finite commutative group then the $\mathbb{Z}$-module M is *not* free since the only independent sets are empty and the module generated by the empty set has only one element: zero.

Example 5.21. The $\mathbb{Z}$-module $\mathbb{Q}$ is *not* free: If it were it would have a basis $\mathcal{E}$. If $\mathcal{E}$ had more than one element then it would contain two linearly independent elements contradicting the conclusion of Example 5.18; if it had strictly fewer than two elements then $\mathbb{Q}$ would be cyclic contradicting the conclusion of Example 5.5.

Example 5.22 (Example 5.3, contd.)**.** In view of Examples 5.3 & 5.16, $\{e_1, \ldots, e_n\}$ is a basis for the R-module R^n and $\{e_1^t, \ldots, e_n^t\}$ is a basis for the R-module R^n_{COL} – these are both free modules.

Example 5.23 (Example 4.15, contd.)**.** If $\{x_1, \ldots, x_n\}$ is a basis for the R-module M then the linear map $\Phi_x : R^n \to M; r \mapsto r_1.x_1 + \cdots + r_n.x_n$ is injective since $\{x_1, \ldots, x_n\}$ is linearly independent, and surjective since $\{x_1, \ldots, x_n\}$ is a generating set, hence Φ_x is an R-linear isomorphism.

Proposition 5.24. *Suppose that M is an R-module with a basis $\mathcal{E}$. Then $\mathcal{E}$ is a minimal generating set. In particular, if M is finitely generated then $\mathcal{E}$ is finite.*

Proof. Suppose that $\mathcal{E}' \subset \mathcal{E}$ generates M and $e \in \mathcal{E} \setminus \mathcal{E}'$. Since $\mathcal{E}'$ generates M, $1.e = e \in \langle \mathcal{E}' \rangle$ and so $\{e\} \cup \mathcal{E}'$ is linearly dependent. But this is contained in $\mathcal{E}$ which is linearly independent and linear independence is preserved under passing to subsets. This contradiction establishes the first claim. The last part follows by Proposition 5.13. □

Example 5.25. The set $\{2, 3\}$ is a minimal generating set for the $\mathbb{Z}$-module $\mathbb{Z}$, but it is not linearly independent and so not a basis.

Example 5.26 (Vector spaces, contd.)**.** A minimal generating set in a vector space is linearly independent and so a basis. In particular, any finitely generated vector space has a minimal generating set and so has a finite basis. In other words, every finitely generated vector space has a basis, by contrast with the case of more general modules (Example 5.20).

Remark 5.27. ⚠ In a vector space any two finite bases have the same size – this is sometimes called the Dimension theorem. For more general rings, finite bases of modules over those rings need not have the same size: the zero module over the trivial ring has $\varnothing$ and $\{0\}$ as bases of sizes 0 and 1 respectively; Exercise III.9 gives an example of a non-trivial ring and a module over that ring with bases of sizes 1 and 2.

Presentations

A quotient of a finitely generated module is finitely generated, but the same is not true of submodules:

Example 5.28. In Exercise II.4 we saw that $\overline{\mathbb{Z}}$ contains an ideal that is not finitely generated, and this ideal is therefore a submodule of the cyclic $\overline{\mathbb{Z}}$-module $\overline{\mathbb{Z}}$ that is not finitely generated.

A matrix $A \in M_n(R)$ is said to be **upper triangular** if $A_{i,j} = 0$ whenever $j < i$.

Proposition 5.29. *Suppose that R is a PID and $M \leqslant R^n$. Then there is an upper triangular $A \in M_n(R)$ such that $M = R^n A$.*

Proof. For each $1 \leqslant i \leqslant n$ the set $M_i := \{x_i \in R : x \in M \text{ and } x_1, \dots, x_{i-1} = 0\}$ is a submodule of the R-module R and since R is a PID every such submodule is an ideal and generated by an element of R. Let $A \in M_n(R)$ be such that the ith row of A is $(0, \dots, 0, A_{i,i}, \dots, A_{i,n}) \in M$ where $A_{i,i}$ generates M_i.

By design A is upper triangular and every row of A is in M, so any linear combination of rows of A is in M – in other words $R^n A \subset M$. In the other direction, suppose that $M \setminus R^n A$ is non-empty. The zero vector is not in this set, and so if it is non-empty then it contains an element x with $i \leqslant n$ maximal such that $x_1, \dots, x_{i-1} = 0$. By design $A_{i,i} \mid x_i$, say $x_i = zA_{i,i}$. Then $x' := x - (0, \dots, 0, z, 0, \dots, 0)A \in M \setminus R^n A$, but $x'_1, \dots, x'_i = 0$ contradicting the maximality of i. □

Remark 5.30. Being free and finitely generated are properties that are preserved by isomorphisms so in particular, if M is a submodule of a free and finitely generated module over a PID then it is finitely generated.

Example 5.31 (Vector spaces, contd.)**.** For V a subspace of $\mathbb{F}^n$ the above tells us that V is generated by at most n vectors since any field is a PID and so $\dim W \leqslant n$.

An R-module M has a **finite presentation with presentation matrix** $A \in M_{m,n}(R)$ if there is an R-linear isomorphism $\Phi : R^n / R^m A \to M$.

Example 5.32. For M an R-module with basis $x_1, \dots, x_n$, the linear map $R^n \to M; r \mapsto r_1.x_1 + \cdots + r_n.x_n$ is an R-linear isomorphism. For any $m \in \mathbb{N}_0$, we have $R^m 0_{m\times n} = \{0_{R^n}\}$ and hence by the first isomorphism theorem M has a finite presentation with presentation matrix $0_{m\times n}$.

Observation 5.33. A module with a finite presentation is finitely generated. On the other hand, Exercise III.6 gives an example of a finitely generated module that does *not* have a finite presentation.

Example 5.34. For R a PID and M an R-module generated by $x_1, \ldots, x_n$ there is an R-linear surjection $R^n \to M; r \mapsto r_1.x_1 + \cdots + r_n.x_n$. By Proposition 5.29 the kernel of this map is $R^n A$ for some upper triangular $A \in M_n(R)$, and hence by the first isomorphism theorem M has a finite presentation with presentation matrix A.

6 Elementary operations and the Smith normal form

There are three types of **elementary column (resp. row) operation** that can be applied to matrices in $M_{n,m}(R)$ – transvections, dilations, and interchanges – and these correspond to right (resp. left) multiplication by matrices from $M_m(R)$ and $M_n(R)$ respectively.

Write $E_n(i,j)$ for the matrix in $M_n(R)$ with 0_Rs everywhere except for row i and column j where the entry is 1_R. Then $E_n(i,j)E_n(k,l) = E_n(i,l)$ if $j = k$ and $E_n(i,j)E_n(k,l) = 0_{n\times n}$ if $j \neq k$.

Transvections

For $1 \leqslant i, j \leqslant m$ with $i \neq j$ and $\lambda \in R$ put $T_m(i,j;\lambda) = I_m + \lambda.E_m(i,j)$ (where . is the scalar multiplication of the R-module $M_m(R)$) so that

$$T_m(i,j;\lambda)T_m(i,j;-\lambda) = I_m = T_m(i,j;-\lambda)T_m(i,j;\lambda).$$

Given $A \in M_{n,m}(R)$, the matrix $AT_m(i,j;\lambda)$ is the matrix A with the ith column times λ added to the jth column; we write this

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

Similarly the matrix $T_n(i,j;\lambda)A$ is the matrix A with λ times the jth row added to the ith row; we write this

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

Dilations

For $1 \leqslant i \leqslant m$ and $u \in U(R)$ let $D_m(i;u) = I_m + (u-1).E_m(i,i)$ so that

$$D_m(i;u)D_m(i;u^{-1}) = I_m = D_m(i;u^{-1})D_m(i;u).$$

The matrix $AD_m(i;u)$ is the matrix with the ith column replaced by the ith column times u and as above we write this and the corresponding row operation as

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

Interchanges

For $1 \leqslant i, j \leqslant m$ let $S_m(i,j) = I_m + E_m(i,j) + E_m(j,i) - E_m(i,i) - E_m(j,j)$ so that $S_m(i,j)^2 = I_m$. The matrix $AS_m(i,j)$ is the matrix A with columns i and j swapped and as above we write

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

for this and the corresponding row operation.

Remark 6.1. We write $\mathrm{GL}_n(R)$ for the group $U(M_n(R))$, and $\mathrm{GE}_n(R)$ for the subgroup of $\mathrm{GL}_n(R)$ generated by the transvections, dilations, and interchanges.

In general $\mathrm{GL}_2(R) \neq \mathrm{GE}_2(R)$, though this can be hard to show. An example, taken from [Coh66, p23], is the ring $\mathbb{Z}[\theta]$ where $\theta^2 - \theta + 5 = 0$. Here the matrix

$$A := \begin{pmatrix} 3-\theta & 2+\theta \\ -3-2\theta & 5-2\theta \end{pmatrix}$$

is in $\mathrm{GL}_2(\mathbb{Z}[\theta])$ but not in $\mathrm{GE}_2(\mathbb{Z}[\theta])$.

We say that $A, B \in M_{n,m}(R)$ are **equivalent by elementary operations** and write $A \sim_{\mathcal{E}} B$ if there is a sequence $A =: A_0 \to A_1 \to \cdots \to A_{k-1} \to A_k := B$ such that A_{i+1} is the result of an elementary row or column operation applied to A_i for all $0 \leqslant i < k$.

We say that $A, B \in M_{n,m}(R)$ are **equivalent** and write $A \sim B$ if there are matrices $S \in \mathrm{GL}_n(R)$ and $T \in \mathrm{GL}_m(R)$ such that $A = SBT$.

Observation 6.2. Both $\sim_{\mathcal{E}}$ and $\sim$ are equivalence relations, and in view of the definition of $\mathrm{GE}_n(R)$ we have $A \sim_{\mathcal{E}} B$ if and only if there is $P, Q \in \mathrm{GE}_n(R)$ such that $A = PBQ$, so that $A \sim_{\mathcal{E}} B$ implies $A \sim B$.

Example 6.3. For $A \in M_{n,m}(R)$ write $r_1, \ldots, r_n$ for its rows, and $c_1, \ldots, c_m$ for its columns. For any $\sigma \in S_n$ and $\tau \in S_m$ we have

$$\begin{pmatrix} r_1 \\ \vdots \\ r_n \end{pmatrix} \sim_{\mathcal{E}} \begin{pmatrix} r_{\sigma(1)} \\ \vdots \\ r_{\sigma(n)} \end{pmatrix} \text{ and } \begin{pmatrix} c_1 & \cdots & c_m \end{pmatrix} \sim_{\mathcal{E}} \begin{pmatrix} c_{\tau(1)} & \cdots & c_{\tau(m)} \end{pmatrix},$$

since σ (resp. τ) is generated by transpositions, and interchanging rows (resp. columns) i and j corresponds to apply the transposition (ij) to the row (resp. column) indices.

We say that $A \in M_{n,m}(R)$ is **diagonal** if $A_{i,j} = 0$ whenever $i \neq j$. ⚠ In particular, we do *not* insist that that A be square.

Example 6.4. If $A \in M_{n,m}(R)$ is diagonal, interchanging rows i and j and columns i and j gives the matrix A with $A_{i,i}$ and $A_{j,j}$ interchanged. Hence for any $\sigma \in S_{\min\{n,m\}}$ we have

$$A \sim_{\mathcal{E}} \begin{pmatrix} A_{\sigma(1),\sigma(1)} & 0 & \cdots \\ 0 & A_{\sigma(2),\sigma(2)} & \ddots \\ \vdots & \ddots & \ddots \end{pmatrix}.$$

Example 6.5. Two matrices $A, B \in M_n(\mathbb{F})$ are said to be **similar** if there is $P \in \mathrm{GL}_n(\mathbb{F})$ such that $A = P^{-1}BP$, and so if A and B are similar then $A \sim B$. However,

$$A := \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \xrightarrow{c_1 \leftrightarrow c_2} \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} =: B,$$

so that here $A \sim_{\mathcal{E}} B$, but B is diagonal and A is not similar to a diagonal matrix *i.e.* it is not diagonalisable.

Theorem 6.6. *Suppose that R is a Euclidean domain. Then every $A \in M_{n,m}(R)$ is equivalent by elementary operations to a diagonal matrix.*

Proof. Let $\mathcal{A}_k$ be those matrices $B \sim_{\mathcal{E}} A$ with the additional property that whenever $i < k$ and $j \neq i$, or $j < k$ and $i \neq j$, we have $B_{i,j} = 0$. We shall show by induction that $\mathcal{A}_k$ is non-empty for $k \leqslant \min\{m, n\} + 1$; $\mathcal{A}_1$ contains A and so is certainly non-empty.

Let f be a Euclidean function for R, and suppose that $\mathcal{A}_k \neq \varnothing$ and $k \leqslant \min\{m, n\}$. Let $B \in \mathcal{A}_k$ be a matrix with $f(B_{k,k})$ minimal (with the convention that $f(0) = \infty$). First we show that $B_{k,k} \mid B_{k,i}$ for all $i > k$ (including in the case $B_{k,k} = 0$): if not, there is some $i > k$ with $B_{k,i} = qB_{k,k} + r$ and $f(r) < f(B_{k,k})$, so we apply the elementary operations

$$B = \begin{pmatrix} B_{1,1} & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots & & \vdots & & \vdots \\ \vdots & \ddots & \ddots & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \cdots & 0 & B_{k,k} & \cdots & B_{k,i} & \cdots & B_{k,m} \\ \vdots & & \vdots & \vdots & & \vdots & & \vdots \\ 0 & \cdots & 0 & B_{n,k} & \cdots & B_{n,i} & \cdots & B_{n,m} \end{pmatrix}$$

$$\xrightarrow{c_i \mapsto c_i - c_k q} \begin{pmatrix} B_{1,1} & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots & & \vdots & & \vdots \\ \vdots & \ddots & \ddots & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \cdots & 0 & B_{k,k} & \cdots & B_{k,i} - B_{k,k}q & \cdots & B_{k,m} \\ \vdots & & \vdots & \vdots & & \vdots & & \vdots \\ 0 & \cdots & 0 & B_{n,k} & \cdots & B_{n,i} - B_{n,k}q & \cdots & B_{n,m} \end{pmatrix}$$

$$\xrightarrow{c_k \leftrightarrow c_i} \begin{pmatrix} B_{1,1} & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots & & \vdots & & \vdots \\ \vdots & \ddots & \ddots & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \cdots & 0 & B_{k,i} - B_{k,k}q & \cdots & B_{k,k} & \cdots & B_{k,m} \\ \vdots & & \vdots & \vdots & & \vdots & & \vdots \\ 0 & \cdots & 0 & B_{n,i} - B_{n,k}q & \cdots & B_{n,k} & \cdots & B_{n,m} \end{pmatrix} =: B'.$$

Then $B' \in \mathcal{A}_k$ has $B'_{k,k} = B_{k,i} - qB_{k,k} = r$, but $f(B'_{k,k}) = f(r) < f(B_{k,k})$ which contradicts the minimality in our choice of B. Similarly, but with row operations in place of column operations, $B_{k,k} \mid B_{i,k}$ for all $i > k$.

For $k < i \leqslant m$ let q_i be such that $B_{k,i} = B_{k,k}q_i$. Apply elementary column operations

$$B \xrightarrow{c_{k+1} \mapsto c_{k+1} - c_k q_{k+1}} \begin{pmatrix} B_{1,1} & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots & \vdots & \vdots & & \vdots \\ \vdots & \ddots & \ddots & 0 & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & B_{k,k} & 0 & B_{k,k+2} & \cdots & B_{k,m} \\ 0 & \cdots & 0 & B_{k+1,k} & B_{k+1,k+1} - B_{k+1,k}q_{k+1} & B_{k+1,k+2} & \cdots & B_{k+1,m} \\ \vdots & & \vdots & \vdots & \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & B_{n,k} & B_{n,k+1} - B_{n,k}q_{k+1} & B_{n,k+2} & \cdots & B_{n,m} \end{pmatrix}$$

$$\vdots$$

$$\xrightarrow{c_m \mapsto c_m - c_k q_m} \begin{pmatrix} B_{1,1} & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots & \vdots & & \vdots \\ \vdots & \ddots & \ddots & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & B_{k,k} & 0 & \cdots & 0 \\ 0 & \cdots & 0 & B_{k+1,k} & B_{k+1,k+1} - B_{k+1,k}q_{k+1} & \cdots & B_{k+1,m} - B_{k+1,k}q_m \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & B_{n,k} & B_{n,k+1} - B_{n,k}q_{k+1} & \cdots & B_{n,m} - B_{n,k}q_m \end{pmatrix} =: B'.$$

For $k < i \leqslant n$ let p_i be such that $B_{i,k} = p_i B_{k,k}$. Apply elementary row operations

$$B' \xrightarrow{r_{k+1} \mapsto r_{k+1} - p_{k+1} r_k} \begin{pmatrix} B_{1,1} & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots & \vdots & & \vdots \\ \vdots & \ddots & \ddots & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & B_{k,k} & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & B'_{k+1,k+1} & \cdots & B'_{k+1,m} \\ 0 & \cdots & 0 & B_{k+1,k} & B'_{k+2,k+1} & \cdots & B'_{k+2,m} \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & B_{n,k} & B'_{n,k+1} & \cdots & B'_{n,m} \end{pmatrix}$$

$$\cdots \xrightarrow{r_n \mapsto r_n - p_n r_k} \begin{pmatrix} B_{1,1} & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots & \vdots & & \vdots \\ \vdots & \ddots & \ddots & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & B_{k,k} & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & B'_{k+1,k+1} & \cdots & B'_{k+1,m} \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & 0 & B'_{n,k+1} & \cdots & B'_{n,m} \end{pmatrix} =: B''.$$

Then $B'' \sim_{\mathcal{E}} B' \sim_{\mathcal{E}} B \sim_{\mathcal{E}} A$ and $B'' \in \mathcal{A}_{k+1}$. The inductive step is complete. It follows that $\mathcal{A}_{\min\{m,n\}+1} \neq \varnothing$; any B in this set is diagonal and equivalent to A. □

For $d_1, \ldots, d_n \in \mathbb{N}_0$ and $B_1 \in M_{d_1}(R), \ldots, B_n \in M_{d_n}(R)$ we write

$$B_1 \oplus \cdots \oplus B_n := \begin{pmatrix} B_1 & 0_{d_1 \times d_2} & \cdots & 0_{d_1 \times d_n} \\ 0_{d_2 \times d_1} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0_{d_{n-1} \times d_n} \\ 0_{d_n \times d_1} & \cdots & 0_{d_n \times d_{n-1}} & B_n \end{pmatrix}.$$

We call the B_is the **blocks** of the matrix $B_1 \oplus \cdots \oplus B_n$, and it will be useful to allow 'degenerate' 0×0 blocks.

Example 6.7. A matrix $A \in M_n(R)$ is diagonal with entries $d_1, \ldots, d_n$ if $A = (d_1) \oplus \cdots \oplus (d_n)$.

Example 6.8. If $d_1 + \cdots + d_n = n$ then $I_n = I_{d_1} \oplus \cdots \oplus I_{d_n}$. ⚠ This is *not* a special case of the previous example because we are allowing 0×0 blocks.

Observation 6.9. If $B_i \sim B'_i$ (resp. $B_i \sim_{\mathcal{E}} B'_i$) for $1 \leqslant i \leqslant n$ then $B_1 \oplus \cdots \oplus B_n \sim B'_1 \oplus \cdots \oplus B'_n$ (resp. $B_1 \oplus \cdots \oplus B_n \sim_{\mathcal{E}} B'_1 \oplus \cdots \oplus B'_n$).

Lemma 6.10. *Suppose that R is an integral domain, $k < l$ and $A \in M_{n,m}(R)$ is diagonal with $\langle d \rangle = \langle A_{k,k} \rangle + \langle A_{l,l} \rangle$ for some $d \neq 0$, and $md = A_{l,l} A_{k,k}$. Then A is equivalent by elementary operations to the matrix A with $A_{k,k}$ replaced by m and $A_{l,l}$ replaced by d.*

Proof. Let $\alpha, \beta, p, q \in R$ be such that $A_{k,k}\alpha + \beta A_{l,l} = d$ and $A_{k,k} = dp$, $A_{l,l} = qd$, and so $m = qdp$. Then

$$A \xrightarrow{c_l \mapsto c_l + c_k \alpha} \begin{pmatrix} \ddots & & & & \\ & A_{k,k} & & A_{k,k}\alpha & \\ & & \ddots & & \\ & & & A_{l,l} & \\ & & & & \ddots \end{pmatrix} \xrightarrow{r_k \mapsto r_k + \beta r_l} \begin{pmatrix} \ddots & & & & \\ & A_{k,k} & & d & \\ & & \ddots & & \\ & & & A_{l,l} & \\ & & & & \ddots \end{pmatrix} \xrightarrow{c_k \mapsto c_k - c_l p}$$

$$\begin{pmatrix} \ddots & & & & \\ & 0 & & d & \\ & & \ddots & & \\ & -m & & A_{l,l} & \\ & & & & \ddots \end{pmatrix} \xrightarrow{r_l \mapsto r_l - q r_k} \begin{pmatrix} \ddots & & & & \\ & 0 & & d & \\ & & \ddots & & \\ & -m & & 0 & \\ & & & & \ddots \end{pmatrix} \xrightarrow{r_l \leftrightarrow r_k}$$

$$\begin{pmatrix} \ddots & & & & \\ & -m & & & \\ & & \ddots & & \\ & & & d & \\ & & & & \ddots \end{pmatrix} \xrightarrow{r_k \leftrightarrow (-1) r_k} \begin{pmatrix} \ddots & & & & \\ & m & & & \\ & & \ddots & & \\ & & & d & \\ & & & & \ddots \end{pmatrix}.$$

The result is proved. □

We say that $A \in M_{n,m}(R)$ is in **Smith normal form over** R if it is diagonal *and* $A_{i,i} \mid A_{i+1,i+1}$ for all $1 \leqslant i < \min\{n, m\}$.

Proposition 6.11. *Suppose that R is a Bezout domain. Then every diagonal matrix $A \in M_{n,m}(R)$ is equivalent by elementary operations to a matrix in Smith normal form.*

Proof. Let $\mathcal{A}_k$ be the set of diagonal matrices that are elementarily equivalent to A, and such that if the diagonal entries are denoted $a_1, a_2, \ldots, a_{\min\{m,n\}}$, then $a_i \mid a_j$ whenever $1 \leqslant i \leqslant j$ and $i \leqslant k$. Certainly $A \in \mathcal{A}_0$ since the hypotheses on the entries is vacuous then, so there is a maximal $k \in \mathbb{N}^*$ with $k - 1 \leqslant \min\{m, n\}$ such that $\mathcal{A}_{k-1}$ is non-empty.

By maximality of k for each matrix in $\mathcal{A}_{k-1}$ with diagonal entries $a_1, a_2, \ldots, a_{\min\{m,n\}}$ there is a minimal $l > k$ with $a_k \nmid a_l$; let $B \in \mathcal{A}_{k-1}$ have l maximal with this property. By Lemma 6.10 and Proposition 3.22 we can replace a_k and a_l by the greatest common divisor and least common multiple respectively of a_k and a_l, to get a matrix C that is equivalent to B by elementary operations.

Write $a'_1, \ldots, a'_{\min\{m,n\}}$ for the diagonal entries of C, so that for $i \notin \{k, l\}$ we have $a'_i = a_i$. a'_k and a'_l are linear combinations of a_k and a_l and so for $i \leqslant k-1$, a'_i divides them both, and hence for $1 \leqslant i \leqslant j$ we have we have $a'_i \mid a'_j$. It follows that $C \in \mathcal{A}_{k-1}$. Finally $a'_k \mid a_k$ and so $a'_k \mid a'_j$ for $k \leqslant j < l$, but also $a'_k \mid a'_l$ contradicting maximality of l. The result is proved. □

Theorem 6.12. *Suppose that R is a Euclidean domain. Then every $A \in M_{n,m}(R)$ is equivalent by elementary operations to a matrix in Smith normal form.*

Proof. This follows from Theorem 6.6 and Proposition 6.11. □

Remark 6.13. Following the work of Kaplanksy [Kap49] an integral domain R for which every $A \in M_{n,m}(R)$ is equivalent to a matrix in Smith normal form, is called an **elementary**

divisor domain, so in this language Theorem 6.12 shows that every Euclidean domain is an elementary divisor domain.

In the other direction Kaplansky showed [LLS74, Theorem 3.1] that every elementary divisor domain is a Bezout domain, and it is an open problem [Lor12] (going back at least to [Hel43]) to give an example of a Bezout domain that that is not an elementary divisor domain.

7 Applications of Smith normal form

With these tools we are in a position to describe the structure of finitely generated modules over a Euclidean domain:

Theorem 7.1. *Suppose that R is a Euclidean domain and M is generated by $x_1, \dots, x_n$. Then there are elements $a_1 \mid a_2 \mid \cdots \mid a_n$ in R and a matrix $Q \in \mathrm{GL}_n(R)$ such that*

$$\begin{array}{ccc} (R/\langle a_1\rangle) \oplus \cdots \oplus (R/\langle a_n\rangle) & \to & M \\ (r_1 + \langle a_1\rangle, \dots, r_n + \langle a_n\rangle) & \mapsto & (rQ).x_1 + \cdots + (rQ).x_n \end{array}$$

is a well-defined R-linear isomorphism.

Proof. The map

$$\Phi_x : R^n \to M; r \mapsto r_1.x_1 + \cdots + r_n.x_n$$

is an R-linear surjection. Since R is a Euclidean domain it is a PID and hence by Proposition 5.29 there is $A \in M_n(R)$ such that the kernel of this map is R^nA. By Theorem 6.12 there is a diagonal matrix $B \in M_n(R)$ with entries $a_1 \mid \cdots \mid a_n$ and $P, Q \in \mathrm{GL}_n(R)$ such that $A = PBQ$. The map

$$R^n \to M; r \mapsto (rQ).x_1 + \cdots + (rQ).x_n$$

is an R-linear map which is surjective because Q is invertible and Φ_x is surjective. The kernel is the set of $r \in R^n$ for which $rQ \in \ker \Phi_x$ *i.e.* for which there is $r' \in R^n$ such that $rQ = r'A$. This is true if and only if $r = (r'P)(P^{-1}AQ^{-1}) = (r'P)B$. Since P is invertible r is in the kernel if and only if $r \in R^nB = \langle a_1\rangle \oplus \cdots \oplus \langle a_n\rangle$. Finally, the composition of maps

$$\begin{array}{ccccc} (R/\langle a_1\rangle) \oplus \cdots \oplus (R/\langle a_n\rangle) & \to & R^n/R^nB & \to & M \\ (r_1 + \langle a_1\rangle, \dots, r_n + \langle a_n\rangle) & \mapsto & r + \langle a_1\rangle \oplus \cdots \oplus \langle a_n\rangle & \mapsto & (rQ).x_1 + \cdots + (rQ).x_n \end{array}$$

is a composition of well-defined R-linear isomorphisms by the first isomorphism theorem. □

This in turn lets us describe the structure of finitely generated commutative groups:

Corollary 7.2. *Suppose that G is a commutative group generated by $x_1, \dots, x_n$. Then there are natural numbers $d_1 \mid d_2 \mid \cdots \mid d_n$ (which may be 0) such that G is isomorphic to $\mathbb{Z}/\langle d_1\rangle \oplus \cdots \oplus \mathbb{Z}/\langle d_n\rangle$.*

Proof. This is a corollary of Theorem 7.1 since a commutative group is a $\mathbb{Z}$-module, and $\mathbb{Z}$ is a Euclidean domain. We may ensure the d_is are natural numbers by multiplying by a unit in $\mathbb{Z}$ as necessary. □

Matrix forms

In this section we work with matrices multiplying columns on the left rather than rows on the right. Equivalent matrices induce isomorphisms in the same way as in the proof of Theorem 7.1:

Proposition 7.3. *Suppose that $A, B \in M_n(\mathbb{F}[X])$, and $P, Q \in \mathrm{GL}_n(\mathbb{F}[X])$ are such that $A = PBQ$. Then the map*

$$\mathbb{F}[X]^n_{\mathrm{COL}}/A\mathbb{F}[X]^n_{\mathrm{COL}} \to \mathbb{F}[X]^n_{\mathrm{COL}}/B\mathbb{F}[X]^m_{\mathrm{COL}}; x + A\mathbb{F}[X]^n_{\mathrm{COL}} \mapsto P^{-1}x + B\mathbb{F}[X]^n_{\mathrm{COL}}$$

is a well-defined $\mathbb{F}[X]$-linear isomorphism.

Proof. Since $\mathbb{F}[X]$ is commutative $B\mathbb{F}[X]^n_{\mathrm{COL}}$ is an $\mathbb{F}[X]$-module, and hence $\mathbb{F}[X]^n_{\mathrm{COL}} \to \mathbb{F}[X]^n_{\mathrm{COL}}/B\mathbb{F}[X]^n_{\mathrm{COL}}; x \mapsto P^{-1}x + B\mathbb{F}[X]^n_{\mathrm{COL}}$ is a well-defined $\mathbb{F}[X]$-linear surjection. It has kernel $A\mathbb{F}[X]^n_{\mathrm{COL}}$, since $P^{-1}x \in B\mathbb{F}[X]^n_{\mathrm{COL}}$ if and only if $P^{-1}x = Bx'$ for some $x' \in \mathbb{F}[X]^n_{\mathrm{COL}}$, but $P^{-1}x = Bx'$ if and only if $x = (PBQ)(Q^{-1}x') = A(Q^{-1}x')$, and hence $P^{-1}x \in B\mathbb{F}[X]^n_{\mathrm{COL}}$ if and only if $x = Ax''$ for some $x'' \in \mathbb{F}[X]^n_{\mathrm{COL}}$ since Q is invertible. The result then follows by the first isomorphism theorem. □

For $p = a_0 + \cdots + a_d X^d \in \mathbb{F}[X]$ and $C \in M_n(\mathbb{F}[X])$ write $p.C$ for the matrix with $(p.C)_{i,j} = p(X)C_{i,j}(X)$ – the . is the scalar multiplication in the $\mathbb{F}[X]$-module $M_n(\mathbb{F}[X])$ – and write $p(C)$ the evaluation homomorphism at C extending the ring homomorphism $\mathbb{F} \to M_n(\mathbb{F}[X])$, which is a composition of the inclusion homomorphism $\mathbb{F} \to \mathbb{F}[X]$ and the diagonal homomorphism $\mathbb{F}[X] \to M_n(\mathbb{F}[X])$ *i.e.*

$$p(C) = a_0.I_n + \cdots + a_d.C^d.$$

Lemma 7.4. *Suppose that $A \in M_n(\mathbb{F})$. Then $e_1^t + (X.I_n - A)\mathbb{F}[X]^n_{\mathrm{COL}}, \ldots, e_n^t + (X.I_n - A)\mathbb{F}[X]^n_{\mathrm{COL}}$ is a basis of the $\mathbb{F}$-vector space $\mathbb{F}[X]^n_{\mathrm{COL}}/(X.I_n - A)\mathbb{F}[X]^n_{\mathrm{COL}}$.*

Proof. Since the matrix $X.I_n$ is in the centre of $M_n(\mathbb{F}[X])$,

$$(X.I_n)^i - A^i = (X.I_n - A)((X.I_n)^{i-1} + \cdots + A^{i-1});$$

and since $\mathbb{F}[X]$ is commutative, left multiplication in $M_n(\mathbb{F}[X])$ is $\mathbb{F}[X]$-linear, so

$$p(X.I_n) - p(A) = (X.I_n - A)\sum_{i=1}^{d} a_i.(A^{i-1} + \cdots + (X.I_n)^{i-1}) = (X.I_n - A)Q$$

for some $Q \in M_n(\mathbb{F}[X])$. Now, the map

$$\Phi : \mathbb{F}[X]^n_{\mathrm{COL}} \to \mathbb{F}^n_{\mathrm{COL}}; p \mapsto p_1(A)e_1^t + \cdots + p_n(A)e_n^t$$

is $\mathbb{F}$-linear, and to identify its kernel we use the same method of proof as for the Factor theorem: Specifically, for $p \in \ker \Phi$ we have

$$\begin{aligned} p = p - \Phi(p) &= (p_1(X.I_n) - p_1(A))e_1^t + \cdots + (p_n(X.I_n) - p_n(A))e_n^t \\ &= (X.I_n - A)(Q_1e_1^t + \cdots + Q_ne_n^t), \end{aligned}$$

for some $Q_1, \ldots, Q_n \in M_n(\mathbb{F}[X])$. In particular, $p \in (X.I_n - A)\mathbb{F}[X]^n_{\mathrm{COL}}$.

In the other direction, $e_1^t + (X.I_n - A)\mathbb{F}[X]^n_{\mathrm{COL}}, \ldots, e_n^t + (X.I_n - A)\mathbb{F}[X]^n_{\mathrm{COL}}$ is $\mathbb{F}$-linearly independent as a subsequence of the subspace $\mathbb{F}[X]^n_{\mathrm{COL}}/(X.I_n - A)\mathbb{F}[X]^n_{\mathrm{COL}}$. To see this, suppose $\lambda_1, \ldots, \lambda_n \in \mathbb{F}$ have $\lambda_1.e_1^t + \cdots + \lambda_n.e_n^t \in (X.I_n - A)\mathbb{F}[X]^n_{\mathrm{COL}}$, say $\lambda_1.e_1^t + \cdots + \lambda_n.e_n^t = (X.I_n - A)q$ for some $q \in \mathbb{F}[X]^n_{\mathrm{COL}}$. If q is not the zero vector then there is i with $\deg q_i \geqslant 0$ maximal, and so the degree of the ith entry of $(X.I_n - A)q$ is $\deg q_i + 1 > 0$, a contradiction. Hence $\lambda_1.e_1^t + \cdots + \lambda_n.e_n^t = 0$ and so $\lambda_1, \ldots, \lambda_n = 0$.

Finally, the vectors $e_1^t + (X.I_n - A)\mathbb{F}[X]^n_{\mathrm{COL}}, \ldots, e_n^t + (X.I_n - A)\mathbb{F}[X]^n_{\mathrm{COL}}$ are also spanning since $\Phi(e_i^t) = e_i^t$ for all $1 \leqslant i \leqslant n$, and $e_1^t, \ldots, e_n^t$ is a spanning subset of $\mathbb{F}^n_{\mathrm{COL}}$. The result is proved. □

Proposition 7.5. *Suppose that $A, B \in M_n(\mathbb{F})$. Then $X.I_n - A$ and $X.I_n - B$ are equivalent as matrices in $M_n(\mathbb{F}[X])$ if and only if A and B are similar as matrices in $M_n(\mathbb{F})$.*

Proof. If A and B are similar then there is $P \in \mathrm{GL}_n(\mathbb{F})$ such that $A = PBP^{-1}$, but then $X.I_n - A = P(X.I_n - B)P^{-1}$ and $X.I_n - A$ is similar, and so equivalent, to $X.I_n - B$ as matrices in $M_n(\mathbb{F}[X])$.

In the other direction, since $X.I_n - A \sim X.I_n - B$, Proposition 7.3 gives an $\mathbb{F}[X]$-linear isomorphism

$$\Phi : \mathbb{F}[X]^n_{\mathrm{COL}}/(X.I_n - A)\mathbb{F}[X]^n_{\mathrm{COL}} \to \mathbb{F}[X]^n_{\mathrm{COL}}/(X.I_n - B)\mathbb{F}[X]^n_{\mathrm{COL}}.$$

By Lemma 7.4 we know $e_1^t + (X.I_n - A)\mathbb{F}[X]^n_{\mathrm{COL}}, \ldots, e_n^t + (X.I_n - A)\mathbb{F}[X]^n_{\mathrm{COL}}$ is an $\mathbb{F}$-basis for $\mathbb{F}[X]^n_{\mathrm{COL}}/(X.I_n - A)\mathbb{F}[X]^n_{\mathrm{COL}}$, and similarly with A replaced by B. Since Φ is, in particular, an $\mathbb{F}$-linear bijection we conclude that there is $P \in \mathrm{GL}_n(\mathbb{F})$ such that

$$\Phi(v + (X.I_n - A)\mathbb{F}[X]^n_{\mathrm{COL}}) = Pv + (X.I_n - B)\mathbb{F}[X]^n_{\mathrm{COL}} \text{ for all } v \in \mathbb{F}^n_{\mathrm{COL}}.$$

Now, Φ is $\mathbb{F}[X]$-linear, so for $v \in \mathbb{F}^n_{\mathrm{COL}}$ we have

$$\begin{aligned} 0 &= \Phi((X.I_n - A)v + (X.I_n - A)\mathbb{F}[X]^n_{\mathrm{COL}}) \\ &= X\Phi(v + (X.I_n - A)\mathbb{F}[X]^n_{\mathrm{COL}}) - \Phi(Av + (X.I_n - A)\mathbb{F}[X]^n_{\mathrm{COL}}) \\ &= XPv - PAv + (X.I_n - B)\mathbb{F}[X]^n_{\mathrm{COL}}. \end{aligned}$$

In other words, $XPv - PAv = Xw - Bw$ for some $w \in \mathbb{F}[X]^n_{\mathrm{COL}}$. Since $v \in \mathbb{F}^n_{\mathrm{COL}}$, no entry on the left can have a non-zero coefficient of X^i for any $i \geqslant 2$, and hence $w \in \mathbb{F}^n_{\mathrm{COL}}$. Equating coefficients we have $w = Pv$ and $PAv = Bw$, and hence $Av = P^{-1}BPv$. Since v was arbitrary we have that $A = P^{-1}BP$ as claimed. □

Given a monic polynomial $f(X) = X^d + a_{d-1}X^{d-1} + \cdots + a_0 \in \mathbb{F}[X]^*$ we define the $d \times d$ matrices

$$C(f) := \begin{pmatrix} 0 & \cdots & \cdots & 0 & -a_0 \\ 1 & \ddots & & \vdots & \vdots \\ 0 & \ddots & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & 0 & \vdots \\ 0 & \cdots & 0 & 1 & -a_{d-1} \end{pmatrix} \text{ and } D(f) := \begin{pmatrix} f(X) & 0 & \cdots & 0 \\ 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 \end{pmatrix}.$$

The matrix $C(f)$ is called the **companion matrix** to f. ⚠ We allow $d = 0$ when these are 'empty' 0×0 matrices.

Example 7.6. For $f(X) \in \mathbb{F}[X]^*$ we have $X.I_d - C(f) \sim_{\mathcal{E}} D(f)$. To see this write $f(X) = X^d + a_{d-1}X^{d-1} + \cdots + a_0$, and put $f_0(X) = 1$ and $f_i = Xf_{i-1}(X) + a_{d-i}$ for $1 \leqslant i \leqslant d$ so that $f_1(X) = X + a_{d-1}$ and $f_d(X) = f(X)$; and apply row and column operations in four groups:

$$X.I_d - C(f) = \begin{pmatrix} X & 0 & \cdots & 0 & a_0 \\ -1 & \ddots & \ddots & \vdots & \vdots \\ 0 & \ddots & \ddots & 0 & \vdots \\ \vdots & \ddots & \ddots & X & a_{d-2} \\ 0 & \cdots & 0 & -1 & X + a_{d-1} \end{pmatrix} \xrightarrow[r_1 \mapsto r_1 + Xr_2]{r_{d-1} \mapsto r_{d-1} + Xr_d} \begin{pmatrix} 0 & 0 & \cdots & 0 & f_d(X) \\ -1 & \ddots & \ddots & \vdots & \vdots \\ 0 & \ddots & \ddots & 0 & \vdots \\ \vdots & \ddots & \ddots & 0 & f_2(X) \\ 0 & \cdots & 0 & -1 & f_1(X) \end{pmatrix}$$

$$\xrightarrow[c_d \mapsto c_d + f_{d-1}(X)c_1]{c_d \mapsto c_d + f_1(X)c_{d-1}} \begin{pmatrix} 0 & 0 & \cdots & 0 & f_d(X) \\ -1 & \ddots & \ddots & \vdots & 0 \\ 0 & \ddots & \ddots & 0 & \vdots \\ \vdots & \ddots & \ddots & 0 & 0 \\ 0 & \cdots & 0 & -1 & 0 \end{pmatrix}$$

$$\xrightarrow[c_{d-1} \leftrightarrow c_d]{c_1 \leftrightarrow c_d} \begin{pmatrix} f_d(X) & 0 & \cdots & 0 \\ 0 & -1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & -1 \end{pmatrix} \xrightarrow[c_d \leftrightarrow (-1)c_d]{c_2 \leftrightarrow (-1)c_2} D(f_d) = D(f).$$

⚠ The order of the row operations in the first group and the column operations in the third group matter, so we do $r_{d-1} \mapsto r_{d-1} + Xr_d$ first and $r_1 \mapsto r_1 + Xr_2$ last in the first group, and

$c_d \mapsto c_d + f_1(X)c_{d-1}$ first and $c_d \mapsto c_d + f_{d-1}(X)c_1$ last in the third group; in the other two groups the operations commute.

For $\lambda \in \mathbb{F}$ and $d \in \mathbb{N}_0$ define the **d-dimensional Jordan matrix with eigenvalue λ** to be the (possibly empty) $d \times d$-matrix

$$J(\lambda, d) := \begin{pmatrix} \lambda & 1 & 0 & \cdots & 0 \\ 0 & \lambda & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \lambda & 1 \\ 0 & \cdots & \cdots & 0 & \lambda \end{pmatrix}; \text{ and recall } D((X-\lambda)^d) = \begin{pmatrix} (X-\lambda)^d & 0 & \cdots & 0 \\ 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 \end{pmatrix}.$$

Example 7.7. For $\lambda \in \mathbb{F}$ we have $X.I_d - J(\lambda, d) \sim_{\mathcal{E}} D((X-\lambda)^d)$.

To see this note that $X.I_d - J(\lambda, d)$ equals

$$\begin{pmatrix} X-\lambda & -1 & 0 & \cdots & 0 \\ 0 & X-\lambda & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & X-\lambda & -1 \\ 0 & \cdots & \cdots & 0 & X-\lambda \end{pmatrix} \xrightarrow[c_1 \mapsto c_1 + (X-\lambda)c_2]{\substack{c_{d-1} \mapsto c_{d-1} + (X-\lambda)c_d \\ \ddots}} \begin{pmatrix} 0 & -1 & 0 & \cdots & 0 \\ 0 & 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & 0 & -1 \\ (X-\lambda)^d & \cdots & \cdots & 0 & 0 \end{pmatrix}$$

$$\xrightarrow[r_{d-1} \leftrightarrow r_d]{\substack{r_1 \leftrightarrow r_d \\ \ddots}} \begin{pmatrix} (X-\lambda)^d & 0 & \cdots & 0 \\ 0 & -1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & -1 \end{pmatrix} \xrightarrow[c_d \leftrightarrow (-1)c_d]{\substack{c_2 \leftrightarrow (-1)c_2 \\ \ddots}} D((X-\lambda)^d)$$

Theorem 7.8. *Suppose that $A \in M_n(\mathbb{F})$. Then there are monic polynomials $f_1 \mid \cdots \mid f_n$ such that $A \sim_{\mathcal{E}} C(f_1) \oplus \cdots \oplus C(f_n)$.*

Proof. By Theorem 6.12 there are polynomials $f_1 \mid \cdots \mid f_n$ such that $X.I_n - A \sim_{\mathcal{E}} \Delta(X)$ where $\Delta(X)$ is the diagonal matrix with entries $f_1, \ldots, f_n$. In particular, there are $P, Q \in \mathrm{GL}_n(\mathbb{F}[X])$ such that $P(X)(X.I_n - A)Q(X) = \Delta(X)$. Since $P(X)$ and $Q(X)$ are invertible we have $\det P(X) \det P(X)^{-1} = 1$ and hence $\det P(X), \det Q(X) \in U(R)$ (see Exercise IV.7 for the proof that determinant is multiplicative), hence $\det(X.I_n - A)$ are associates $\det \Delta(X)$ in $\mathbb{F}[X]$. In particular, since $\det(X.I_n - A)$ is monic and of degree n, none of $f_1, \ldots, f_n$ is identically 0 and so they all have degrees which we denote $d_1, \ldots, d_n$ respectively and satisfy $n = d_1 + \cdots + d_n$. Moreover, by multiplying by units we may assume that $f_1, \ldots, f_n$ are monic.

By permuting columns and rows as necessary we have $\Delta(X) \sim_{\mathcal{E}} D(f_1) \oplus \cdots \oplus D(f_n)$. The calculation in Example 7.6 shows us that $D(f_i) \sim_{\mathcal{E}} X.I_{d_i} - C(f_i)$ and hence $X.I_n - A \sim_{\mathcal{E}} X.I_n - C(f_1) \oplus \cdots \oplus C(f_n)$. The result now follows from Proposition 7.5. □

A matrix is said to be in **rational canonical form** if it is a block diagonal matrix with blocks $C(f_1)$, ..., $C(f_n)$ for monic polynomials $f_1 \mid \cdots \mid f_n$. In particular, the above says that every matrix is similar to a matrix in rational canonical form.

Remark 7.9. Although we shall not prove it, if two matrices in rational canonical form are similar then they are equal.

Example 7.10. For $f(X) = (X-\lambda_1)^{d_1}\cdots(X-\lambda_n)^{d_n}$ with $\lambda_1, \ldots, \lambda_n$ pairwise distinct and $d_1 + \cdots + d_n = n$, we have

$$D(f) \sim_{\mathcal{E}} D((X-\lambda_1)^{d_1}) \oplus \cdots \oplus D((X-\lambda_n)^{d_n}).$$

To see this, we use induction on r to show that

$$D((X-\lambda_1)^{d_1}) \oplus \cdots \oplus D((X-\lambda_n)^{d_n}) \sim_{\mathcal{E}} D(f_r) \oplus D((X-\lambda_{r+1})^{d_{r+1}}) \oplus \cdots \oplus D((X-\lambda_n)^{d_n})$$

where $f_r(X) = (X-\lambda_r)^{d_r} f_{r-1}(X)$ and $f_0(X) = 1$. This is certainly true when $r = 0$, and for the inductive step when $r < n$ note that the ideal generated by f_r and $(X-\lambda_{r+1})^{d_{r+1}}$ is principal, say generated by g_r. If g_r has a root then it is a root of $(X-\lambda_{r+1})^{d_{r+1}}$ and also of f_r, hence g_r has no root and $\langle f_r \rangle + \langle (X-\lambda_{r+1})^{d_{r+1}} \rangle = \langle 1 \rangle$. By Lemma 6.10 we can replace the first element on the diagonal – that is f_r – by $f_r(X-\lambda_{r+1})^{d_{r+1}} = f_{r+1}$, and the $(\deg f_r + 1)$st element on the diagonal – that is $(X-\lambda_{r+1})^{d_{r+1}}$ by 1. The resulting matrix has $(\deg f_r) - 1 + d_{r+1} = (\deg f_{r+1}) - 1$ copies of 1 on the diagonal after the initial f_{r+1}, and hence equals $D(f_{r+1}) \oplus D((X-\lambda_{r+1})^{d_{r+1}}) \oplus \cdots \oplus D((X-\lambda_n)^{d_n})$. The example is complete.

Theorem 7.11. *Suppose that $A \in M_n(\mathbb{C})$. Then there are $\lambda_1, \ldots, \lambda_n \in \mathbb{C}$ and $t_1, \ldots, t_n \in \mathbb{N}_0$ with $t_1 + \cdots + t_n = n$ such that $A \sim_{\mathcal{E}} J(\lambda_1, t_1) \oplus \cdots \oplus J(\lambda_n, t_n)$.*

Proof. By Theorem 6.6 there are polynomials $f_1, \ldots, f_n$ such that $X.I_n - A \sim_{\mathcal{E}} \Delta(X)$ where $\Delta(X)$ is the diagonal matrix with entries $f_1, \ldots, f_n$. As in the proof of Theorem 7.8 we conclude that we may suppose each f_i is monic and write d_i for its degree, and $n = d_1 + \cdots + d_n$. By permuting columns and rows as necessary we have $\Delta(X) \sim_{\mathcal{E}} D(f_1) \oplus \cdots \oplus D(f_n)$.

If $f \in \mathbb{C}[X]$ is irreducible then $f(X) \sim X - \lambda$ for some $\lambda \in \mathbb{C}$ – this is where we use the fact that the field is the complex numbers rather than a more general field – so since $\mathbb{C}[X]$ is a Factorisation domain, we conclude that $f_i(X) = (X-\lambda_{i,1})^{d_{i,1}}\cdots(X-\lambda_{i,r_i})^{d_{i,r_i}}$ with $\lambda_{i,1}, \ldots, \lambda_{i,r_i}$ pairwise distinct and $d_{i,1} + \cdots + d_{i,r_i} = d_i$. In view of the calculation in Examples 7.7 & 7.10 we have

$$\begin{aligned} D(f_i) &\sim_{\mathcal{E}} D((X-\lambda_{i,1})^{d_{i,1}}) \oplus \cdots \oplus D((X-\lambda_{i,r_i})^{d_{i,r_i}}) \\ &\sim_{\mathcal{E}} (X.I_{d_{i,1}} - J(\lambda_{i,1}, d_{i,1})) \oplus \cdots \oplus (X.I_{d_{i,r_i}} - J(\lambda_{i,r_i}, d_{i,r_i})). \end{aligned}$$

Finally, let $\lambda_1, \ldots, \lambda_n$ be $\lambda_{1,1}, \ldots, \lambda_{1,r_1}, \lambda_{2,1}, \ldots, \lambda_{2,r_2}, \ldots, \lambda_{n,1}, \ldots, \lambda_{n,r_n}$ in order and similarly for $t_1, \ldots, t_n$. The result is proved by Proposition 7.5. □

A matrix is said to be in **Jordan normal form** if it is a block diagonal matrix with blocks $J(\lambda_1, d_1)$, ..., $J(\lambda_n, d_n)$ for $\lambda_1, \ldots, \lambda_n \in \mathbb{F}$ and $d_1, \ldots, d_n \in \mathbb{N}_0$. In particular, the above theorem says that every matrix over $\mathbb{C}$ is similar to a matrix in Jordan normal form.

References

[Ber14] D. Berlyne. Ideal theory in rings (Translation of "Idealtheorie in Ringbereichen" by Emmy Noether). 2014, arXiv:1401.2577.

[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.

[Coh66] P. M. Cohn. On the structure of the GL_2 of a ring. *Inst. Hautes Études Sci. Publ. Math.*, (30):5–53, 1966. URL `http://www.numdam.org/item?id=PMIHES_1966__30__5_0`.

[Con] K. Conrad. Remarks about Euclidean domains. URL `https://kconrad.math.uconn.edu/blurbs/ringtheory/euclideanrk.pdf`.

[Fuc58] L. Fuchs. *Abelian groups*. Publishing House of the Hungarian Academy of Sciences, Budapest, 1958.

[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.

[Hel43] O. Helmer. The elementary divisor theorem for certain rings without chain condition. *Bull. Amer. Math. Soc.*, 49(4):225–236, 04 1943. doi:10.1090/S0002-9904-1943-07886-X.

[Kap49] I. Kaplansky. Elementary divisors and modules. *Trans. Amer. Math. Soc.*, 66:464–491, 1949. doi:10.2307/1990591.

[Kap70] I. Kaplansky. *Commutative rings*. Allyn and Bacon, Inc., Boston, Mass., 1970.

[Kea98] M. E. Keating. *A First Course in Module Theory*. Imperial College Press, 1998. doi:https://doi.org/10.1142/p082.

[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.

[LLS74] M. D. Larsen, W. J. Lewis, and T. S. Shores. Elementary divisor rings and finitely presented modules. *Transactions of the American Mathematical Society*, 187:231–248, 1974. doi:10.2307/1997051.

[Lor12] D. Lorenzini. Elementary divisor domains and Bézout domains. *J. Algebra*, 371:609–619, 2012. doi:10.1016/j.jalgebra.2012.08.020.

[Noe21] E. Noether. Idealtheorie in Ringbereichen. *Math. Ann.*, 83(1-2):24–66, 1921. doi:10.1007/BF01464225.

[Poo19] B. Poonen. Why all rings should have a 1. *Math. Mag.*, 92(1):58–62, 2019. doi:10.1080/0025570X.2018.1538714.

[She88] K. Shen. The historical development of the Chinese remainder theorem. *J. Hangzhou Univ. Natur. Sci. Ed.*, 15(3):270–282, 1988.

[Sou20] K. Soundararajan. Bertrand's postulate and the existence of finite fields. 2020, arXiv:2007.01389.

[Tol04] J. R. R. Tolkein. *The Fellowship of the Ring. The Lord of the Rings Part I.* HarperCollins e-books, 50th anniversary edition, 2004. URL `https://s3.amazonaws.com/scschoolfiles/112/j-r-r-tolkien-lord-of-the-rings-01-the-fellowship-of-the-ring-retail-pdf.pdf`.