A3: RINGS AND MODULES, 2019–2020

TOM SANDERS

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

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

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

Course Synopsis: Recap on rings (not necessarily commutative or with an identity) 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 characterization (proofs non-examinable) [2]

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

Last updated: 1st April, 2020.

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

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

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

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

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

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

Blue text indicates non-examinable material.

1. Rings: a recap and motivating examples

We begin by fixing some terminology for some concepts which have been introduced in previous courses. A **ring** R is a set (also denoted R and called the **carrier set**) equipped with two binary operations $+$, called **addition**, and $\times$, called **multiplication**, such that

(i) the set R equipped with $+$ is a commutative group, called the **additive group**;
(ii) $\times$ is an associative operation on R;
(iii) $\times$ **distributes** over $+$, meaning that

$$x \times (y + z) = x \times y + x \times z \text{ and } (x + y) \times z = (x \times z) + (y \times z) \text{ for all } x, y, z \in R.$$

The additive group of a ring has a unique identity called the **zero** of the ring and denoted 0. We write $-x$ for the **additive inverse** of $x \in R$; the map $R \to R; x \mapsto -x$ is called **negation**.

We say R is a **commutative** ring if multiplication is commutative.

We shall often write xy in place of $x \times y$ in a ring.[1] We respect the usual precedence of multiplication over addition so by $x + yz$ we mean $x + (y \times z)$, and not $(x + y) \times z$.

We write R^* for the set of non-zero elements of the ring R.

Lemma 1.1. *Suppose that R is a ring.*

(i) (Zero annihilates) $0x = x0 = 0$ *for all* $x \in R$;
(ii) (Negation distributes) $-(xy) = (-x)y = x(-y)$ *for all* $x, y \in R$.

Proof. First $0 = 0x + (-(0x)) = (0 + 0)x + (-(0x)) = (0x + 0x) + (-(0x)) = 0x + 0 = 0x$ for all $x \in R$ and similarly for $x0$. Secondly, $(xy) + ((-x)y) = (x + (-x))y = 0y = 0$ and so by uniqueness of additive inverses $-(xy) = (-x)y$, and similarly $-(xy) = x(-y)$. □

It may happen that multiplication has an identity. If it does then this identity is unique and we denote it 1 and call it the **multiplicative identity**. A ring with a multiplicative identity is called **unital**.

⚠ Take care here as some authors define a ring to be what we call unital ring; see *e.g.* [Poo19] for some motivation for this point of view.

If R is a unital ring, then we say that $x \in R$ is a **unit** if it has an inverse with respect to multiplication, and we write $U(R)$ for the set of units of R. If x does have an inverse with respect to multiplication then it is unique; we call it the **multiplicative inverse** of x and denote it x^{-1}.

⚠ Some authors write R^* for $U(R)$ – we reserve R^* for the non-zero elements of the ring – and some write $R^\times$ for $U(R)$.

Proposition 1.2 (Unit group)**.** *Suppose that R is a unital ring. Then multiplication on R restricts to a group operation on $U(R)$ with identity* 1 *and the inverse of x under this restricted operation is also x^{-1}.*

[1] Since $x-y$, which is shorthand for $x+(-y)$, and $x-y$, which is shorthand for $x\times(-y)$, look remarkably similar, we shall avoid the latter.

Proof. Certainly $1 \in U(R)$ since $1 \times 1 = 1$. If $x, y \in U(R)$ then there are elements $u, v \in R$ such that $xu = ux = 1$ and $yv = vy = 1$, whence $(xy)(vu) = x((yv)u) = x(1u) = xu = 1$ and similarly $(vu)(xy) = 1$ so $xy \in U(R)$, which means multiplication on R restricts to a binary operation on $U(R)$.

Associativity of this restricted operation is inherited from multiplication on R, as is the fact that 1 is an identity. It remains to note that if $x \in U(R)$ then $xx^{-1} = 1 = x^{-1}x$ and so $x^{-1} \in U(R)$ and x^{-1} is an inverse for x under this multiplication map. □

A key example of a commutative unital ring, and the source of the terminology above, is the integers $\mathbb{Z}$. For them we have $U(\mathbb{Z}) = \{-1, 1\}$. On the other hand $2\mathbb{Z}$, the set of even integers, with operations inherited from $\mathbb{Z}$ is an example of a ring that is not a unital ring.

Given a ring R we say that S is a **subring** of R if it is a subset of R, and a ring when the addition and multiplication on R are restricted to S. We say that S is a **unital subring** of R if S is a subring of R, R is unital and S contains the multiplicative identity of R.

⚠ $\{0\}$ is a subring of $\mathbb{Z}$ and both are unital rings, but the former is *not* a unital subring of the latter.

Lemma 1.3. *Suppose that R is a ring and $\mathcal{S}$ is a non-empty set of (unital) subrings of R. Then $\bigcap_{S \in \mathcal{S}} S$ is a (unital) subring of R.*

Proof. Note that $0 \in S$ for all $S \in \mathcal{S}$ so $0 \in \bigcap_{S \in \mathcal{S}} S$. Similarly, if $x, y \in \bigcap_{S \in \mathcal{S}} S$ then $x, y \in S$ for all $S \in \mathcal{S}$, and hence $x + (-y) \in S$ for all $S \in \mathcal{S}$ and so $x + (-y) \in \bigcap_{S \in \mathcal{S}} S$. It follows that the intersection is an additive subgroup. Similarly it is multiplicatively closed and associativity of multiplication and distributivity of multiplication over addition are inherited from the operations on R. Finally if all the rings in $\mathcal{S}$ are unital, they all contain 1 and so the intersection does too. □

Given a subring S of R, and elements $\lambda_1, \dots, \lambda_n \in R$ we write $S[\lambda_1, \dots, \lambda_n]$ for the intersection of all subrings containing S and $\lambda_1, \dots, \lambda_n$, which is a subring since R is certainly a subring of R containing S and $\lambda_1, \dots, \lambda_n$. $S[\lambda_1, \dots, \lambda_n]$ is a unital subring of R if S is a unital subring of R.

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

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

a **unital ring homomorphism** is a ring homomorphism $\phi : R \to S$ between two unital rings with the additional property that $\phi(1) = 1$.

If S is a (unital) subring of R then the inclusion map $S \to R$ is a (unital) homomorphism.

⚠If R and S are rings and there is an obvious injective (unital) homomorphism $j : S \to R$ we shall frequently identify S with $j(S)$, speak of S as a (unital) subring of R ,and write $S[\lambda_1, \dots, \lambda_n]$ for what would properly be written $j(S)[\lambda_1, \dots, \lambda_n]$.

Lemma 1.4. *Suppose that $\phi : R \to S$ is a ring homomorphism. Then $\phi(0) = 0$ and $\phi(-x) = -\phi(x)$ for all $x \in R$. If ϕ is a unital ring homomorphism and $x \in U(R)$ then $\phi(x) \in U(S)$ and $\phi(x^{-1}) = \phi(x)^{-1}$.*

Proof. For $x \in R$ we have

$$\begin{aligned}\phi(-x) = \phi(-x) + 0 &= \phi(-x) + (\phi(x) + (-\phi(x)))\\ &= (\phi(-x) + \phi(x)) + (-\phi(x)) = \phi((-x) + x) + (-\phi(x)) = \phi(0) + (-\phi(x)).\end{aligned}$$

Setting $x = 0$ above gives $\phi(0) = 0$, which then gives the second fact. A similar argument[2] shows that if ϕ is a unital homomorphism and $x \in U(R)$ then $\phi(x) \in U(S)$ and $\phi(x^{-1}) = \phi(x)^{-1}$. □

In particular, if $\phi : R \to S$ is a unital ring homomorphism then $\phi(U(R)) \leqslant U(S)$.

⚠ The inclusion $j : \mathbb{Z} \to \mathbb{Q}$ is a unital homomorphism and $\phi(\mathbb{Z}) \cap U(\mathbb{Q}) = \mathbb{Z} \neq \{-1, 1\} = \phi(U(\mathbb{Z}))$.

Any commutative group can be given a ring structure by setting all products to be 0; we call this a **trivial multiplicative structure**. We call a ring **trivial** if it has one element – it is the one element additive group with the aforementioned multiplicative structure. A trivial ring is unital with $0 = 1$, and a unital ring is trivial if $0 = 1$.[3]

It follows from Lemma 1.1 that unless a unital ring is trivial we must have $U(R) \subset R^*$. A **field** is a (non-trivial) commutative unital ring in which $U(R) = R^*$, and this gives us some more examples of commutative unital rings: $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$, and $\mathbb{F}_p$ (the integers modulo p for p a prime).

The above examples are all commutative but they can be used to produce non-commutative rings. Given a unital ring R and $n \in \mathbb{N}$ we write $M_n(R)$ for the set of $n \times n$ matrices with entries in R. We define addition and multiplication of two elements $A, B \in M_n(R)$ by

$$A + B := (A_{ij} + B_{ij})_{i,j=1}^{n} \text{ and } AB := \left(\sum_{k=1}^{n} A_{ik}B_{kj}\right)_{i,j=1}^{n},$$

and these operations make $M_n(R)$ into a unital ring, where $0_{M_n(R)}$ is the matrix with 0_R in every entry and $1_{M_n(R)}$ is the matrix with 1_R on the main diagonal and 0_R elsewhere.

These rings are called **matrix rings** and they are not commutative (provided either R is not commutative or $n > 1$).

For $A_1, \ldots, A_k \in M_n(R)$ we write $R[A_1, \ldots, A_k]$ for the unital subring of $M_n(R)$ generated by the scalar multiples of the identity – that is the matrices with r on the main diagonal for some $r \in R$ and 0_R elsewhere – and the matrices $A_1, \ldots, A_k$.

2. Integral domains and polynomial rings

Suppose that R is a ring. We say that $x \in R$ is a **(left) zero divisor** if there is some $y \in R^*$ such that $xy = 0$, and similarly for right zero divisors.

[2] Note that we require $\phi(1) = 1$ so that we know $\phi(1)^{-1}$ exists; we know that $-\phi(0)$ exists in the first argument because S under addition is a group.

[3] Lemma 1.1 shows that if $0 = 1$ then $0 = 0x = 1x = x$ for all $x \in R$ and hence that R has only one element.

Given a field $\mathbb{F}$ the ring $M_2(\mathbb{F})$ has non-zero zero divisors *e.g.*

$$\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix};$$

and the ring of integers mod N has non-zero zero-divisors if (and only if) N is composite: if $N = pq$ for $1 < p, q < N$ then $p, q \not\equiv 0 \pmod{N}$ but $pq = N \equiv 0 \pmod{N}$.

We say that a ring R is an **integral domain** if it is a non-trivial commutative unital ring with no non-zero zero divisors.

Lemma 2.1 (Cancellation lemma)**.** *Suppose that R is an integral domain, and $x \in R^*$ and $y, z \in R$ have $xy = xz$. Then $y = z$.*

Proof. By distributivity we have $0 = (xy) - (xz) = x(y - z)$. Since $x \in R^*$ and R is an integral domain it follows that $y - z = 0$ as required. □

This immediately gives the following cute proposition.

Proposition 2.2. *Suppose that R is a finite integral domain. Then R is a field.*

Proof. Since R is non-trivial R^* is non-empty. For $a \in R^*$ consider the map $R \to R; x \mapsto ax$. This is an injection by the cancellation lemma, and since R is finite it is a surjection. It follows that there is some $x \in R$ such that $ax = 1$. Since R is commutative, we conclude that $xa = ax = 1$ so $a \in U(R)$ and hence R is a field. □

Integral domains and polynomial rings are closely related. Suppose that R is a commutative unital ring (but not necessarily an integral domain). We write $R[X]$ for the set of R**-polynomials** in the variable X with coefficients in R, that is the set of expressions of the form

$$p(X) = \sum_{i=0}^{\infty} a_i X^i, \tag{2.1}$$

where $a_i \in R$ for all i, and $a_i \in R^*$ for finitely many $i \in \mathbb{N}_0$. We call the a_is the **coefficients** of the polynomial; two polynomials are equal if and only if their coefficients are equal. Given polynomials p and q with coefficients $(a_i)_{i=0}^{\infty}$ and $(b_i)_{i=0}^{\infty}$ we have

$$(p+q)(X) = \sum_{i=0}^{\infty} (a_i + b_i) X^i \text{ and } (pq)(X) := \sum_{i=0}^{\infty} \left(\sum_{j+k=i} a_j b_k \right) X^i.$$

With these operations $R[X]$ is a commutative unital ring.

The map $j : R \to R[X]$ taking elements of R to the corresponding constant polynomial is an injective unital homomorphism, and we shall write r both for an element of R and the constant polynomial r.

⚠ This is a very important map: if $\mathbb{F}$ is a field then the map $j : \mathbb{F} \to \mathbb{F}[X]$ above gives $\mathbb{F}[X]$ the structure of an $\mathbb{F}$-vector space.

Note that the sum in (2.1) is just a notation and should be thought of as a way of recording the coefficients of the polynomial. It is written this way to give a suggestion for how to evaluate the polynomial: if $\lambda \in R$ and $p \in R[X]$ then the map

$$R[X] \to R; p \mapsto p(\lambda) := \sum_{i:a_i \neq 0} a_i\lambda^i,$$

with the usual convention that the empty sum is 0, is a well-defined (since the sum is finite) unital ring homomorphism.

⚠ The maps $\mathbb{F}_p \to \mathbb{F}_p; \lambda \mapsto \lambda^p$ and $\mathbb{F}_p \to \mathbb{F}_p; \lambda \mapsto \lambda$ are the same by Fermat's Little Theorem, but the polynomials $q(X) = X^p$ and $q(X) = X$ are distinct.

If $p \in R[X]^*$ then we define its **degree**[4], denoted $\deg p$, to be the largest $d \in \mathbb{N}_0$ such that the coefficient of X^d in p is non-zero, and we say that $\lambda \in R$ is a **root** of p if $p(\lambda) = 0$.

Proposition 2.3. *Suppose that R is a non-trivial commutative unital ring. Then the following are equivalent:*

(i) R is an integral domain;
(ii) $R[X]$ is an integral domain;
(iii) for every $p, q \in R[X]^*$ we have $\deg pq = \deg p + \deg q$;
(iv) every $p \in R[X]^*$ of degree at most d has at most d roots.

Proof. Certainly (ii) implies (i) by looking at the constant polynomials, and (iii) implies (ii) since the latter is just the former with the second part forgotten.

To see (i) implies (iii) suppose that $p, q \in R[X]^*$ so we can write

$$p(X) = a_nX^n + \cdots + a_0 \text{ and } q(X) = b_mX^m + \cdots + b_0$$

where $a_n, \ldots, a_0, b_m, \ldots, b_0 \in R$, $a_n, b_m \in R^*$ and $n = \deg p$ and $m = \deg q$. Then

$$(pq)(X) = p(X)q(X) = \sum_{r=0}^{n+m} \sum_{\substack{i+j=r \\ i \leqslant n \\ j \leqslant m}} a_ib_jX^{i+j}.$$

Thus the coefficient of X^{n+m} is a_nb_m, and it is non-zero since R is an integral domain. We conclude that $pq \in R[X]^*$ and $\deg pq = n + m = \deg p + \deg q$.

If R is not an integral domain then there are $a, b \in R^*$ such that $ab = 0$, but then the polynomial aX has degree 1 but at least two roots since $ab = 0$ and $a0 = 0$ by Lemma 1.1. This shows (iv) implies (i).

In the other direction, we proceed by induction on the degree d assuming, as we may, (i) and (ii). If $d = 0$ then the polynomial is a non-zero constant and so has no roots as

[4] Occasionally it is useful to define the degree of the zero polynomial to be $-\infty$ and adopt the arithmetic convention that $-\infty + n = -\infty$ for all $n \in \mathbb{N}_0$ so that the degree identity in Proposition 2.3 (iii) continues to hold.

required. Now suppose that $d > 0$ and p has a root $\lambda \in R$. Then

$$\begin{aligned} p(X) = p(X) - p(\lambda) &= \sum_{n=0}^{d} a_n(X^n - \lambda^n) \\ &= (X-\lambda)\sum_{n=1}^{d} a_n(X^{n-1} + \cdots + \lambda^{n-1}) \\ &= (X-\lambda)\sum_{m=0}^{d-1}\left(\sum_{l=0}^{d-m-1} a_{l+m+1}\lambda^l\right)X^m =: (X-\lambda)q(X). \end{aligned}$$

Since $R[X]$ is an integral domain, $q \in R[X]^*$, and from its definition we see that $\deg q \leqslant d-1$. If $\lambda' \neq \lambda$ then $p(\lambda') = 0$ if and only if $q(\lambda') = 0$ since R is an integral domain. The result follows. □

In view of (iii) above, when R is an integral domain, the units of $U(R[X])$ have to be polynomials of degree 0, and the map $U(R) \rightarrow U(R[X])$ taking r to the degree zero polynomial with constant coefficient r is an isomorphism.

Note that $(1+2X)^2 = 1$ in $(\mathbb{Z}/4\mathbb{Z})[X]$, so that $1+2X \in U((\mathbb{Z}/4\mathbb{Z})[X])$. More generally, if R is a commutative unital ring then $U(R[X])$ is exactly the set of polynomials $a_0+a_1X+\cdots+a_dX^d$ where $a_0 \in U(R)$ and there is some $n \in \mathbb{N}$ such that $a_i^n = 0$ for all $1 \leqslant i \leqslant d$. ($x \in R$ is called **nilpotent** if $x^d = 0$ for some $d \in \mathbb{N}$.)

We write $R[X_1, \ldots, X_n]$ for the ring of polynomials in the variables $X_1, \ldots, X_n$ with coefficients in R. Already for $n = 2$ this gives two ways of viewing the resulting ring. If R is an integral domain then Proposition 2.3 tells us that $R[X]$ and $R[Y]$ are both integral domains and so elements of $R[X,Y]$ have both an X-degree, when considered as elements of $R[Y][X]$, and a Y-degree when considered as elements of $R[X][Y]$.

The degree bound on the number of roots in Proposition 2.3 (iv) is important so we give an application. The first part of the proof below has a lot in common with the method we shall use later to establish the existence of Smith Normal Form, and is an adaptation of Schenkman's proof of the basis theorem for finitely generated commutative groups in [Sch60].

Proposition 2.4. *Suppose that $\mathbb{F}$ is a finite field. Then $U(\mathbb{F})$ is cyclic.*

Proof. Suppose that $\{x_1, \ldots, x_n\}$ is a smallest set of generators for $U(\mathbb{F})$. If $n > 1$ let G be the group generated by x_{n-1} and x_n, and (x,y) be a pair of generators for G (not necessarily from X) with the order of x minimal out of all such pairs.

Let au be the order of x and bu be the order of y with $\mathrm{hcf}(a,b) = 1$. By Bezout's Theorem there are $\alpha, \beta \in \mathbb{Z}$ such that $a\alpha + b\beta = 1$. Put $z := x^a y^{-b}$ and $w := x^\beta y^\alpha$, so $x = z^\alpha w^b$ and $y = z^{-\beta}w^a$ and the pair (z,w) generates G. Moreover, $z^u = x^{au}y^{-bu} = 1 \cdot 1 = 1$ and so $u \geqslant au$ by minimality of the order of x.

Since y has order bu, the elements $y^b, y^{2b}, \ldots, y^{ub}$ are u distinct elements, and they are all roots of $X^u - 1 \in \mathbb{F}[X]^*$. Hence by Proposition 2.3 (iv) they are the only roots of $X^u - 1$, but then x is also a root of $X^u - 1$ and so $x = y^{bi}$ for some $1 \leqslant i \leqslant u$. We conclude G is generated by the one element y, and hence $\{x_1, \ldots, x_{n-2}, y\}$ is a generating set for $U(\mathbb{F})$, contradicting the minimality of n. Hence $n = 1$ and $U(\mathbb{F})$ is cyclic. □

Conrad collects together many different proofs of the above result in [Con].

The construction of $\mathbb{Q}$ from $\mathbb{Z}$ only uses the fact that $\mathbb{Z}$ is an integral domain. Given an integral domain R let $\mathrm{Frac}(R)$ be the pairs $(a, b) \in R \times R^*$ subject to the equivalence relation

$$(a, b) \sim (a', b') \text{ if and only if } ab' = a'b.$$

Addition and multiplication are defined by

$$(a, b) + (a', b') := (ab' + a'b, bb') \text{ and } (a, b)(a', b') := (aa', bb')$$

for all $a, a' \in R$ and $b, b' \in R^*$ which are well-defined by the cancellation lemma. The relevant features are summarised in the following theorem whose proof is just a check.

Theorem 2.5 (Field of fractions)**.** *Suppose that R is an integral domain. Then there is a field* $\mathrm{Frac}(R)$ *and an injective unital ring homomorphism $\iota : R \to \mathrm{Frac}(R)$ such that for any field $\mathbb{F}$ and injective unital ring homomorphism $\phi : R \to \mathbb{F}$ there is a unique injective unital ring homomorphism $\psi : \mathrm{Frac}(R) \to \mathbb{F}$ such that $\psi \circ \iota = \phi$* i.e. *so the following diagram commutes*

$$\begin{tikzcd} R \arrow[r, "\iota"] \arrow[dr, "\phi"] & \mathrm{Frac}\, R \arrow[d, "\psi"] \\ & \mathbb{F} \end{tikzcd}$$

The field of fractions may be quite large: the field of fractions of $\mathbb{R}[X]$ is denoted $\mathbb{R}(X)$. It is the set of rational functions, that is ratios $\frac{p(X)}{q(X)}$ where $p \in \mathbb{R}[X]$ and $q \in \mathbb{R}[X]^*$.

Viewing $\mathbb{R}[X]$ as a vector space over $\mathbb{R}$, the set $\{1, X, X^2, \ldots\}$ is a basis. $\mathbb{R}(X)$ is also a vector space over $\mathbb{R}$ but in this case it is much larger.

Lemma 2.6. *The set $\{(X - \lambda)^{-1} : \lambda \in \mathbb{R}\}$ is (uncountable and) linearly independent in $\mathbb{R}(X)$.*

Proof. To check linear independence suppose that there were distinct reals $\lambda_1, \ldots, \lambda_k$ and $\alpha_1, \ldots, \alpha_k$ such that

$$\alpha_1(X - \lambda_1)^{-1} + \cdots + \alpha_k(X - \lambda_k)^{-1} = 0.$$

Rearranging we get

$$\alpha_1 \prod_{i \neq 1} (X - \lambda_i) + \cdots + \alpha_k \prod_{i \neq k} (X - \lambda_i) = 0;$$

and then evaluating successively at $\lambda_1, \ldots, \lambda_k$ we get $\alpha_j \prod_{i \neq j} (\lambda_j - \lambda_i) = 0$ for $1 \leqslant j \leqslant k$ which in turn implies $\alpha_j = 0$ for $1 \leqslant j \leqslant k$. The lemma is proved. □

3. Homomorphisms and ideals

A ring homomorphism between rings R and S is called an **isomorphism** if it has an inverse map that is also a homomorphism; if such a function exists we say that R and S are **isomorphic**.

Lemma 3.1. *Suppose that $\phi : R \to S$ is a bijective ring homomorphism. Then the inverse is also a homomorphism. Moreover, if R or S is unital then they are both unital and ϕ and its inverse are both unital homomorphisms.*

Proof. Suppose that $x, y \in S$. Since ϕ is bijective there are elements $u, v \in R$ such that $x = \phi(u)$ and $y = \phi(v)$. Hence

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

Similarly $\phi^{-1}(xy) = \phi^{-1}(x)\phi^{-1}(y)$.

Now suppose that one of R and S is unital; we may assume R is unital by switching R and S and replacing ϕ by ϕ^{-1} if necessary. Bijectivity of ϕ means for all $s \in S$ there is some $r \in R$ such that $\phi(r) = s$, and hence $\phi(1_R)s = \phi(1_R)\phi(r) = \phi(r) = s = \phi(r) = \phi(r)\phi(1_R) = s\phi(1_R)$. It follows that $\phi(1_R)$ is a multiplicative identity in S, so S is unital with $1_S = \phi(1_R)$, and ϕ and ϕ^{-1} are unital homomorphisms. □

⚠ This bootstrapping of bijections occurs for groups, vector spaces, and many other algebraic structures. On the other hand $f : [0,1) \cup \{2\} \to [0,1]$ with $f(x) = x$ if $x < 1$ and $f(2) = 1$ is a continuous bijection, but the inverse function is *not* continuous.

Lemma 3.2. *Suppose that $\phi : S \to R$ is a (unital) ring homomorphism. Then $\phi(S)$ is a (unital) subring of R.*

Proof. Since S is nonempty, $\phi(S)$ is non-empty. Moreover if $x, y \in \phi(S)$ then there are elements $u, v \in S$ such that $x = \phi(u)$ and $y = \phi(v)$. Then $xy = \phi(u)\phi(v) = \phi(uv) \in \phi(S)$ since S is multiplicatively closed. Additionally, by Lemma 1.4, $x+(-y) = \phi(u)+(-\phi(v)) = \phi(u) + \phi(-v) = \phi(u + (-v)) \in \phi(S)$ since S is an additive group. Since associativity and distributivity are inherited from R we conclude that $\phi(S)$ is a subring of R. If ϕ is a unital homomorphism then $1 = \phi(1) \in \phi(S)$ and so $\phi(S)$ is a unital subring of R. □

Suppose that R is a ring. We say that I is an **ideal**[5] if it is an additive subgroup of R with $xr, rx \in I$ for all $r \in R$ and $x \in I$. The notation $I \lhd R$ is used in places (*e.g.* [Coh00, p12]) to mean I is an ideal of R.

The sets $\{0\}$ and R are always ideals in any ring, and so the only ring with fewer than two ideals is a trivial ring, where $R = \{0\}$.

Given a ring homomorphism $\phi : R \to S$, the **kernel of** ϕ is the set $\ker \phi := \{x \in R : \phi(x) = 0\}$.

Lemma 3.3. *Suppose that $\phi : R \to S$ is a homomorphism. Then $\ker \phi$ is an ideal in R.*

[5] One might also call these two-sided ideals by way of contrast with left and right ideals but this will not be of concern to us here.

Proof. By Lemma 1.4 we have $\phi(0) = 0$ so $0 \in \ker\phi$ and $\phi(-x) = -\phi(x)$ for all $x \in R$. Hence if $x, y \in \ker\phi$ we have $\phi(x + (-y)) = \phi(x) + (-\phi(y)) = 0 - 0 = 0$ and so $x + (-y) \in \ker\phi$ and it is an additive subgroup by the subgroup test.

Now suppose $x \in \ker\phi$ and $r \in R$. Then $\phi(xr) = \phi(x)\phi(r) = 0\phi(r) = 0$ by Lemma 1.1, and similarly $\phi(rx) = 0$. It follows that $xr, rx \in \ker\phi$. The lemma is proved. □

There are two important operations on ideals: intersection and summation. If $I_1, \dots, I_k$ are ideals in R then we write $I_1 + \cdots + I_k$ for the sum of these sets, that is the set $\{x_1 + \cdots + x_k : x_i \in I_i \text{ for } 1 \leqslant i \leqslant k\}$. There is an infinite version of this[6] which is notationally more complicated but not otherwise more problematic.

Lemma 3.4. *Suppose that R is a ring. If $I_1, \dots, I_k$ are ideals of R then so is $I_1 + \cdots + I_k$. If $\mathcal{I}$ is a non-empty family of ideals of R, then $\bigcap_{I \in \mathcal{I}} I$ is an ideal of R.*

Proof. This is essentially just unpacking notation. Doing this for the intersection is easiest since $x \in \bigcap_{I \in \mathcal{I}} I$ if and only if $x \in I$ for all $I \in \mathcal{I}$. For the sum, it follows since addition is commutative. □

The above may be used to define the ideal generated by a set: if R is a ring and V is a subset of R then the ideal generated by V is[7]

$$\langle V \rangle := \bigcap \{I : V \subset I \text{ and } I \text{ is an ideal in } R\}.$$

Note that this intersection is well-defined since $V \subset R$ and R is an ideal in R, and $\langle V \rangle$ is an ideal by Lemma 3.4.

We give a few examples:

(i) If $n \in \mathbb{Z}$ then $\langle n \rangle$ is the set of multiples of n in $\mathbb{Z}$. We shall see later that every ideal in $\mathbb{Z}$ is of this form. ⚠If $n \neq 0$ then $\langle n \rangle = \mathbb{Q}$ in $\mathbb{Q}$.

(ii) For $\lambda \in R$, the ideal $\langle X - \lambda \rangle$ in $R[X]$ is the set of polynomials p with $p(\lambda) = 0$. We proved that the polynomials with λ as a root are in this ideal in the course of the proof of Proposition 2.3; the other direction follows from Lemma 1.1.

(iii) For $\mathbb{F}$ a field and $\lambda, \lambda' \in \mathbb{F}$ distinct, the ideal $\langle X - \lambda, X - \lambda' \rangle = \mathbb{F}[X]$ in $\mathbb{F}[X]$. This follows since any ideal containing $X - \lambda$ and $X - \lambda'$ must contain their difference $\lambda - \lambda'$ which is a unit in $\mathbb{F}$.

(iv) The ideal $\langle 2, X \rangle$ in $\mathbb{Z}[X]$ is the set of polynomials with even constant term. Certainly the polynomials with even constant term are an ideal in $\mathbb{Z}[X]$, and every such polynomial has the from $2q + Xp(X)$ for some $p \in \mathbb{Z}[X]$ and constant polynomial $q \in \mathbb{Z}[X]$, and hence all such polynomials are in this ideal.

Lemma 3.5. *Suppose that R is a non-trivial commutative unital ring. Then R is a field if and only if the only ideals in R are $\{0\}$ and R.*

[6] For $\mathcal{I}$ a set of ideals in R we put $\sum_{I \in \mathcal{I}} I := \bigcup_{S \subset \mathcal{I} : |S| < \infty} \sum_{I \in S} I$.

[7] We also write $\langle v_1, \dots, v_n, V_1, \dots, V_m \rangle := \langle \{v_1, \dots, v_n\} \cup V_1 \cup \cdots \cup V_m \rangle$ where $v_1, \dots, v_n \in R$ and $V_1, \dots, V_m \subset R$.

Proof. Suppose that R is a field, and I is an ideal with $x \in I$ non-zero. Then $x \in U(R)$ by definition and so there is some $y \in R$ such that $yx = 1$. Hence if $z \in R$ then $z = (zy)x \in I$ since $rx \in I$ for all $r \in R$. We conclude that $I = R$ as required.

In the other direction, suppose that $\{0\}$ and R are the only ideals. For $x \in R$ the set $xR := \{xr : r \in R\}$ is an ideal in R since R is commutative. If $x \in R^*$ then $xR \neq \{0\}$ and so $xR = R$ and hence there is some $y \in R$ such that $xy = 1$, and since $xy = yx$ we conclude $x \in U(R)$. On the other hand since R is non-trivial we have $U(R) \subset R^*$, and hence $U(R) = R$ and commutativity seals the deal: R is a field. □

Corollary 3.6. *Suppose that $\mathbb{F}$ is a field, R is a ring, and $\phi : \mathbb{F} \to R$ is a ring homomorphism. Then either ϕ is identically 0 or ϕ is injective.*

Proof. By Lemma 3.3 the kernel of ϕ is an ideal in $\mathbb{F}$, and hence by Lemma 3.5 we have $\ker\phi = \{0\}$ or $\ker\phi = \mathbb{F}$. If ϕ is not identically 0 then there is some $x \in \mathbb{F}$ with $x \notin \ker\phi$, so that $\ker\phi = \{0\}$ and ϕ is injective as claimed. □

It is worth checking that some of the classical examples of commutative unital rings are not secretly the same *i.e.* are not isomorphic. In fact something rather stronger is true: while the inclusion maps

$$\mathbb{Z} \hookrightarrow \mathbb{Q} \hookrightarrow \mathbb{R} \hookrightarrow \mathbb{C}$$

are injective unital ring homomorphisms, in the other direction there are only zero homomorphisms, and so no unital homomorphisms (since none of these rings is trivial).

(i) If $\phi : \mathbb{C} \to \mathbb{R}$ is a ring homomorphism then $\phi(i)^2 + \phi(1)^2 = \phi(0) = 0$ and so $\phi(1) = \phi(i) = 0$ and hence $\phi(z) = 0$ for all $z \in \mathbb{C}$ by Lemma 1.1.

(ii) If $\phi : \mathbb{R} \to \mathbb{Q}$ is a ring homomorphism then it is certainly not injective since $\mathbb{R}$ is uncountable, so by Corollary 3.6 we have $\phi(x) = 0$ for all $x \in \mathbb{R}$.

(iii) If $\phi : \mathbb{Q} \to \mathbb{Z}$ is a ring homomorphism then $\phi(1)(2\phi(1/2) - 1) = \phi(1)(\phi(1/2) + \phi(1/2)) - \phi(1) = 0$, but $2\phi(1/2) - 1$ is odd and so non-zero. Thus $\phi(1) = 0$ and hence $\phi(q) = 0$ for all $q \in \mathbb{Q}$ by Lemma 1.1.[8]

4. Quotient rings and the isomorphism theorems

Kernels of homomorphisms are a key source of ideals and, as the next proposition shows, all ideals arise in this way.

Proposition 4.1 (Quotient rings)**.** *Suppose that R is a unital ring, and I is an ideal in R. Then the set $R/I := \{a + I : a \in R\}$ may be given the structure of a unital ring such that $q : R \to R/I; a \mapsto a + I$ is a unital ring homomorphism.*

Proof. We should like to define addition and multiplication on R/I by

$$(a + I)\hat{+}(b + I) := (a + b) + I \text{ and } (a + I)\hat{\times}(b + I) := (ab) + I \tag{4.1}$$

[8] Alternatively, by Corollary 3.6 if ϕ is not identically 0 then ϕ is injective. Now, $\phi(1)(\phi(1) - 1) = 0$ and since $\mathbb{Z}$ is an integral domain $\phi(1) = 1$ and so ϕ is unital and by Lemma 1.4 we have $\phi(U(\mathbb{Q})) \leqslant U(\mathbb{Z}) = \{-1, 1\}$, which is a contradiction since ϕ is injective and $U(\mathbb{Q}) = \mathbb{Q}^*$ is infinite.

for $a, b \in R$. To see that this is well-defined suppose that $a + I = a' + I$ and $b + I = b' + I$. Then

$$(a + b) - (a' + b') = (a - a') + (b - b') \in I + I = I$$

since I is a group, and hence $(a + b) + I = (a' + b') + I$. Similarly

$$(ab) - (a'b') = (a(b + (-b'))) + ((a + (-a'))b') \in aI + Ib' = I + I = I,$$

by Lemma 1.1, then that I is a group, then the second property of ideals, and finally that I is a group again. It follows that $(ab) + I = (a'b') + I$.

We have shown that there are well-defined binary operations $\widehat{+}$ and $\widehat{\times}$ on R/I, and this is the substance of the argument. We complete the demonstration that R/I is a unital ring by saying that the remaining ring axioms are inherited from the corresponding axioms in R.

For clarity we record some details though these were unlectured and the blue text here will not be examined. If we define a unary operation $\widehat{-} : R/I \to R/I; x + I \mapsto (-x) + I$ and two constants $\widehat{0} := I$ and $\widehat{1} := 1 + I$ then R/I is a unital ring with addition $\widehat{+}$, multiplication $\widehat{\times}$, negation $\widehat{-}$, zero $\widehat{0}$, and multiplicative identity $\widehat{1}$ if (and only if)

(R1) $$U\widehat{+}(V\widehat{+}W) = (U\widehat{+}V)\widehat{+}W \text{ for all } U, V, W \in R/I;$$

(R2) $$\widehat{0}\widehat{+}U = U\widehat{+}\widehat{0} = U \text{ for all } U \in R/I;$$

(R3) $$U\widehat{+}(\widehat{-}U) = (\widehat{-}U)\widehat{+}U = \widehat{0} \text{ for all } U \in R/I;$$

(R4) $$U\widehat{+}V = V\widehat{+}U \text{ for all } U, V \in R/I;$$

(R5) $$U\widehat{\times}(V\widehat{\times}W) = (U\widehat{\times}V)\widehat{\times}W \text{ for all } U, V, W \in R/I;$$

(R6) $$U\widehat{\times}(V\widehat{+}W) = (U\widehat{\times}V)\widehat{+}(U\widehat{\times}W) \text{ for all } U, V, W \in R/I;$$

(R7) $$(U\widehat{+}V)\widehat{\times}W = (U\widehat{\times}W)\widehat{+}(V\widehat{\times}W) \text{ for all } U, V, W \in R/I;$$

(R8) $$\widehat{1}\widehat{\times}U = U\widehat{\times}\widehat{1} = U \text{ for all } U \in R/I.$$

These can be verified from the corresponding identities for R since for any $U, V, W \in R/I$ there are elements $x, y, z \in R$ with $U = x + I$, $V = y + I$ and $W = z + I$. Then, for example, associativity of $\widehat{+}$ (that is (R1)) follows from associativity of $+$ by noting that

$$\begin{aligned}
U\widehat{+}(V\widehat{+}W) &= (x + I)\widehat{+}((y + I)\widehat{+}(z + I)) && \text{Definition of } \widehat{+}\\
&= (x + I)\widehat{+}((y + z) + I) && \text{Definition of } \widehat{+}\\
&= (x + (y + z)) + I && \text{Associativity of } +\\
&= ((x + y) + z) + I && \text{Definition of } \widehat{+}\\
&= ((x + y) + I)\widehat{+}(z + I) && \text{Definition of } \widehat{+}\\
&= ((x + I)\widehat{+}(y + I))\widehat{+}(z + I)\\
&= (U\widehat{+}V)\widehat{+}W
\end{aligned}$$

and all the others in the same way as they have the same form.

Finally, q is a unital homomorphism since it is a homomorphism by the design of (4.1), and unital since $q(1) = \widehat{1} = 1 + I$ is the multiplicative identity of R/I. □

Note that the kernel of the projection map q is exactly the ideal I.

⚠ A similar result is true for rings that are not necessarily unital, and also groups, vector spaces, and modules which we shall encounter later in the course. This is because of the fact that all these structures can be defined in terms of some data – that is some operations and some constants – and some axioms which take the form of some equations in these operations and constants which hold for *all* values of the variables. Fields and integral domains do not (in general) have quotients that are fields or integral domains because their axiomatisation requires equations quantified over R^* as well as R (and in a sense this is unavoidable). Concretely, if $\mathbb{F}$ is a field and $I = \mathbb{F}$ then $\mathbb{F}/I$ is the trivial ring and so not a field.

One may think of the quotient of a ring by an ideal as the ring in which the elements of the ideal are set to 0. We consider some examples:

(i) The quotient $\mathbb{Z}/\langle n\rangle$ where $n \in \mathbb{N}$ is just the ring integers modulo n, often written as $\mathbb{Z}/n\mathbb{Z}$ when discussing groups. (For n composite this serves an an example of a quotient of an integral domain by an ideal which is not an integral domain.)

(ii) The elements of the ring $\mathbb{R}[X]/\langle X^2\rangle$ are the polynomials in $\mathbb{R}[X]$ with all quadratic and higher terms 'set to zero'. For $f \in \mathbb{R}[X]$ we write $f'(0)$ for the coefficient of X in f. We know from the definition of multiplication of polynomials how to work out $(fg)'(0)$ from f and g. It is also possible to do this by considering the equivalence class of polynomials where we set X^2 equal to 0. Specifically, since multiplication in $\mathbb{R}[X]/\langle X^2\rangle$ is well-defined we have

$$\begin{aligned}
&(fg)(0) + (fg)'(0)X + \langle X^2\rangle \\
&\qquad = (fg)(X) + \langle X^2\rangle \\
&\qquad = (f(0) + f'(0)X)(g(0) + g'(0)X) + \langle X^2\rangle \\
&\qquad = f(0)g(0) + (f(0)g'(0) + f'(0)g(0))X + f'(0)g'(0)X^2 + \langle X^2\rangle \\
&\qquad = f(0)g(0) + (f(0)g'(0) + f'(0)g(0))X + \langle X^2\rangle,
\end{aligned}$$

where the passage between the last lines is because $\langle X^2 + 1\rangle$ is a group and $f'(0)g'(0)X^2 \in \langle X^2\rangle$. We conclude that

$$(fg)(0) - f(0)g(0) + ((fg)'(0) - (f(0)g'(0) + f'(0)g(0)))X \in \langle X^2\rangle. \tag{4.2}$$

If the left hand side is not identically 0 then it has a degree which is at most 1. On the other hand, any element of $\langle X^2\rangle$ has the form $X^2q(X)$ for some $q \in \mathbb{R}[X]$. If the left hand side of (4.2) is not zero then q is not identically 0 by Lemma 1.1 and so $1 \geqslant \deg X^2 + \deg q \geqslant 2$, a contradiction. We conclude that $(fg)(0) = f(0)g(0)$ (as expected) and we also recover Leibniz's identity that $(fg)'(0) = (fg' + f'g)(0)$ (at least for polynomials).

(iii) The ring $\mathbb{R}[X]/\langle X^2 + 1\rangle$ can be thought of as $\mathbb{R}$ with an additional element $X + \langle X^2 + 1\rangle$ – more commonly denoted i – such that $(X + \langle X^2 + 1\rangle)^2 + 1 = 0$. It is a 'realisation' of $\mathbb{C}$; indeed, the map

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

is an isomorphism. First, suppose $z, w \in \mathbb{C}$ and write $z = a+bi$ and $w = c+di$ for $a,b,c,d \in \mathbb{R}$ so that

$$\begin{aligned}\psi(z+w) &= \psi((a+c)+(b+d)i)\\ &= (a+c)+(b+d)X+\langle X^2+1\rangle\\ &= (a+cX)+\langle X^2+1\rangle+(b+dX)+\langle X^2+1\rangle = \psi(z)+\psi(w).\end{aligned}$$

Since $zw = (ac-bd)+(bc+ad)i$ we have

$$\begin{aligned}\psi(zw) &= \psi(ac-bd+(bc+ad)i)\\ &= (ac-bd)+(bc+ad)X+\langle X^2+1\rangle\\ &= ac+(bc+ad)X+bdX^2-(X^2+1)bd+\langle X^2+1\rangle\\ &= (a+bX)(c+dX)+\langle X^2+1\rangle = \psi(z)\psi(w).\end{aligned}$$

Thus ψ is a ring homomorphism.

ψ is surjective: any element of $\mathbb{R}[X]/\langle X^2+1\rangle$ – recall an element in this case is a coset – contains a polynomial of minimal degree, say $q(X)$ with lead coefficient a, and if $\deg q \geqslant 2$ then $q(X) - aX^{\deg q-2}(X^2+1)$ has smaller degree and is in the same coset, so we know that this element of minimal degree has degree at most 1 *i.e.* is of the form $a+bX$. However, $\psi(a+bi) = a+bX+\langle X^2+1\rangle$ and so every coset on the codomain has a preimage.

Finally, ψ is injective: since it is surjective it is not identically 0 and so by Corollary 3.6 it is injective since $\mathbb{C}$ is a field.[9]. We conclude that ψ is a bijective ring homomorphism and so by Lemma 3.1 it is an isomorphism.

The First Isomorphism Theorem is a more general result by which we can access such isomorphisms.

Theorem 4.2 (First Isomorphism Theorem)**.** *Suppose that $\phi : R \to S$ is a unital homomorphism. Then $\phi(R)$ is a unital subring of S; $\ker\phi$ is an ideal in R; and the map*

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

is a well-defined injective unital homomorophism with image $\phi(R)$ i.e. *ψ is an injective unital homomorphism such that the following diagram commutes*

$$\begin{tikzcd} R \arrow[r, "q"] \arrow[dr, "\phi"] & R/\ker\phi \arrow[d, "\psi"] \\ & S \end{tikzcd}$$

[9] In lectures we proceeded by examining degree: if $a+bX \in \langle X^2+1\rangle$ then $a+bX = (X^2+1)p(X)$ for some $p \in \mathbb{R}[X]$, and if $(a,b) \neq (0,0)$ then $p(X) \in \mathbb{R}[X]^*$ and so $1 \geqslant \deg(a+bX) = \deg(X^2+1)+\deg p \geqslant 2$, a contradiction. Hence $\ker\phi = \{0\}$, and ϕ is injective as claimed.

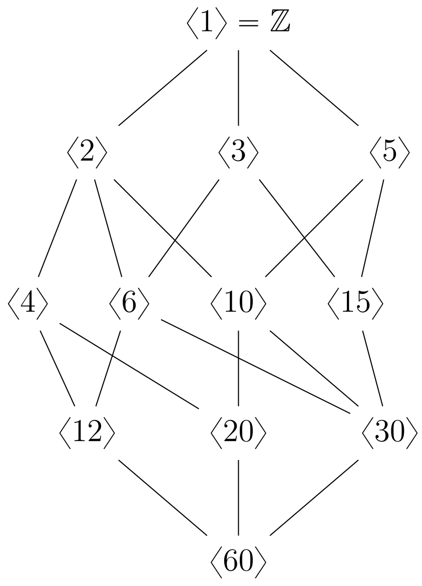

Figure 1. Lattice of ideals in $\mathbb{Z}/\langle 60\rangle$

Proof. The first two conclusions are Lemma 3.2 and Lemma 3.3 respectively. By Proposition 4.1 $R/\ker\phi$ is a unital ring.

Now, $x + \ker\phi = y + \ker\phi$ if and only if $x + (-y) \in \ker\phi$ which is true if and only if $\phi(x) + (-\phi(y)) = \phi(x + (-y)) = 0$ by Lemma 1.4, which in turn is true if and only if $\phi(x) = \phi(y)$. It follows that ψ is a well-defined injection; its image is clearly $\phi(R)$. ψ is a ring homomorphism since

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

and

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

Finally, $\psi(1 + \ker\phi) = \phi(1) = 1$ and so ψ a unital ring homomorphism. The result is proved. □

The ideals of a ring form a lattice as do the ideal in a ring containing a particular lattice. The figure shows the lattice of ideals in $\mathbb{Z}$ containing the ideal $\langle 60\rangle$. These ideals are in one to one correspondence with the ideals in $\mathbb{Z}/\langle 60\rangle$. The next theorem establishes this in general.

Theorem 4.3 (Relationship between ideals in R and R/I)**.** *Suppose that R is a ring and I is an ideal in R. Write $\mathcal{I}$ for the set of ideals in R containing I, and $\mathcal{J}$ for the set of* *ideals in R/I. Then the map*

$$\phi : \mathcal{I} \to \mathcal{J}; I' \mapsto \{x + I : x \in I'\}.$$

is a well-defined inclusion-preserving bijection.

Proof. First, we show the map is well-defined. Suppose that $I' \in \mathcal{I}$, and $S, T \in \phi(I')$. Then there are elements $x, y \in I'$ such that $S = x + I$ and $T = y + I$ so

$$S + (-T) = (x + I) + ((-y) + I) = (x + (-y)) + I \in \phi(I').$$

Since $\phi(I')$ is non-empty, the subgroup test $\phi(I')$ is an additive subgroup of R/I. Furthermore, if $x + I \in R/I$ and $y \in I'$ then

$$(x + I) \times (y + I) = (xy) + I \in \phi(I') \text{ and } (y + I) \times (x + I) = (yx) + I \in \phi(I')$$

since $xy, yx \in I'$. Thus $\phi(I')$ is genuinely an ideal in R/I.

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

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

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

In lectures we discussed the case $R = \mathbb{Z}$ and $I = \langle p \rangle$ for a prime p. Let $\mathcal{I}$ and $\mathcal{J}$ as defined in Theorem 4.3.

(i) If $I' \in \mathcal{I}$ and $I' \neq \langle p \rangle$ then let $x \in I' \backslash \langle p \rangle$. Since x is not a multiple of p – recall that $\langle p \rangle$ is exactly the multiples of p – then $\operatorname{hcf}(x, p) = 1$ and so by Bezout's Theorem there are $\alpha, \beta \in \mathbb{Z}$ such that $\alpha x + \beta p = 1$. But then

$$1 = \alpha x + \beta p \in \alpha I' + \beta\langle p \rangle \subset \alpha I' + \beta I' = I'.$$

However, $1 \in I'$ means $I' = \mathbb{Z}$. Hence $\mathcal{I} = \{\langle p \rangle, \mathbb{Z}\}$

(ii) On the other hand $\mathbb{F}_p = \mathbb{Z}/\langle p \rangle$ is a field and so by Lemma 3.5 the only ideals are the zero ideal and the whole field so $\mathcal{J} = \{\{0_{\mathbb{F}_p}\}, \mathbb{F}_p\}$.

Since $\mathbb{F}_p = \mathbb{Z}/\langle p \rangle$ we have $0_{\mathbb{F}_p} = \langle p \rangle$. The correspondence in Theorem 4.3 takes $\langle p \rangle$ to $\{\langle p \rangle\}$ and $\mathbb{Z}$ to $\{x + \langle p \rangle : x \in \mathbb{Z}\} = \mathbb{F}_p$.

⚠ The appearance of Bezout's Theorem in (i) should not be too much of a surprise: it is hiding in the fact that $\mathbb{F}_p$ is a field in the second part. The main component of proving this is showing that every $x \in \mathbb{F}_p^*$ has a multiplicative inverse. If $x \in \mathbb{F}_p^*$ then $x = x_0 + \langle p \rangle$ for some $x_0 \notin \langle p \rangle$. Thus $\operatorname{hcf}(x_0, p) = 1$ and hence by Bezout's Theorem there are $\alpha, \beta \in \mathbb{Z}$ such that $\alpha x_0 + \beta p = 1$, whence $(\alpha + \langle p \rangle)x = 1 + \langle p \rangle$ and x has an inverse as required.

5. The Chinese Remainder Theorem

Given a family $(R_i)_{i\in I}$ of unital rings we write $\prod_{i\in I} R_i$ (or $R_1\times\cdots\times R_k$ if $I=\{1,\dots,k\}$) for the direct product of these rings, that is the set[10] $\prod_{i\in I} R_i$ endowed with pointwise operations:

$$a+b:=(a_i+b_i)_{i\in I} \text{ and } ab:=(a_ib_i)_{i\in I} \text{ for all } a,b\in\prod_{i\in I} R_i.$$

This is a unital ring with $0_{\prod_i R_i}=(0_{R_i})_{i\in I}$, $1_{\prod_i R_i}=(1_{R_i})_{i\in I}$, and

$$U\left(\prod_{i\in I} R_i\right)=\prod_{i\in I} U(R_i). \tag{5.1}$$

The projection maps

$$\pi_j:\prod_{i\in I} R_i\rightarrow R_j; x\mapsto x_j$$

are all unital homomorphism; their existence is what really captures the product structure.

⚠ The maps $\iota_j:R_j\rightarrow\prod_{i\in I} R_i$ (which are defined so that $\pi_j(\iota_j(x))=x$ and $\pi_i(\iota_j(x))=0_{R_i}$ if $i\neq j$) are ring homomorphisms, but they are not in general unital ring homomorphisms.

⚠ Direct products do *not* preserve the property of being an integral domain: in the ring $\mathbb{R}\times\mathbb{R}$ we have $(a,0)\times(0,b)=0$ for all $a,b\in\mathbb{R}$.

We say that ideals I and J in a ring R are **coprime** if $I+J=R$. To explain the terminology recall that Bezout's Theorem can be phrased as saying that if $R=\mathbb{Z}$ then $1\in\langle x\rangle+\langle y\rangle$ if (and only if) x and y are coprime.

Theorem 5.1 (Chinese Remainder Theorem). *Suppose that R is a commutative unital ring and $I_1,\dots,I_k$ are pairwise coprime ideals in R. Then the map*

$$R/(I_1\cap\cdots\cap I_k)\rightarrow(R/I_1)\times\cdots\times(R/I_k); r+I_1\cap\cdots\cap I_k\mapsto(r+I_1,\dots,r+I_k)$$

is an isomorphism.

Proof. It is enough to show that the map

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

is a surjective unital homomorphism with kernel $I_1\cap\cdots\cap I_k$. The result then follows by the First Isomorphism Theorem. Quotient maps are all unital homomorphisms and so is this map. The kernel is exactly the set of $r\in R$ such that $r+I_i=I_i$ for all $1\leqslant i\leqslant k$ which is to say $\ker\phi=I_1\cap\cdots\cap I_k$ as required.

[10] This is the Cartesian product. We take it to be the set of functions $f:I\rightarrow\bigcup_{i\in I} R_i$ such that $f(i)\in R_i$ for all $i\in I$. Such functions are sometimes called choice functions in the literature, the idea being that for each $i\in I$ we choose some $f(i)\in R_i$. If I is an initial segment of the natural numbers we often write f_i instead of $f(i)$.

Proving that the map is surjective is the rub and is perhaps most easily done in the $k = 2$ case first. In general, note that[11]

$$R = \bigcap_{i \neq j} (I_j + I_i) = I_j + \bigcap_{i \neq j} I_i,$$

so we can take $x_j \in I_j$ and $y_j \in \bigcap_{i \neq j} I_i$ with $x_j + y_j = 1$. For $u \in (R/I_1) \times \cdots \times (R/I_k)$ we have

$$\phi\left(u_1 y_1 + \cdots + u_k y_k\right) = (u_1 y_1 + I_1, \dots, u_k y_k + I_k)$$

and the map is surjective as required. □

This result immediately gives the usual formulation where we are trying to solve simultaneous congruences: if $m_1, \dots, m_k$ are pairwise coprime naturals and $a_1, \dots, a_k \in \mathbb{Z}$ then there is some $a \in \mathbb{Z}$ such that $a \equiv a_i \pmod{m_i}$ for $1 \leqslant i \leqslant k$.

Similarly, Euler's totient function is $\phi(n) := |U(\mathbb{Z}/\langle n \rangle)|$ and the above coupled with (5.1) shows that this function is multiplicative, meaning $\phi(mn) = \phi(m)\phi(n)$ whenever $\mathrm{hcf}(m, n) = 1$.

⚠ ϕ is *not* totally multiplicative, for example $\phi(4) = 2 \neq 1 = \phi(2)^2$.

Corollary 5.2 (Polynomial interpolation)**.** *Suppose that $\mathbb{F}$ is a field, $\lambda_1, \dots, \lambda_k \in \mathbb{F}$ are pairwise distinct and $a_1, \dots, a_k \in \mathbb{F}$. Then there is a polynomial $p \in \mathbb{F}[X]$ of degree at most $k - 1$ such that $p(\lambda_i) = a_i$ for all $1 \leqslant i \leqslant k$.*

Proof. Since $\lambda_i \neq \lambda_j$ and $\mathbb{F}$ is a field we see that $(\lambda_j - \lambda_i)^{-1}(X - \lambda_i) - (\lambda_j - \lambda_i)^{-1}(X - \lambda_j) = 1$ and hence the ideals $(\langle X - \lambda_i \rangle)_{i=1}^k$ are pairwise coprime. Write $I := \bigcap_{i=1}^k \langle X - \lambda_i \rangle$ and apply the Chinese Remainder Theorem to $\mathbb{F}[X]$ to see that there is a polynomial $q \in \mathbb{F}[X]/I$ such that $q(X) \in a_i + \langle X - \lambda_i \rangle$ for all $1 \leqslant i \leqslant k$.

Let $p(X) \in q(X) + I$ have minimal degree so that $p(X) \in q(X) + I \subset a_i + \langle X - \lambda_i \rangle$ for all $1 \leqslant i \leqslant k$. If $d := \deg p$ has $d \geqslant k$ then write a_d for the lead coefficient of p and note that $p(X) - a_d X^{d-k} \prod_{i=1}^k (X - \lambda_i) \in q(X) + I$ and has strictly smaller degree. The result is proved. □

⚠Note that the obvious extension of Theorem 5.1 to infinitely many rings fails: if $R = \mathbb{Z}$ and $I_i := \langle p_i \rangle$ where p_i is the ith prime then the product $\prod_i (\mathbb{Z}/\langle p_i \rangle)$ is uncountable, but any quotient of $\mathbb{Z}$ is countable so there cannot be a surjection from a quotient of $\mathbb{Z}$ to this product.

[11] ⚠The second equality here, while true, is misleading as it makes use of the coprimality condition and is *not* true for general ideals. Specifically, since I_j is coprime to I_i for all $i \neq j$, there are elements $z_i \in I_j$ and $w_i \in I_i$ with $z_i + w_i = 1$. Thus $1 = \left(1 - \prod_{i \neq j} (1 - z_i)\right) + \left(\prod_{i \neq j} w_i\right) \in I_j + \bigcap_{i \neq j} I_i$. Thanks to Terry Song for asking for more explanation here. These details are not needed in the $k = 2$ case and I do not regard the details for $k > 2$ in this footnote as bookwork for the exam.

For a specific example of ideals I, J, and K with $(I + J) \cap (I + K) \neq I + J \cap K$ consider $R = \mathbb{Z}[X]$, $I = \langle 2 \rangle$, $J = \langle X + 1 \rangle$ and $K = \langle X - 1 \rangle$. Here $X + 1 \in I + J = I + K$ whereas $J \cap K = \langle X^2 - 1 \rangle$ and so if $X + 1 \in I + J \cap K$ then $X + 1 = p(X)(X^2 - 1) + 2q(X)$ for $p, q \in \mathbb{Z}[X]$. Degree considerations show that $p \equiv 0$ and then $2 \mid X + 1$ which is a contradiction.

⚠ Take care with the meaning of coprime for ideals: there is no non-unit $q(X)$ in $\mathbb{Z}[X]$ such that $X-1$ and $X+1$ are multiples of $q(X)$, but the ideals $\langle X-1\rangle$ and $\langle X+1\rangle$ are *not* coprime. This is reflected in the failure of Corollary 5.2 if $\mathbb{F}$ is replaced by $\mathbb{Z}$ where, for any $p(X) \in \mathbb{Z}[X]$ we must have $2 \mid p(1)-p(-1)$ so we cannot specify the value of these two points arbitrarily.

6. The Integers and characteristic

The ideal structure of the integers is well-behaved. We say that an ideal in a ring R is **principal** if it is generated by one element.

Proposition 6.1. *Every ideal in $\mathbb{Z}$ is principal.*

Proof. Suppose that I is a non-zero ideal and let $\mu > 0$ be its smallest positive element. If $I \neq \langle \mu\rangle$ then there is a minimal positive $\nu \in I \backslash \langle \mu\rangle$. By minimality of μ we have $\nu > \mu$. By the ideal property of I we have $\nu - \mu \in I$, and by minimality of ν we have $\nu - \mu \in \langle \mu\rangle$ and hence $\nu \in \langle \mu\rangle$, a contradiction. The result is proved. □

The integers play a uniquely important role amongst unital rings:

Proposition 6.2. *Suppose that R is a unital ring. Then there is a unique unital ring homomorphism $\phi : \mathbb{Z} \to R$.*

Proof. For existence we define ϕ recursively on $\mathbb{N}_0$ by $\phi(0) = 0$ and $\phi(n+1) := \phi(n)+1$ for $n \in \mathbb{N}_0$, and then put $\phi(-n) := -\phi(n)$ for $n \in \mathbb{N}$. Certainly $\phi(1) = 1$, and we can use induction to show that ϕ is a ring homomorphism.

In the other direction if ϕ and ψ are unital ring homomorphisms we can show $\phi(x) = \psi(x)$ for all $x \in \mathbb{N}_0$ by induction (since $\phi(1) = 1 = \psi(1)$), which extends to the whole of $\mathbb{Z}$ since $\phi(-x) = -\phi(x)$ and $\psi(-x) = -\psi(x)$ by Lemma 1.4. □

⚠ A function f is said to be right cancellable if whenever $g \circ f = h \circ f$ we have $g = h$. It can be shown that a function is right cancellable if and only if it is surjective. This remains true if we restrict f, g, and h to be linear maps between vector spaces; or homomorphisms between groups; or continuous maps between compact Hausdorff spaces; amongst many other things.

By contrast, let $f : \mathbb{Z} \to \mathbb{Q}$ be the inclusion map (which is an injective unital homomorphism), and suppose $g, h : \mathbb{Q} \to R$ are unital ring homomorphisms with $g \circ f = h \circ f$. Then for all $a \in \mathbb{Z}$ and $b \in \mathbb{Z}^*$ we have $g(ab^{-1}) = g(a)g(b)^{-1} = h(a)h(b)^{-1} = h(ab^{-1})$ so $g = h$ despite the fact that f is *not* surjective.

Proposition 6.2 tells us that for any unital ring R there is a unique unital homomorphism $\phi : \mathbb{Z} \to R$. This map has a kernel which is an ideal by Lemma 3.3 and principal by Proposition 6.1, say $\ker\phi = \langle x\rangle$. If $\langle y\rangle = \ker\phi$ then we have $x \mid y$ and $y \mid x$ and so $x = \pm y$, thus there is a unique element of $n \in \mathbb{N}_0$ such that $\ker\phi = \langle n\rangle$. This is called the **characterstic** of the ring R.

The inclusion map from $\mathbb{Z}$ into $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$ is a unital ring homomorphism in each case, and hence *the* unital ring homomorphism and we see that the characteristic at these rings is 0.

Proposition 6.3. *Suppose that R is an integral domain of non-zero characteristic. Then R has prime characteristic p and is a vector space over $\mathbb{F}_p$.*

Proof. Let $\phi : \mathbb{Z} \to R$ be the unital homomorphism of Proposition 6.2 and suppose that the characteristic is p. If $p = ab$ for $a, b \geqslant 1$ then $0 = \phi(p) = \phi(a)\phi(b)$ and since R is an integral domain we conclude that $\phi(a) = 0$ or $\phi(b) = 0$; say the former. Then $a \in \langle p \rangle$ whence $a = 0$ or $a \geqslant p$. It must be the latter and hence p is prime.

The First Isomorphism Theorem gives an injective unital homomorphism $\mathbb{Z}/\langle p \rangle \to R$, and so R is a vector space over $\mathbb{F}_p = \mathbb{Z}/\langle p \rangle$ as required. □

In view of Theorem 2.5 this means that any integral domain R is of prime characteristic p and sits between two fields of characteristic p.

While integral domains of characteristic 0 need not be vector spaces, a field of characteristic 0 is a vector space over $\mathbb{Q}$ by Theorem 2.5.

7. Prime and maximal ideals

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

Lemma 7.1. *Suppose that R is a commutative unital ring. Then I is proper if and only if R/I is non-trivial.*

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

For example, if R is an integral domain then $\langle X \rangle$ is prime in $R[X]$. To see this note that $p \in \langle X \rangle$ if and only if $p(0) = 0$, whence the primality of $\langle X \rangle$ follows from the fact that R is an integral domain. This is a close connection which manifests more generally:

Proposition 7.2. *Suppose that R is a commutative unital ring and I is a proper ideal. Then I is a prime ideal if and only if R/I is an integral domain.*

Proof. Note that $a + I, b + I \in R/I$ has $(a + I)(b + I) = (ab) + I = I$ if and only if $ab \in I$.

$\Rightarrow$: $(a + I)(b + I) = 0_{R/I} = I$ implies $ab \in I$ implies, and $a \in I$ 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. (Note R/I is non-trivial since I is proper.)

$\Leftarrow$: If $ab \in I$ then $(a+I)(b+I) = I = 0_{R/I}$ and hence $a+I = 0_{R/I} = I$ or $b+I = 0_{R/I} = I$ so $a \in I$ or $b \in I$ *i.e.* I is prime. □

⚠Note that R is an integral domain if and only if $\{0\}$ is prime.

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 R.

Proposition 7.3. *Suppose that R is a commutative unital ring and I is a proper ideal in R. Then I is maximal ideal if and only if R/I is a field.*

Proof. By Theorem 4.3 there is an ideal $I \subsetneq J \subsetneq R$ if and only if there is an ideal $\{0\} \subsetneq \tilde{J} \subsetneq R/I$. The result follows from Lemma 3.5 since R/I is non-trivial. □

It follows immediately from this and Proposition 7.2 that every maximal ideal is prime, but this can also be proved directly.

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

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

Theorem 7.4. *Suppose that R is a commutative unital ring and I is a proper ideal in R. Then there is a maximal ideal J in R containing I.*

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

We say that an element $x \in R$ is **prime** if $\langle x\rangle$ is a prime ideal. On the face of it this seems different to the 'usual' notion of prime in the naturals when they are considered as elements of the ring of integers. To explain the connection we shall need a little more notation.

In a commutative unital ring, we have $\langle x\rangle = \{xr : r \in R\}$, which we sometimes write as xR (or Rx).

⚠In the commutative (but not unital) ring $2\mathbb{Z}$, all the elements of $\{2r : r \in 2\mathbb{Z}\}$ are divisible by 4 and so, in particular, this set does not contain 2, and we have $\langle 2\rangle \neq \{2r : r \in 2\mathbb{Z}\}$.

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

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

Note that for all $x \in R$ we have $x \mid 0$ (including $0 \mid 0$) and $u \mid x$ for all $u \in U(R)$. We say that a and b are **associates** and write $a \sim b$ if $\langle a\rangle = \langle b\rangle$, which is trivially an equivalence relation.

We say that $a \in R^*$ is **irreducible** if either of the following equivalent properties holds:

(i) $\langle a\rangle$ is maximal amongst proper principal ideals;
(ii) whenever $x \mid a$ we have $x \sim a$ xor $x \sim 1$.

⚠Note that units are *not* irreducible since the ideal generated by a unit is not proper.

Lemma 7.5. *Suppose that R is an integral domain. Then*

(i) $a \sim b$ if and only if there is some $x \in U(R)$ such that $a = xb$;
(ii) $a \in R^*$ is irreducible if and only if whenever $a = xy$ we have $x \sim 1$ or $y \sim 1$;
(iii) $a \in R^*$ is irreducible if and only if whenever $a = xy$ we have $x \sim a$ or $y \sim a$;
(iv) if $a \in R^*$ is prime then it is irreducible.

Proof. $\Leftarrow$ from (i): If $x \in U(R)$ then by closure $xR \subset R$, and if $z \in R$ implies $(zx^{-1})x \in Rx$, whence $R = Rx$ and hence $Ra = Rb$.

$\Rightarrow$ from (i): If $Ra = Rb$ then there are $x, y \in R$ such that $a = xb$ and $b = ya$, whence $a = xya$ and by the Cancellation Lemma $xy = 1$ so $x \in U(R)$.

$\Leftarrow$ from (ii): Suppose that $\langle a\rangle \subset \langle x\rangle$ then $a = xy$ for some $y \in R$ and either $x \sim 1$, or $y \sim 1$ whence $x \sim a$ by (i).

$\Rightarrow$ from (ii): Suppose a is irreducible and $a = xy$. Then $\langle a\rangle \subset \langle y\rangle$ and either $y \sim 1$ (and we are done) or $y \sim a$. In the latter case by (i) we have $a = zy$ for some $z \in U(R)$ and hence $zy = xy$ so by the Cancellation Lemma $z = x$ and hence $x \sim 1$ by (i).

The proof of (iii) is similar to the proof of (ii).

For (iv), suppose that a is prime and $a = yz$ for $y, z \in R$. Then $a \mid yz$ and so by primality, either $a \mid y$ meaning $\langle y\rangle \subset \langle a\rangle \subset \langle y\rangle$ and $y \sim a$, or $a \mid x$ and the same argument gives $x \sim a$. The result follows from (iii). □

⚠ The right to left implication in part (i) of the Lemma is true in any commutative unital ring, but the left to right implication may fail if R is not an integral domain: Consider the ideal $I = \langle Z - XYZ\rangle$ in the ring $\mathbb{F}[X, Y, Z]$ viewed as polynomials in Z with coefficients in $\mathbb{F}[X, Y]$. Then we may think of the (commutative unital) ring $R := \mathbb{F}[X, Y, Z]/I$ as the polynomials in Z with constant coefficient from $\mathbb{F}[X, Y]$ and all other coefficients from $\mathbb{F}[X, Y]/\langle 1 - XY\rangle$. In R we have $\langle Z\rangle = \langle YZ\rangle$. But if $u \in U(R)$ then $u = a + Zb$ where $b \in R$ and $a \in U(\mathbb{F}[X, Y]) = \mathbb{F}^*$, so if $Z = uYZ \pmod{I}$ then equating coefficients of Z we have $1 - XY \mid 1 - aY$ in $\mathbb{F}[X, Y]$ which is a contradiction when we look at the X-degree. We conclude that there is no unit $u \in U(R)$ such that $Z = uYZ \pmod{I}$.

⚠ Even in integral domains, irreducible elements need not be primes: The ring $\mathbb{F}[X^2, X^3]$, which is the unital subring of $\mathbb{F}[X]$ consisting of polynomials whose coefficient of X is 0, is an integral domain. In this ring X^3 is irreducible, but $(X^3)^2 \in \langle X^2\rangle$ while $X^3 \notin \langle X^2\rangle$.

An integral domain is said to be a **principal ideal domain** or **PID** if every ideal is principal. Every field is a PID, and we saw in Proposition 6.1 that $\mathbb{Z}$ is also a PID. In PIDs we have the following complement to Lemma 7.5 part (iv).

Proposition 7.6. *Suppose that R is a PID and $x \in R^*$. Then x is irreducible if and only if $R/\langle x\rangle$ is a field. In particular, any non-zero prime ideal is maximal.*

Proof. $\langle x\rangle$ is maximal amongst proper principal ideals if and only if $\langle x\rangle$ is maximal amongst all proper ideals in R (since R is a PID) which is true if and only if $R/\langle x\rangle$ is a field by Proposition 7.3.

For the last part if I is a non-zero prime ideal then since R is a PID, $I = \langle x\rangle$ for some $x \in R^*$. Hence x is irreducible by Lemma 7.5 (iv), and so $\langle x\rangle$ is maximal amongst proper principal ideals, but these are the only proper ideals in R so I is maximal amongst all proper ideals. □

Since Proposition 6.1 established that $\mathbb{Z}$ is a PID, we have from Proposition 7.2 and Proposition 7.6 that $n \in \mathbb{N}$ is prime in the old sense if and only if it is prime in the new sense.

Proposition 7.6 immediately explains our existing supply of finite fields: the fields $\mathbb{F}_p$ of the integers mod p are all quotients of the principal ideal of integers by an irreducible. To get more we need some more PIDs; the proof that $\mathbb{Z}$ is a PID in Proposition 6.1 adapts to give the following.

Proposition 7.7. *Suppose that $\mathbb{F}$ is a field. Then $\mathbb{F}[X]$ is a PID.*

Proof. Suppose that I is an ideal in $\mathbb{F}[X]$. We may assume that it is non-zero and since the units of $\mathbb{F}[X]$ are the elements of $\mathbb{F}^*$ we may take $p \in I$ a monic polynomial of minimum degree. If I is not principal then there is an element $q \in I \backslash \langle p \rangle$, also monic, and of minimal degree (in this complement). By minimality of the degree of p we have $\deg p \leqslant \deg q$. Since I is an ideal $q(X) - p(X)X^{\deg q - \deg p} \in I$, and since p and q are monic this difference has degree less than q. By minimality of q it follows that $q(X) - p(X)X^{\deg q - \deg p} \in \langle p \rangle$, but then $q \in \langle p \rangle$ – a contradiction. □

8. Fields and adjunction of elements

We say that a field $\mathbb{F}$ is a **subfield** of a field $\mathbb{K}$ or $\mathbb{K}$ is a **field extension** of $\mathbb{F}$ if $\mathbb{F}$ is a unital subring of $\mathbb{K}$.[12] In this situation $\mathbb{K}$ has the structure of a vector space over $\mathbb{F}$ and we call its $\mathbb{F}$-dimension the **degree** of the field extension, also denoted $|\mathbb{K} : \mathbb{F}|$.

Theorem 8.1. *Suppose that $\mathbb{F}$ is a field and $f \in \mathbb{F}[X]$ is irreducible of degree d. Then $\mathbb{K} := \mathbb{F}[X]/\langle f \rangle$ is a degree d field extension of $\mathbb{F}$, there is $\alpha \in \mathbb{K}$ such that $\mathbb{K} = \mathbb{F}[\alpha]$ and the set of $\mathbb{F}$-polynomials with α as a root is the ideal generated by f.*

Proof. Proposition 7.7 tells us $\mathbb{F}[X]$ is a PID and hence Proposition 7.6 tells us that $\mathbb{F}[X]/\langle f \rangle$ is a field. The map $\mathbb{F} \to \mathbb{F}[X] \to \mathbb{F}[X]/\langle f \rangle$ is a composition of the embedding of $\mathbb{F}$ as the constant functions of a polynomial ring, and then the quotient map. This is a unital homomorphism which is injective since f is non-constant (being maximal), so $\mathbb{K}$ is an $\mathbb{F}$-vector space. Let $\alpha := X + \langle f \rangle$. The set of $\mathbb{F}$-polynomials with α as a root is an ideal and since $\mathbb{F}[X]$ is a PID it is generated by some g. Since $f(\alpha) = f(X) + \langle f \rangle = 0$ we see that $f \in \langle g \rangle$, but f is irreducible so $\langle f \rangle = \langle g \rangle$.

The elements $1, \alpha, \ldots, \alpha^{d-1}$ are $\mathbb{F}$-independent in $\mathbb{K}$; if they were not then there would be a polynomial $g \in \mathbb{F}[X]^*$ of degree at most $d-1$ such that $g(\alpha) = 0$, but all non-zero polynomials with this property are in the ideal generated by f and so have degree at least d.

By induction $\alpha^n \in \operatorname{Span}(1, \alpha, \ldots, \alpha^{d-1})$ for all $n \geqslant d$, so $\mathbb{K}$ is a degree d extension and $\mathbb{K} = \mathbb{F}[\alpha]$. □

We think of $\mathbb{K}$ as the field $\mathbb{F}$ with the element α **adjoined**.

[12] Note that not all unital subrings of fields are subfields *e.g.* $\mathbb{Z}$ is a unital subring of $\mathbb{C}$, but $\mathbb{Z}$ is not a field.

Suppose that $\mathbb{K}$ is a field extension of $\mathbb{F}$ and $\alpha \in \mathbb{K}$. The set $\{f \in \mathbb{F}[X] : f(\alpha) = 0\}$ is an ideal in $\mathbb{F}[X]$. If it is non-trivial we say that α is **$\mathbb{F}$-algebraic**, and since $\mathbb{F}[X]$ is a PID there is a unique monic generator, which we call the **minimal polynomial** of α. Note that if $|\mathbb{K} : \mathbb{F}| = d$ then $1, \alpha, \dots, \alpha^d$ must be linearly dependent for any $\alpha \in \mathbb{K}$, so *every* such α is $\mathbb{F}$-algebraic.

All degree one polynomials in $\mathbb{F}[X]$ are irreducible, but in view of Theorem 8.1 they do not give us any new fields. A quadratic is irreducible if and only if it does not have a root in $\mathbb{F}$ which leads to a couple of examples:

(i) $X^2 + 1$ is irreducible over $\mathbb{R}$. Hence $\mathbb{R}[X]/\langle X^2 + 1\rangle$ is a field, as we saw directly in example (iii) after Proposition 4.1.

(ii) Suppose that p is an odd prime. The map[13] $U(\mathbb{F}_p) \to U(\mathbb{F}_p); x \mapsto x^2$ is not injective since $(-1)^2 = 1^2$ (and $-1 \neq 1$ in for odd p), but the domain and codomain are finite and of the same size, so the map is not surjective. Thus there is some $a_p \in \mathbb{F}_p$ such that $q(X) := X^2 - a_p$ has no roots over $\mathbb{F}_p$, and hence $\mathbb{F}_p[X]/\langle X^2 - a_p\rangle$ is a field of order p^2.

(iii) Any quadratic $q \in \mathbb{F}_2[X]$ must have the form $q(X) = X^2 + aX + b$. q is irreducible if and only if $q(0) = q(1) = 1$, whence $X^2 + X + 1$ is the only quadratic irreducible in $\mathbb{F}_2[X]$. $\mathbb{F}_2[X]/\langle X^2 + X + 1\rangle$ is then a field of order 4.

⚠The field in this last example is the unique (up to isomorphism) field with 4 elements and is denoted $\mathbb{F}_4$. It is *not* equal to the ring $\mathbb{Z}/4\mathbb{Z}$ – $2 \times 2 = 0$ in the latter.

Corollary 8.2. *Suppose that $\mathbb{F}$ is a finite field extension of $\mathbb{R}$. Then $\mathbb{F}$ has degree at most* 2.

Proof. Suppose that $\alpha \in \mathbb{F}\backslash\mathbb{R}$. Since $\mathbb{F}$ is a finite extension of $\mathbb{R}$, α is $\mathbb{R}$-algebraic and it has a minimal polynomial $m \in \mathbb{R}[X]$. By the Fundamental Theorem of Algebra m is either linear or quadratic; since $\alpha \notin \mathbb{R}$ it is quadratic. In particular there is some element $\beta \in \mathbb{F}$ such that $\beta^2 + 1 = 0$ and hence $\mathbb{C}$ is (isomorphic) to a subfield of $\mathbb{F}$. $\mathbb{F}$ is a finite extension of $\mathbb{C}$ (since it is a finite extension of $\mathbb{R}$). However, if $\alpha \in \mathbb{F}\backslash\mathbb{C}$ then it is $\mathbb{C}$-algebraic and so has a minimal polynomial $m' \in \mathbb{C}[X]$. Again, by the Fundamental Theorem of Algebra m' is linear, contradicting the fact that $\alpha \notin \mathbb{C}$. The result is proved. □

⚠Note that the finiteness here is critical: $\mathbb{R}(X)$ is a field extension of $\mathbb{R}$ that is certainly not finite.

Theorem 8.3 (Tower Law)**.** *Suppose that $\mathbb{L}$ is a field extension of $\mathbb{K}$ and $\mathbb{K}$ is a field extension of $\mathbb{F}$. Then $|\mathbb{L} : \mathbb{F}| = |\mathbb{L} : \mathbb{K}||\mathbb{K} : \mathbb{F}|$.*

Proof. Let $e_1, \dots e_n$ be a basis for $\mathbb{K}$ as a vector space over $\mathbb{F}$ and $f_1, \dots, f_m$ be a basis for $\mathbb{L}$ as a vector space over $\mathbb{K}$. We shall show that $(e_i f_j : 1 \leqslant i \leqslant n, 1 \leqslant j \leqslant m)$ is a basis for $\mathbb{L}$ as a vector space over $\mathbb{F}$. There are two things to check:

[13] In fact it is a homomorphism.

Independence: Suppose that $\sum_{i,j} \lambda_{i,j} e_i f_j = 0$ for some $\lambda_{i,j} \in \mathbb{F}$. Then

$$\sum_{j=1}^{m} \left(\sum_{i=1}^{n} \lambda_{i,j} e_i \right) f_j = 0,$$

but each coefficient of f_j is an element of $\mathbb{K}$, and so by linear independence of $(f_1, \dots, f_m)$ we see that $\sum_{i=1}^{n} \lambda_{i,j} e_i = 0$ for all $1 \leqslant j \leqslant m$. But then by linear independence of $(e_1, \dots, e_n)$ we see that $\lambda_{i,j} = 0$ for all $1 \leqslant i \leqslant n$ and $1 \leqslant j \leqslant m$.

Spanning: If $x \in \mathbb{L}$ then since $(f_1, \dots, f_m)$ is a basis for $\mathbb{L}$ over $\mathbb{K}$ we have elements $\mu_1, \dots, \mu_m \in \mathbb{K}$ such that $x = \mu_1 f_1 + \cdots + \mu_m f_m$. Since $(e_1, \dots, e_n)$ is a basis for $\mathbb{K}$ over $\mathbb{F}$, for each $1 \leqslant i \leqslant m$ we have $\lambda_{i,1}, \dots, \lambda_{i,n}$ such that $\mu_i = \lambda_{i,1} e_1 + \cdots + \lambda_{i,n} e_n$ and hence

$$x = \sum_{i=1}^{m} \mu_i f_i = \sum_{i,j} \lambda_{i,j} e_i f_j$$

as required. □

⚠ The polynomial $X^3 + X + 1$ is irreducible[14] so $\mathbb{L} := \mathbb{F}_2[X]/\langle X^3 + X + 1\rangle$ is a field of order 8. However, $\mathbb{L}$ does not have a subfield of order 4: suppose it did, and call it $\mathbb{K}$. Since $1 \in \mathbb{K}$, we have that $\mathbb{F}_2$ is a subfield of $\mathbb{K}$. However, $|\mathbb{K} : \mathbb{F}_2| = 2$ and $|\mathbb{L} : \mathbb{F}_2| = 3$ and the Tower Law then gives us $|\mathbb{L} : \mathbb{K}| \times 2 = 3$, a contradiction.

9. Irreducibility tests for polynomials

Suppose that $\phi : R \to S$ is a unital homomorphism between two integral domains. Then

$$\tilde{\phi} : R[X] \to S[X]; a_d X^d + \cdots + a_1 X + a_0 X \mapsto \phi(a_d) X^d + \cdots + \phi(a_1) X + \phi(a_0) \tag{9.1}$$

is also a unital homomorphism between integral domains. We can use this homomorphism to examine irreducibility in $S[X]$ and $R[X]$ through each other. We begin with the case when ϕ is the embedding map from Theorem 2.5 of an integral domain into its field of fractions.

⚠ $2X$ is reducible in $\mathbb{Z}[X]$ but irreducible in $\mathbb{Q}[X]$. We say that $f \in \mathbb{Z}[X]$ is **primitive** if there is no prime p dividing all of the coefficients of f.

Theorem 9.1 (Gauss' Lemma)**.** *A non-constant polynomial $f \in \mathbb{Z}[X]$ is irreducible in $\mathbb{Z}[X]$ if and only if it is primitive, and irreducible in $\mathbb{Q}[X]$.*

Proof. $\Leftarrow$: Suppose that f is primitive and irreducible in $\mathbb{Q}[X]$. Write $f = gh$ for $g, h \in \mathbb{Z}[X]$. Since f is irreducible in $\mathbb{Q}[X]$ we see that either $\deg g = 0$ or $\deg h = 0$, and since f is primitive we then conclude that either g or h is ± 1.

$\Rightarrow$: Suppose that f is irreducible in $\mathbb{Z}[X]$, and $f = gh$ for some $g, h \in \mathbb{Q}[X]$. Let $\lambda \in \mathbb{N}$ be minimal such that there is some $q \in \mathbb{Q}^*$ with $\lambda q^{-1} g \in \mathbb{Z}[X]$ and $qh \in \mathbb{Z}[X]$. Let $q \in \mathbb{Q}^*$ be such that $g' := \lambda q^{-1} g \in \mathbb{Z}[X]$ and $h' := qh \in \mathbb{Z}[X]$. Suppose that p is a prime dividing λ. Then $p \mid g'h'$ and applying the modulo p reduction map to $g'h'$ we get $(g'$

[14] There are only two irreducible cubics in $\mathbb{F}_2[X]$, with the other being $X^3 + X^2 + 1$. To see this note that any reducible cubic p has a linear factor *i.e.* $p(0) = 0$ or $p(1) = 0$.

$\pmod p)(h' \pmod p) = 0$ in the integral domain $\mathbb{F}_p[X]$. Hence $g' \equiv 0 \pmod p$ or $h' \equiv 0 \pmod p$, but then $p \mid \lambda q^{-1}g$ or $p \mid qh$ (in $\mathbb{Z}[X]$) contradicting minimality of λ since either $(\lambda p^{-1})q^{-1}g \in \mathbb{Z}[X]$ and $qh \in \mathbb{Z}[X]$, or $(\lambda p^{-1})(qp^{-1})^{-1} \in \mathbb{Z}[X]$ and $(qp^{-1})h \in \mathbb{Z}[X]$. We conclude that $\lambda = 1$ and so g or h is a unit in $\mathbb{Z}[X]$ and hence in $\mathbb{Q}[X]$ as required. □

In the proof above we used the reduction $\pmod p$ map which itself gives rise to a useful test.

Theorem 9.2 (Reduction test). *Suppose that $f \in \mathbb{Z}[X]$ is monic, and p is a prime such that $f \pmod p$ is irreducible. Then f is irreducible.*

Proof. Write $\widetilde{\cdot}$ for the homomorphism $\mathbb{Z}[X] \to \mathbb{F}_p[X]; g \mapsto g \pmod p$. Suppose that $f = gh$ for $g, h \in \mathbb{Z}[X]$, so that $\widetilde{f} = \widetilde{g}\widetilde{h}$. Since $\widetilde{f}$ is irreducible we see that either $\widetilde{g}$ or $\widetilde{h}$ is a unit in $\mathbb{F}_p[X]$ which means that exactly one of them has degree 0. Since f is monic we have

$$\deg g + \deg h = \deg f = \deg \widetilde{f} = \deg \widetilde{g} + \deg \widetilde{h}.$$

However $\deg \widetilde{g} \leqslant \deg g$ and $\deg \widetilde{h} \leqslant \deg h$, hence $\deg \widetilde{g} = \deg g$ and $\deg \widetilde{h} = \deg h$ and so exactly one of the polynomials g and h has degree 0; say g. Since the lead coefficient of f is 1 we conclude that $g \mid 1$ and hence g is a unit as required. □

For example, the polynomial $p(X) = X^3 - 34X^2 + 17X + 289$ is irreducible in $\mathbb{Z}[X]$ because $p(X) \pmod 2 = X^3 + X + 1$ is irreducible in $\mathbb{F}_2[X]$.

⚠ We need some condition like f being monic: for example, if $f(X) = (2X+1)^2X$ then $f \pmod 2 = X$ which is irreducible but f is not.

Sometimes the reduction test is not enough to establish irreducibility, and the following proposition gives another useful approach.

Proposition 9.3 (Eisenstein's Criterion). *Suppose that $f(X) = a_nX^n + \cdots + a_1X + a_0$ is a primitive polynomial in $\mathbb{Z}[X]$ and p is a prime such that $p \mid a_i$ for all $0 \leqslant i < n$; $p \nmid a_n$; and $p^2 \nmid a_0$. Then f is irreducible in $\mathbb{Z}[X]$ and, hence, in $\mathbb{Q}[X]$.*

Proof. Write $\widetilde{\cdot}$ for the homomorphism $\mathbb{Z}[X] \to \mathbb{F}_p[X]; g \mapsto g \pmod p$. The first two hypotheses mean that $\widetilde{f} \sim X^n$, and if $f = gh$ then $\widetilde{g}\widetilde{h} \sim X^n$. Since X is prime in $\mathbb{F}_p[X]$ we conclude that $\widetilde{g} \sim X^i$ and $\widetilde{h} \sim X^{n-i}$ for some $0 \leqslant i \leqslant n$. If $0 < i < n$ then this means that the constant term of g and the constant term of h are both divisible by p and hence the constant term of f is divisible by p^2, a contradiction. Since $\deg g + \deg h = \deg f = \deg \widetilde{g} + \deg \widetilde{h}$ and $\deg \widetilde{g} \leqslant \deg g$ and $\deg \widetilde{h} \leqslant \deg h$ we conclude that $\deg \widetilde{g} = \deg g$ and $\deg \widetilde{h} = \deg h$. Since f is primitive it has no non-unit constant divisors and the result is proved. □

For example, $f(X) = 2X^4 + 3X + 3$ is irreducible in $\mathbb{Z}[X]$ by Eisenstein's Criterion.

The polynomial $q(X) = X^4 + 1$ is irreducible in $\mathbb{Z}[X]$. There are various ways this can be shown and various ways it cannot.

(i) *(Equating coefficients)* q has no degree 1 factors in $\mathbb{Z}[X]$ since it has no roots. Thus if it were reducible then there would be two quadratic factors $f(X) = a_2X^2 + a_1X +$

a_0 and $g(X) = b_2X^2 + b_1X + b_0$ in $\mathbb{Z}[X]$ such that $q = fg$. Equating coefficients gives

$$a_2b_2 = 1, a_2b_1 + a_1b_2 = 0, a_2b_0 + a_1b_1 + a_0b_2 = 0, a_1b_0 + a_0b_1 = 0, \text{ and } a_0b_0 = 1,$$

which gives a contradiction.

(ii) *(Eisenstein's Criterion)* This does not apply directly, however $q(X+1) = (X+1)^4 + 1 = X^4 + 4X^3 + 6X^2 + 4X + 2$ is irreducible by the Criterion at $p = 2$, and hence q is irreducible.

(iii) *(Prime values)* $q(0) = 1$, $q(\pm 1) = 2$, $q(\pm 2) = 17$, $q(\pm 4) = 257$, $q(\pm 6) = 1297$ are all primes or units. As before, the critical case is when q is a product of two quadratic factors in $\mathbb{Z}[X]$. Then one of the factors has to take a value from $\{-1, 1\}$ at every element of $Z := \{-6, -4, -2, -1, 0, 1, 2, 4, 6\}$, hence at least one factor takes a value from $\{-1, 1\}$ at at least five values in Z, and this quadratic either takes the value 1 at least 3 times or -1 at least 3 times. But a quadratic that is the same value at three points is constant. Hence q is irreducible.

⚠By contrast, the polynomial X^2+X+2 is irreducible in $\mathbb{Z}[X]$ (since it has no integer roots) but is also even on the integers and so prime at only finitely many values.

(iv) *(Reduction modulo a prime)* This test does *not* work for any prime. Suppose that p is a prime and write $\widetilde{q}$ for $q \pmod p$. Then for $p = 2$, $\widetilde{q}(X) = (X+1)^4$ and so q is reducible $\pmod 2$. Now suppose that p is odd. We saw in example (ii) after Theorem 8.1 that for every odd prime p there is a field extension $\mathbb{F}$ of $\mathbb{F}_p$ of degree 2. Now $U(\mathbb{F})$ has order $p^2 - 1$ in this case which, since p is odd, is divisible by 8. Since $U(\mathbb{F})$ is cyclic (Proposition 2.4) it follows that it has an element α of multiplicative order exactly 8. Since $0 = \alpha^8 - 1 = (\alpha^4 - 1)(\alpha^4 + 1)$ we see it must be the second factor that is 0 (otherwise α would have order dividing 4). But then $\widetilde{q}(\alpha) = 0$ and so q is in the ideal generated by the minimal polynomial (in $\mathbb{F}_p[X]$) of α which has degree 2. It follows that $\widetilde{q}$ is not irreducible in $\mathbb{F}_p[X]$.

10. Unique factorisation domains

The aim of this section is to establish an analogue of the Fundamental Theorem of Arithmetic for PIDs. It is instructive to keep the case $R = \mathbb{Z}$ in mind for understanding the arguments.

We say that a commutative unital ring R has the **ascending chain condition on principal ideals** or **ACCP** if whenever $(I_i)_{i \in \mathbb{N}_0}$ is an ascending chain (meaning $I_i \subset I_{i+1}$ for all $i \in \mathbb{N}_0$) of principal ideals then there is some $N \in \mathbb{N}_0$ such that $I_n = I_N$ for all $n \geqslant N$.

Lemma 10.1. *Suppose that R is a PID. Then R has the ACCP.*

Proof. Suppose that $(I_i)_{i \in \mathbb{N}_0}$ is an ascending chain of principal ideals. Then $I := \bigcup_{i \in \mathbb{N}_0} I_i$ is an ideal, and so principal say $I = \langle x \rangle$. But then there is some $N \in \mathbb{N}_0$ such that $x \in I_N$ and hence for $n \geqslant n$ we have $\langle x \rangle \subset I_N \subset I_n \subset I = \langle x \rangle$, and hence $I_n = \langle x \rangle = I_N$ for all $n \geqslant N$. The result is proved. □

A ring is said to be **Noetherian** if it satisfies the ascending chain condition on all ideals, meaning that whenever $(I_i)_{i\in\mathbb{N}_0}$ is an ascending chain (meaning $I_i \subset I_{i+1}$ for all $i \in \mathbb{N}_0$) of ideals then there is some $N \in \mathbb{N}_0$ such that $I_n = I_N$ for all $n \geqslant N$. This is a *much* more important concept but will not be a focus of this course.

⚠The fact that the chain of ideals is ascending rather than descending makes a significant difference. We say that R has the **descending chain condition on principal ideals** or **DCCP** if whenever $(I_i)_{i\in\mathbb{N}_0}$ is a descending chain (meaning $I_i \supset I_{i+1}$ for all $i \in \mathbb{N}_0$) of principal ideals then there is some $N \in \mathbb{N}_0$ such that $I_n = I_N$ for all $n \geqslant N$.

While any PID has the ACCP, it turns out that an integral domain has the DCCP if and only if it is a field. The if direction is immediate since there are only two ideals in a field. In the other direction, suppose $x \in R^*$. Then we have a chain of ideals $\langle x\rangle \supset \langle x^2\rangle \supset \cdots \supset \langle x^i\rangle \supset \cdots$. By the DCCP there is some $i \in \mathbb{N}$ such that $\langle x^i\rangle = \langle x^{i+1}\rangle$, and so there is some $r \in R$ such that $x^i = rx^{i+1}$ and hence by the Cancellation Lemma $rx = xr = 1$ and so $x \in U(R)$ as required.

For us the important feature of the ACCP is that it will let us factorise elements of a ring. To formulate this precisely we say an integral domain R is **factorisation domain** or **atomic domain** if for every $x \in R^*$ there is a possibly-empty vector $(x_1, \dots, x_r)$ of irreducible elements of R such that $x \sim x_1 \cdots x_r$ with the convention that the empty product is 1.

Lemma 10.2. *Suppose that R is an integral domain with the ACCP. Then R is a factorisation domain.*

Proof. Write $\mathcal{F}$ for the set of elements of R that can be written as a product of irreducible elements so that $1 \in \mathcal{F}$, all irreducible elements of R are also in $\mathcal{F}$, and $\mathcal{F}$ is closed under multiplication. If $R\backslash\mathcal{F}$ is not empty we can create a sequence $(x_i)_{i\in\mathbb{N}_0}$ of elements of $R\backslash\mathcal{F}$ iteratively with $\langle x_0\rangle \subsetneq \langle x_1\rangle \subsetneq \cdots$ which contradicts the ACCP. Let $x_0 \in R\backslash\mathcal{F}$. At step i we have $x_i \notin \mathcal{F}$ and so it is not irreducible and hence $x_i = y_i z_i$ for some $y_i, z_i \not\sim x_i$. Since $\mathcal{F}$ is closed under multiplication we cannot have both y_i and z_i in $\mathcal{F}$; let $x_{i+1} \in \{y_i, z_i\}$ such that $x_{i+1} \notin \mathcal{F}$. This gives the desired sequence and the result is proved. □

Although we did not mention it, this argument required the Axiom of Dependent Choice, but not the full Axiom of Choice.

There are factorisation domains not having the ACCP but these are not easy to construct; the first example was given by Grams in [Gra74].

Primes are important for ensuring uniqueness of factorisation.

Proposition 10.3. *Suppose that R is an integral domain and $x_1, \dots, x_r, y_1, \dots, y_s$ are primes with $x_1 \cdots x_r \sim y_1 \cdots y_s$ then $r = s$ and there is a permutation π of $[r] := \{1, \dots, r\}$ such that $x_i \sim y_{\pi(i)}$ for all $1 \leqslant i \leqslant r$.*

Proof. It is convenient for this induction to prove something slightly more general: We shall show that if $x_1, \dots, x_r$ are primes and $(y_i)_{i\in I}$ is a sequence of irreducible elements indexed by a finite set I such that $\prod_{i=1}^r x_i \sim \prod_{i\in I} y_i$ then there is a bijection $\pi : [r] \to I$ such that $x_i \sim y_{\pi(i)}$ for all $1 \leqslant i \leqslant r$.

For $r = 0$ we have $\prod_{i \in I} y_i \sim 1$ (by definition of the empty product) and so $y_i \in U(R)$ for all $i \in I$ meaning that I is empty since no unit is irreducible. Now, suppose that $r > 0$. Then x_r is prime and $x_r \mid \prod_{i \in I} y_i$ whence there is some $j \in I$ such that $x_r \mid y_j$. But y_j is irreducible and $x_r \not\sim 1$ and so $x_r \sim y_j$. By the Cancellation Lemma $x_1 \cdots x_{r-1} \sim \prod_{i \in I \setminus \{j\}} y_i$, and by the inductive hypothesis there is a bijection $\tilde{\pi} : [r-1] \to I \setminus \{j\}$ such that $x_i \sim y_{\tilde{\pi}(i)}$ for all $1 \leqslant i \leqslant r-1$. Extend this to a bijection $[r] \to I$ by setting $\pi(r) = j$ and the result is proved. □

A **unique factorisation domain** or **UFD** is a factorisation domain in which all irreducible elements are prime, which leads to a uniqueness of factorisation as described in Proposition 10.3.

Proposition 10.4. *Suppose that R is a PID. Then R is a UFD.*

Proof. By Lemma 10.1 and Lemma 10.2 we have that R is a factorisation domain. That every irreducible is prime follows from Proposition 7.6 and Proposition 7.2. □

In particular, since $\mathbb{Z}$ is a PID by Proposition 6.1 the above gives the Fundamental Theorem of Arithmetic.

Not all UFDs are PIDs. Indeed, $\mathbb{Z}[X]$ is a UFD (we have not proved this) but it is not a PID (by Q2, Examples Sheet 2) since $\mathbb{Z}$ is not a field. In general Gauss' Lemma can be used to show that if R is a UFD then $R[X]$ is a UFD, which gives other examples of UFDs that are not PIDs such as $\mathbb{F}[X, Y]$.

11. Euclidean domains

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

Proposition 11.1 (Division algorithm for integers)**.** *$\mathbb{Z}$ is a Euclidean domain.*

Proof. Take $f(z) := |z|$ for $z \in \mathbb{Z}^*$. Suppose that $a, b \in \mathbb{Z}$ and $b \neq 0$, and consider the set $\{a + bq : q \in \mathbb{Z}, a + bq \geqslant 0\}$. This is a non-empty set of natural numbers and so it has a minimal element, call it r. We certainly have $r \geqslant 0$; suppose $r \geqslant f(b) = |b| = \omega b$ for some $\omega \in \{-1, 1\}$. Then $r = a + bq$, and $0 \leqslant r - f(b) = a + b(q - \omega) < r$ contradicting minimality; hence $r < f(b)$. Since $r \geqslant 0$ we have $r = 0$ or $f(r) = r < f(b)$ as required. □

There is also a division algorithm for polynomials which is captured by the same definition.

Proposition 11.2 (Division algorithm for polynomials)**.** *Suppose that $\mathbb{F}$ is a field. Then $\mathbb{F}[X]$ is a Euclidean domain.*

Proof. Take $f(p) = \deg p$ for $p \in \mathbb{F}[X]^*$. Suppose that $a, b \in \mathbb{F}[X]$ and $b \neq 0$. If $b \mid a$ then we take $r = 0$ and let q be such that $a = bq$; we are done. It not then $P := \{a + bq : q \in \mathbb{F}[X]\}$ does not contain 0; take $r = a + bq$ such that the degree is minimal for polynomials in P.

Suppose that $\deg r \geqslant \deg b$. Then write λ for the coefficient of $X^{\deg r}$ in r and note that $r' := r - b\lambda X^{\deg r - \deg b}$ has $r' \in P$ and $\deg r' < \deg r$, a contradiction. It follows that $\deg r < \deg b$ as required. □

Given an integral domain R, a **Dedekind-Hasse function** on R is a map $N : R^* \to \mathbb{N}_0$ such that whenever $a, b \in R^*$ either $b \mid a$ or there is some non-zero element $c \in \langle a, b\rangle$ such that $N(c) < N(b)$. Put another way either $b \mid a$ or there are elements $p, q \in R$, $c \in R^*$ such that $ap = bq + c$ and $N(c) < N(b)$. The definition of Euclidean function places the additional requirement $p = 1$, so in particular any ring supporting a Euclidean function supports a Dedekind-Hasse function.

Proposition 11.3. *Suppose that R is an integral domain. Then R is a PID if and only if R admits a Dedekind-Hasse function. In particular, R is a PID if it is a Euclidean Domain.*

Proof. $\Rightarrow$: Since R is a UFD we can define $N : R^* \to \mathbb{N}_0$ to be the number of irreducible factors[15] of its argument. For any $a, b \in R^*$ either $b \mid a$ or else $\langle a, b\rangle \supsetneq \langle b\rangle$. Since $\langle a, b\rangle$ is principal, it is generated by some $c \in R^*$ and we have $c \mid b$, and $c \not\sim b$ whence $N(c) < N(b)$.

$\Leftarrow$: Suppose that I is a non-zero ideal in R and let $b \in I$ have $N(b)$ minimal. Now suppose that $a \in I$ so that by the Dedekind-Hasse property either $b \mid a$, or else there is some non-zero $c \in \langle a, b\rangle \subset I$ with $N(c) < N(b)$. The second conclusion is incompatible with the minimality and so the first holds and hence $I = \langle b\rangle$ as required. □

There are integral domains that are not Euclidean domains, for example $\mathbb{F}[X, Y]$ is not even a PID and so by the above it is not an ED.

However, more than this there are examples of PIDs which are not Euclidean domains, but showing this is not easy. The rings $\mathbb{Z}\left[\frac{1}{2}(1 + \sqrt{-D})\right]$ where $D \in \{19, 43, 67, 163\}$ are some classic examples all described in [PV08]. We shall give a different example now.

Write $A := \mathbb{R}[X, Y]/I$ where $I := \langle X^2 + Y^2 + 1\rangle$. Every $F \in A$ has a unique coset representative of the form $p(X) + Yq(X)$ and we write $\widetilde{F} := p(X) - Yq(X) + I$, which has two useful properties:

- $\widetilde{\cdot}$ is multiplicative: *i.e.* $\widetilde{F}\widetilde{G} = \widetilde{FG}$ for all $F, G \in A$.
- If $F = p(X) + Yq(X) + I$ then since the Y-degree of any non-zero element if I is positive we have $F\widetilde{F} \cap \mathbb{R}[X] = p(X)^2 + q(X)^2(X^2 + 1)$.

In view of this, if $F = p(X) + Yq(X) + I$ and $G = s(X) + Yt(X) + I$ and $FG = 0_A$ then $F\widetilde{F}G\widetilde{G} = 0_A$, and so

$$(p(X)^2 + (X^2 + 1)q(X)^2)(s(X)^2 + (X^2 + 1)t(X)^2) = 0$$

whence $p = q = 0$ or $s = t = 0$ and we see that A is an integral domain. Similarly, suppose that $FG = 1 + I$ so that F and G are units of A. The same argument shows that $q = t = 0$ and p and s are constants in $\mathbb{R}^*$.

[15] If one wanted N to be multiplicative one could take 2 to the power of the number of irreducible factors Instead.

Lemma 11.4. *Suppose that $J \neq \{0\}$ is an ideal in A. Then A/J is a finite-dimensional vector space over $\mathbb{R}$.*

Proof. By hypothesis there is some $p(X) + Yq(X) + I \in J$ without both p and q being 0. Thus $p(X)^2 + (X^2 + 1)q(X)^2 + I \in J$; write d for the degree of $p(X)^2 + (X^2 + 1)q(X)^2$ and note that $1, \ldots, X^d$ and $Y, YX, \ldots, YX^d$ are linearly dependent in A/J from which the result follows. □

The maximal ideals in A have a particular structure.

Lemma 11.5. *Suppose that J is a maximal ideal in A. Then $J = \langle \alpha + \beta X + \gamma Y + I\rangle$ for $\alpha, \beta, \gamma \in \mathbb{R}$ with $(\beta, \gamma) \neq (0, 0)$ and $\dim_{\mathbb{R}} A/J = 2$.*

Proof. We know by Proposition 7.3 that A/J is a field and since A and J are both $\mathbb{R}$-vector spaces we know that A/J is a field extension of $\mathbb{R}$, and it is finite by Lemma 11.4. Thus by Corollary 8.2 the extension has degree at most 2. On the other hand $(X + I + J)^2 + (Y + I + J)^2 + (1 + I + J)^2 = 0_{A/J} = I + J$ and $1 + I + J \neq I + J$ so we cannot have $A/J \cong \mathbb{R}$ (since then a sum of three squares of elements in $\mathbb{R}$ would be zero without all of the elements being zero).

The elements $1 + I + J$, $X + I + J$, and $Y + I + J$ must be linearly dependent in A/J and $1 + I + J \neq J$, so there are reals $\alpha, \beta, \gamma \in \mathbb{R}$ with $(\beta, \gamma) \neq (0, 0)$ such that $\alpha(1 + I) + \beta(X + I) + \gamma(Y + I) = (\alpha + \beta X + \gamma Y) + I \in J$. Finally $(\beta^2 + \gamma^2)X^2 + 2\alpha\beta X + (\alpha^2 + \gamma^2) + I \in \langle \alpha + \beta X + \gamma Y + I\rangle$, and so $A/\langle \alpha + \beta X + \gamma Y + I\rangle$ is at most 2-dimensional and $J = \langle \alpha + \beta X + \gamma Y + I\rangle$ as required. □

Proposition 11.6. *The ring A is not a Euclidean domain.*

Proof. Suppose that A supports a Euclidean function f. A in not a field so we may take $F \in A^*$ a non-unit with $f(F)$ minimal (amongst non-units). The ideal $\langle F\rangle$ is maximal and so $F \sim \alpha + \beta X + \gamma Y$ for some $(\beta, \gamma) \neq (0, 0)$. Let (β', γ') be linearly independent of (β, γ), so that by the Euclidean property we have $\beta' X + \gamma' Y \in \mathbb{R} + \langle F\rangle$ and hence $\langle F\rangle = A$, a contradiction. □

Proposition 11.7. *The ring A is a PID.*

Proof. Suppose that J is a non-principal ideal in A. Then J is non-trivial and so $\dim_{\mathbb{R}} A/J < \infty$ by Lemma 11.4. Set $J_0 := J$ and at stage n suppose J_n is non-principal. Then it is contained in a maximal ideal by Theorem 7.4, and this ideal is principal by Lemma 11.5 so there is some non-unit F_n such that $J_n \subset \langle F_n\rangle$; set $J_{n+1} := \{F : FF_n \in J_n\}$ so that $J_{n+1} \supsetneq J_n$ and J_{n+1} is non-principal. The former conclusion ensures that $\dim_{\mathbb{R}} A/J_{n+1} < \dim_{\mathbb{R}} A/J_n$ but this process cannot go on indefinitely since $\dim_{\mathbb{R}} A/J < \infty$, so we have a contradiction. □

12. Modules

Suppose that M is a commutative group. An **endomorphism** of M is a homomorphism $M \to M$; we write $\mathrm{End}(M)$ for the set of all endomorphisms of M. This has the structure of a unital ring with addition defined coordinate-wise and multiplication by composition:

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

The multiplicative identity of this ring is the identity map $M \to M; x \mapsto x$, and the zero is the map $M \to M; x \mapsto 0_M$. There are a number of things to check; we mention the important parts:

- The set of all functions $M \to M$ forms a (commutative) group under pointwise addition, and the fact that $\mathrm{End}(M)$ is a subgroup makes essential use of the commutativity of M.[16]
- The set of all functions $M \to M$ is closed under composition, and this operation is associative. We also need that the composition of homomorphisms is again a homomorphism.
- Finally, composition of functions is right-distributive over coordinate-wise addition, but it is not in general left distributive. This is the point where we make critical use of the fact that endomorphisms are homomorphism because this ensures that composition is left distributive over coordinate-wise addition: $(\phi \circ (\psi + \pi))(x) = \phi(\psi(x) + \pi(x)) = (\phi \circ \psi)(x) + (\phi \circ \pi)(x)$ for all $x \in$M.

Thus $\mathrm{End}(M)$ is another example of a ring in roughly the same way as $\mathrm{Sym}(X)$ – the set of bijections of a set X – is a group. An action of a group G on a set X is a homomorphism $G \to \mathrm{Sym}(X)$, and a module is the same sort of thing for rings. Specifically, given a unital ring R a **left R-module** M is a commutative group also denoted M and a unital homomorphism

$$\rho : R \to \mathrm{End}(M)$$

which we call **scalar multiplication**; we write rx for $\rho(r)(x)$.

We begin by mentioning some examples.

(i) The analogue of Cayley's Theorem[17] for groups is the fact that for any unital ring R there is a unital homomorphism

$$R \to \mathrm{End}(R); r \mapsto (R \to R; x \mapsto rx).$$

(ii) Given a vector space V and a field $\mathbb{F}$ the map

$$\mathbb{F} \to \mathrm{End}(V); \lambda \mapsto (V \to V; v \mapsto \lambda v)$$

is a unital homomorphism giving V the structure of a left $\mathbb{F}$-module. Conversely, if V is a left $\mathbb{F}$-module then it has the structure if an $\mathbb{F}$-vector space with scalar multiplication the same as that for modules.

(iii) A commutative group M is a left $\mathbb{Z}$-module via the map

$$\mathbb{Z} \to \mathrm{End}(M); z \mapsto (M \to M; x \mapsto zx).$$

(iv) Given a vector space V and an endomorphism[18] $T : V \to V$ there are two $\mathbb{F}[X]$-modules associated with T:

[16] Indeed, the set of homomorphisms $G \to G$ with the binary operation $(\phi, \psi) \mapsto (x \mapsto \phi(x)\psi(x))$ forms a group if and only if G is commutative. To see this consider what happens if $x \mapsto x^{-1}$ is a homomorphism.

[17] Cayley's Theorem says that if G is a group then $G \to \mathrm{Sym}(G); g \mapsto (G \to G; x \mapsto gx)$ is a well-defined homomorphism.

[18] Meaning here a linear map $V \to V$.

(a) The commutative group V equipped with
$$\mathbb{F}[X] \to \mathrm{End}(V); p \mapsto (V \to V; x \mapsto p(T)x).$$
(b) The commutative group $\mathbb{F}[T] := \{p(T) : p \in \mathbb{F}[X]\}$ equipped with
$$\mathbb{F}[X] \to \mathrm{End}(\mathbb{F}[T]); p \mapsto (\mathbb{F}[T] \to \mathbb{F}[T]; x \mapsto p(T)x).$$

Modules can be thought of as analogues of vector spaces with the field replaced by a ring.

A **left R-module homomorphism** or **R-linear map** between two left R-modules M and N is a group homomorphism $\phi : M \to N$ with
$$\phi(rx) = r\phi(x) \text{ for all } x \in M, r \in R.$$
If $\mathbb{F}$ is a field this has the same meaning as $\mathbb{F}$-linear in the usual sense.

⚠The rx on the left is the scalar multiplication on M and the $r\phi(x)$ is the scalar multiplication on N.

As before we say that $\phi : M \to N$ is an **isomorphism** if it is a homomorphism with a homomorphic inverse or, equivalently, a bijective homomorphism; we write $M \cong N$.

Given a module M we say that N is a **submodule** of M if N is a subgroup of M as an additive group and $rx \in N$ for all $x \in N$ and $r \in R$.

Proposition 12.1 (Quotient modules)**.** *Suppose that M is a left R-module and N is a submodule of M. Then M/N can be equipped with the scalar multiplication $r(x + N) := rx + N$ for $r \in R$ making it into an R-module.*

Proof. Since N is a commutative subgroup of M we have that M/N is a commutative group. We have already seen that in this case $\mathrm{End}(M/N)$ is a unital ring. We just need to check that the map
$$R \to \mathrm{End}(M/N); r \mapsto (x + N \mapsto rx + N)$$
is a well-defined unital homomorphism. To see it is well-defined note that if $x+N = y+N$ then $x - y \in N$, and hence $r(x - y) \in N$ and so $rx + N = ry + N$ so that the map on the right maps $M/N \to M/N$. It is a homomorphism since
$$r(x + y) + N = ((rx) + (ry)) + N = (rx + N) + (ry + N) \text{ for all } x, y \in M,$$
and so the map on the left really maps into $\mathrm{End}(M/N)$. Since $1x = x$ for all $x \in M$ we have that 1 is mapped to the multiplicative identity in $\mathrm{End}(M/N)$. Finally,
$$(r + s)x + N = ((rx) + (sx)) + N = (rx + N) + (sx + N) \text{ for all } r, s \in R, x \in M,$$
and
$$(rs)x + N = r(sx) + N = r(sx + N) = (rs)(x + N) \text{ for all } r, s \in R, x \in M.$$
The result is proved. □

Theorem 12.2 (First Isomorphism Theorem)**.** *Suppose that $\phi : M \to N$ is an R-linear map between left R-modules M and N. Then $\ker\phi$ is a submodule of M; $\phi(M)$ 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 $\phi(M)$.

Proof. First, $\ker\phi$ and $\operatorname{Im}\phi$ are subgroups of the additive groups of M and N respectively by the First Isomorphism Theorem for groups. Now, if $r \in R$ and $x \in \ker\phi$ then $\phi(rx) = r\phi(x) = r0 = 0$, and so $rx \in \ker\phi$, and $\ker\phi$ is a submodule of M. On the other hand if $r \in R$ and $x \in \phi(M)$ then $x = \phi(y)$ for $y \in M$ so $rx = r\phi(y) = \phi(ry) \in \phi(M)i$ and so $\phi(M)$ is a submodule of N.

By Proposition 12.1 $M/\ker\phi$ is a left R-module. The map is injective and well-defined since $x + \ker\phi = y + \ker\phi$ iff $x - y \in \ker\phi$ iff $\phi(x - y) = 0$ iff $\phi(x) = \phi(y)$. The image is certainly $\phi(M)$. It remains to check the map is R-linear:

$$\widetilde{\phi}((x + y) + \ker\phi) = \phi(x + y) = \phi(x) + \phi(y) = \widetilde{\phi}(x + \ker\phi) + \widetilde{\phi}(y + \ker\phi),$$

and

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

The result is proved. □

Given an indexing set I and left R-modules $(M_i)_{i\in I}$, the **direct sum** of $(M_i)_{i\in I}$ is denoted $\bigoplus_{i\in I} M_i$ and is defined to be the direct sum of the commutative groups M_i, also denoted $\bigoplus_{i\in I} M_i$, endowed with the structure of a left R-module via the multiplication $rx := (rx_i)_{i\in I}$ where rx_i denotes the scalar multiplication of r on x_i as an element of M_i.

We take the usual convention that if $I = \varnothing$ then $\bigoplus_{i\in I} M_i$ is the zero module, and if $M_1, \dots, M_n$ are modules then we write $M_1 \oplus \cdots \oplus M_n$ for $\bigoplus_{i\in\{1,\dots,n\}} M_i$, and finally M^n for the direct sum of M with itself n-times.

⚠Recall that the direct sum of an infinite family $(M_i)_{i\in I}$ of commutative groups is the set of $x \in \prod_{i\in I} M_i$ with at most finitely many non-identity coordinates. For example, if $I = \mathbb{N}_0$ and $M_i = \mathbb{Z}$ then $\bigoplus_{i\in\mathbb{N}_0} M_i$ is the set of integer-valued sequences which are non-zero at a finite number of coordinates under coordinate-wise addition. This is *much* smaller than the set of all sequences.

For $j \in I$, define the map

$$\iota_j : M_j \to \bigoplus_{i\in I} M_i$$

where the jth coordinate of $\iota_j(x)$ is x and the ith coordinate is 0_{M_i} for all $i \neq j$. This map is R-linear.

13. Cyclic modules and the Chinese Remainder Theorem revisited

Vector spaces are an important example of modules, and just as finite dimensional vector spaces were amenable to particularly detailed study so we shall be interested in the analogue for modules. Given $x_1, \dots, x_n$ in a left R-module M we write

$$\langle x_1, \dots, x_n\rangle := \{r_1x_1 + \cdots + r_nx_n : r_1, \dots, r_n \in R\}.$$

This is an R-module, and we say that M is **generated** by $x_1, \dots, x_n$. M is **finitely generated** if there are elements $x_1, \dots, x_n \in M$ such that M is generated by $x_1, \dots, x_n$, or equivalently, if there is an R-linear surjection $R^n \to M$ for some $n \in \mathbb{N}$.

If R is a field so that M is a vector space then M is finitely generated if and only if it is finite dimensional, but defining dimension requires two important theorems.

Theorem 13.1 (Finitely generated vector spaces have a basis)**.** *Suppose that V is a finitely generated vector space over $\mathbb{F}$. Then there is some $n \in \mathbb{N}$ such that $V \cong \mathbb{F}^n$.*

Theorem 13.2 (All bases have the same size)**.** *Suppose that V is a (finitely generated) vector space over $\mathbb{F}$. If $V \cong \mathbb{F}^n$ and $V \cong \mathbb{F}^m$ then $n = m$.*

Informally we think of the first result as saying that finitely generated vector spaces can be built out of copies of $\mathbb{F}$; and the second as saying that this can be done in an essentially unique way. We should like analogues of these theorems for modules, however there are some obstacles.

Suppose that M is a left R-module. We say that $x_1, \dots, x_n$ in M are **linearly independent** if

$$r_1x_1 + \cdots + r_nx_n = 0_M \text{ for } r \in R^n \text{ implies } r_1, \dots, r_n = 0_R,$$

and this coincides with the existing definition for vector spaces. As with vectors, if $x_1, \dots, x_n$ are linearly independent and generate M then we say that $x_1, \dots, x_n$ form a **basis** for M, and any module with a basis is called a **free module**.[19] Put another way, a finitely generated R-module M is a free module if it is R-linearly isomorphic to R^n for some $n \in \mathbb{N}_0$.

Free modules are those that can be built out of copies of their underlying ring so they include vector spaces, but also modules like $\mathbb{Z}^n$. However, some relatively simple modules are not free; indeed, $\mathbb{Z}/\langle N\rangle$ contains no non-empty independent sets[20]. This means that if we are to hope for an analogue of Theorem 13.1 for modules we are going to have to enlarge our class of building blocks to include more than just the underlying ring.

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

(i) For a field $\mathbb{F}$, a cyclic $\mathbb{F}$-module is either the zero module or isomorphic to $\mathbb{F}$.
(ii) A cyclic $\mathbb{Z}$-module is isomorphic to $\mathbb{Z}/\langle N\rangle$ for some $N \in \mathbb{N}_0$. These are the cyclic groups, also denoted $\mathbb{Z}/N\mathbb{Z}$ for $N \neq 0$ and $\mathbb{Z}$ for $\mathbb{Z}/\langle 0\rangle$.

Cyclic modules can be described in terms of something called left ideals: given a unital ring R we say that I is a **left ideal** of R if I is an additive subgroup of R and $ra \in I$ for all $r \in R$ and $a \in I$. Equivalently, I is a submodule of R considered as a left R-module over itself.

If R is commutative then a left ideal is an ideal as defined earlier in the course – sometimes these are called **two-sided ideals** for clarity. In general R/I does *not* have the structure of a ring, but since I is a submodule of R, R/I does have the structure of an R-module and it is cyclic generated by $1 + I$. The scalar multiplication in these instances is given by

$$R \to \mathrm{End}(I); r \mapsto (I \to I; x \mapsto rx) \text{ and } R \to \mathrm{End}(R/I); r \mapsto (R/I \to R/I; xI \mapsto rxI)$$

respectively.

[19] We should be a little careful here about modules that are not finitely generated but we shall not be dealing with those in this course.

[20] As a $\mathbb{Z}$-module. As a $\mathbb{Z}/\langle N\rangle$-module is does contain independent sets.

Given a module M and an element $x \in M$ we write $\mathrm{Ann}_R(x) := \{r \in R : rx = 0_M\}$ and call this the **annihilator** of x.

Lemma 13.3. *Suppose that R is a unital ring, M is a left R-module, and $x \in M$. Then $\mathrm{Ann}_R(x)$ is a left ideal of R and if M is generated by x then $M \cong R/\mathrm{Ann}_R(x)$.*

Proof. That $\mathrm{Ann}_R(x)$ is a left ideal of R is a short check. For the second part, by the First Isomorphism Theorem for modules (Theorem 12.2) applied to the R-linear map $R \to M; r \mapsto rx$, the map

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

is an R-linear isomorphism of left R-modules. □

Informally the lemma tells us that up to isomorphism cyclic R-modules are quotients of R by left ideals.

The intersection and sum of two left ideals is a left ideal (*c.f.* Lemma 3.4), and these operations on ideals provide a way to combine cyclic modules.

Theorem 13.4. *Suppose that R is a unital ring and I and J are left ideals with $I + J = R$. Then $R/(I \cap J) \cong (R/I) \oplus (R/J)$ as left R-modules.*

Proof. We define $\pi : R \to (R/I) \oplus (R/J); r \mapsto (r + I, r + J)$. First, π is a homomorphism of commutative groups:

$$\begin{aligned}\pi(r+s) &= (r+s+I, r+s+J)\\ &= ((r+I)+(s+I), (r+J)+(s+J))\\ &= (r+I, r+J) + (s+I, s+J) = \pi(r) + \pi(s)\end{aligned}$$

for all $r, s \in R$. Secondly,

$$\pi(rs) = (rs+I, rs+J) = (r(s+I), r(s+J)) = r(s+I, s+J) = r\pi(s),$$

for all $r, s \in R$ and so π is an R-linear map.

To show that the map is surjective, suppose that $(x + I, y + J) \in (R/I) \oplus (R/J)$. Since $I + J = R$ there are elements $\alpha \in I$ and $\beta \in J$ such that $\alpha + \beta = 1$. Consider $z := x\beta + y\alpha$. We have

$$z + I = x\beta + y\alpha + I = x + (y - x)\alpha + I,$$

but $y - x \in R$ and $\alpha \in I$ and so $(y - x)\alpha \in I$ since I is a left ideal. We conclude that $z + I = x + I$. Similarly $z + J = y + J$ and hence $\pi(z) = (x + I, y + J)$, so π is surjective.

Finally, the kernel of π is $I \cap J$ and so the result follows by the First Isomorphism Theorem for modules. □

This theorem might be called a 'non-commutative version' of the Chinese Remainder Theorem. It does not in general extend to more than two summands, but if R is commutative then it does.

Theorem 13.5 (Chinese Remainder Theorem for modules)**.** *Suppose that R is a commutative unital ring and $I_1, \ldots, I_k$ are pairwise coprime ideals in R. Then*

$$R/(I_1 \cap \cdots \cap I_k) \cong (R/I_1) \oplus \cdots \oplus (R/I_k)$$

as left R-modules.

Proof. The additive group of the ring $(R/I_1)\times\cdots\times(R/I_k)$ is the same as the additive group of the module $(R/I_1)\oplus\cdots\oplus(R/I_k)$. Theorem 5.1 gives a bijective (group) homomorphism between the additive groups of $R/(I_1\cap\cdots\cap I_k)$ and $(R/I_1)\oplus\cdots\oplus(R/I_k)$ and the explicit form of this homomorphism is easily checked to be R-linear. The result is proved. □

14. Uniqueness of cyclic decompositions

A basic obstacle to an analogue of Theorem 13.2 for modules comes from some non-obvious relationships between cyclic modules *e.g.* the fact that $\mathbb{Z}/\langle 6\rangle\cong\mathbb{Z}/\langle 3\rangle\oplus\mathbb{Z}/\langle 2\rangle$ as $\mathbb{Z}$-modules by Theorem 13.5.

Despite this, there is a way to recover uniqueness at least for modules over commutative rings provided our cyclic modules are suitably nested.

Theorem 14.1 (Uniqueness Theorem)**.** *Suppose that R is a commutative unital ring, M is a (finitely generated) R-module, and $I_1\subset\cdots\subset I_n$ and $J_1\subset\cdots\subset J_m$ are proper ideals such that $M\cong(R/I_1)\oplus\cdots\oplus(R/I_n)$ and $M\cong(R/J_1)\oplus\cdots\oplus(R/J_m)$. Then $n=m$ and $J_k=I_k$ for all $1\leqslant k\leqslant n$.*

We begin with a result which bootstraps the analogous fact for vector spaces.

Lemma 14.2. *Suppose that R is a commutative unital ring, and $I_1\subset\cdots\subset I_n$ are proper ideals. Then $(R/I_1)\oplus\cdots\oplus(R/I_n)$ is generated by a set of size n and by no smaller set.*

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

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

ensures the first part of the lemma. For the second, by Theorem 7.4 there is a maximal ideal $J\supset I_n$ and hence $J\supset I_k$ for all $1\leqslant k\leqslant n$. The R-linear surjection

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

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

$$\begin{aligned}x&=r_1x^{(1)}+\cdots+r_tx^{(t)}\\&=(r_1x^{(1)}_1+\cdots+r_tx^{(t)}_1,\dots,r_1x^{(1)}_n+\cdots+r_tx^{(t)}_n)\\&=((r_1+J)x^{(1)}_1+\cdots+(r_t+J)x^{(t)}_1,\dots,(r_1+J)x^{(1)}_n+\cdots+(r_t+J)x^{(t)}_n)\\&=(r_1+J)x^{(1)}+\cdots+(r_t+J)x^{(t)}.\end{aligned}$$

Proposition 7.3 ensures that R/J is a field and the map $r+J\mapsto((R/J)^n\rightarrow(R/J)^n;x\mapsto(rx_1,\dots,rx_n))$ is a well-defined unital homomorphism so that $(R/J)^n$ is a vector space over R/J and the above calculation shows that $x^{(1)},\dots,x^{(t)}$ is a spanning set for $(R/J)^n$

as an (R/J)-module *i.e.* as a vector space over R/J. Since $(R/J)^n$ is an n-dimensional vector space over R/J any spanning set has size at least n *i.e.* $t \geqslant n$. □

Proof of Theorem 14.1. First, by Lemma 14.2 we have $n = m$. Since the map $M \to M; z \mapsto xz$ is R-linear we have that xM is an R-module. We shall show that for $1 \leqslant k \leqslant n$

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

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

Suppose that $x \in R$. The module $x(R/I_k)$ is generated by $x + I_k$ as an R-module, and so by Lemma 13.3 $x(R/I_k) \cong R/\operatorname{Ann}_R(x + I_k)$. Now

$$\operatorname{Ann}_R(x + I_k) = \{r \in R : r(x + I_k) = I_k\} = \{r : rx \in I_k\},$$

so $x \notin I_k$ if and only if $\operatorname{Ann}_R(x + I_k)$ is proper[21]; and $\operatorname{Ann}_R(x + I_1) \subset \cdots \subset \operatorname{Ann}_R(x + I_n)$ since the $I_1 \subset \cdots \subset I_k$. Let $0 \leqslant j(x) \leqslant n$ be maximal such that $x \notin I_{j(x)}$ (with $j(x) = 0$ if $x \in I_1$) then

$$\begin{aligned} xM &\cong x((R/I_1) \oplus \cdots \oplus (R/I_n))\\ &\cong (R/\operatorname{Ann}_R(x + I_1)) \oplus \cdots \oplus (R/\operatorname{Ann}_R(x + I_n))\\ &\cong (R/\operatorname{Ann}_R(x + I_1)) \oplus \cdots \oplus (R/\operatorname{Ann}_R(x + I_{j(x)})) \end{aligned} \quad \begin{array}{l} \operatorname{Ann}_R(x + I_k) \textit{ not proper}\\ \Rightarrow R/\operatorname{Ann}_R(x + I_k) \cong \{0\}\\ \textit{for } j(x) < k \leqslant n \end{array}$$

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

By Lemma 14.2 we conclude that if $x \notin I_k$ then $j(x) \geqslant k$ and so xM is not generated by strictly fewer than $j(x)$ (and hence k) elements and so $x \notin K_k$. On the other hand if $x \in I_k$ then $j(x) < k$ and so xM *is* generated by at most $j(x)$ (*i.e.* strictly fewer than k) elements and so $x \in K_k$. The result is proved. □

This theorem lets us define the **rank** of a free module to be the size of its basis. Indeed, suppose that we had a free R-module with two bases. Then we have an isomorphism $R^n \to R^m$, and taking $I_1 = \cdots = I_n = \{0\}$ and $J_1 = \cdots = J_m = \{0\}$ we get $n = m$.

⚠ Suppose that M is the direct sum of countably many copies of $\mathbb{Z}$ indexed by $\mathbb{N}_0$ and $R := \operatorname{End}(M)$. Then it can be shown that $R \cong R^2$ as left R-modules. In particular, we cannot hope to extend Theorem 14.1 to all unital rings.

15. Existence of cyclic decompositions

We now turn to the problem of an analogue of Theorem 13.1.

Theorem 15.1. *Suppose that R is a PID and M is a finitely generated R-module. Then there is $n \in \mathbb{N}_0$ and proper ideals $I_1 \subset \cdots \subset I_n$ such that*

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

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

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

[21] I paused here in lectures, and since this was not obvious to me in that moment I am adding some clarification here: If $x \in I_k$ then $rx \in I_k$ for all $r \in R$ since I_k is an ideal, and hence $\operatorname{Ann}_R(x + I_k) = R$. Conversely, if $\operatorname{Ann}_R(x + I_k) = R$ then $1.(x + I_k) = I_k$ and so $x \in I_k$.

Lemma 15.2. *Suppose that R is a PID with elements $a_1, \dots, a_n, h \in R$, and $\langle a_1, \dots, a_n \rangle = \langle h \rangle$, and M is an R-module with elements $x_1, \dots, x_n \in M$. Then there are elements $y_1, \dots, y_n \in M$ with $\langle y_1, \dots, y_n \rangle = \langle x_1, \dots, x_n \rangle$ such that $hy_n = a_1x_1 + \cdots + a_nx_n$.*

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

We proceed by induction on n; $n = 1$ is immediate since $a_1 \sim h$ in that case. For $n > 1$ let h' be a generator of $\langle a_1, \dots, a_{n-1} \rangle$. By the inductive hypothesis we may take $y_1, \dots, y_{n-2}, y^*_{n-1}$ such that $\langle y_1, \dots, y_{n-2}, y^*_{n-1} \rangle = \langle x_1, \dots, x_{n-1} \rangle$ and $h'y^*_{n-1} = a_1x_1 + \cdots + a_{n-1}x_{n-1}$.

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

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

Finally, $hy_n = h'y^*_{n-1} + a_nx_n = a_1x_1 + \cdots + a_nx_n$ and the result is proved. □

Proof of Theorem 15.1. We proceed inductively to show that there are elements $z_1, \dots, z_n$ generating M such that

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

Since M is finitely generated there is a minimal $n \in \mathbb{N}$ such that M is generated by a set of size n. Let $x_1, \dots, x_n$ be a set of generators in which $\operatorname{Ann}_R(x_n)$ is generated by an element r_n (possibly 0_R) with the smallest[22] number of irreducible factors *i.e.* for every generating set $y_1, \dots, y_n$ of M, any generator of the ideal $\operatorname{Ann}_R(y_n)$ has at least as many irreducible factors as r_n, and hence any $r \in \operatorname{Ann}_R(y_n)$ has at least as many irreducible factors as r_n. Note that $\operatorname{Ann}_R(x_n)$ is proper since otherwise $x_1, \dots, x_{n-1}$ would generate M contradicting minimality of n.

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

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

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

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

Proof. Suppose that $x + y = 0$ for some $x \in M'$ and $y \in \langle x_n \rangle$ so that $x = a_1x_1 + \cdots + a_{n-1}x_{n-1}$ and $y = a_nx_n$ for some $a_1, \dots, a_n \in R$. Let a^*_n be such that $\langle a^*_n \rangle = \langle a_n, r_n \rangle$; $\alpha, \beta \in R$ be such that $a^*_n = \alpha a_n + \beta r_n$; and h be such that $\langle \alpha a_1, \dots, \alpha a_{n-1}, a^*_n \rangle = \langle h \rangle$. Apply Lemma 15.2 to get $y_1, \dots, y_n \in M$ with $\langle y_1, \dots, y_n \rangle = \langle x_1, \dots, x_n \rangle = M$ and

$$hy_n = \alpha a_1x_1 + \cdots + \alpha a_{n-1}x_{n-1} + a^*_nx_n = \alpha(a_1x_1 + \cdots + a_nx_n) + \beta r_nx_n = 0.$$

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

[22] Where we count the number of irreducible factors of 0 as ∞ and order $\mathbb{N}_0 \cup \{\infty\}$ in the usual way.

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

To see this last claim, suppose that $r \in \operatorname{Ann}_R(z_{n-1})$ and let h be such that $\langle h\rangle = \langle r, r_n\rangle$. Apply Lemma 15.2 to get $y_1,\dots,y_n$ with $\langle y_1,\dots,y_n\rangle = \langle z_1,\dots,z_n\rangle = M$ and $hy_n = rz_{n-1} + r_nz_n = 0$. But $h \mid r_n$ and so by minimality of the number of irreducible factors of r_n we have $h \sim r_n$ and hence $r_n \mid r$ *i.e.* $r \in \langle r_n\rangle = \operatorname{Ann}_R(z_n)$. □

16. The structure theorem for modules over PIDs and applications

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

Theorem 16.1 (Structure Theorem for modules over PIDs). *Suppose that R is a PID and M is a finitely generated R-module. Then*

(i) (Invariant factor form) *There is a sequence $a_r \mid \cdots \mid a_1$ of elements*[23] *of R with $a_r \not\sim 1$ such that*
$$M \cong (R/\langle a_1\rangle) \oplus \cdots \oplus (R/\langle a_r\rangle)$$
and the sequence $(a_i)_{i=1}^r$ is unique up to associates.

(ii) (Primary form) *There are some $s,t \in \mathbb{N}_0$, irreducible elements $p_1,\dots,p_t \in R$, and $e_1,\dots,e_t \in \mathbb{N}$, such that*
$$M \cong R^s \oplus (R/\langle p_1^{e_1}\rangle) \oplus \cdots \oplus (R/\langle p_t^{e_t}\rangle).$$

Proof. The first part is just the combination of Theorems 14.1 & 15.1, and the fact that every ideal in a PID is generated by one element.

For the second part, apply the first and then decompose each factor $R/\langle a\rangle$ further: If $a = 0$ then $R/\langle a\rangle \cong R$.

If not then since $a \not\sim 1$ and R is a UFD we have irreducible elements $q_1,\dots,q_l \in R^*$ with $q_i \not\sim q_j$ for $i \neq j$ and naturals $c_1,\dots,c_l$ such that $a \sim q_1^{c_1}\cdots q_l^{c_l}$.

Let h be a generator of $\langle q_i^{c_i}, q_j^{c_j}\rangle$ for $i \neq j$. Since R is a UFD, any prime factor of h must be an associate of something in $\{q_i\}$ and $\{q_j\}$, but since $q_i \not\sim q_j$ we have $h \sim 1$ *i..e* $\langle q_i^{c_i}\rangle + \langle q_j^{c_j}\rangle = R$ for all $i \neq j$. Thus by the Chinese Remainder Theorem for modules we have
$$R/\langle a\rangle \cong (R/\langle q_1^{c_1}\rangle) \oplus \cdots \oplus (R/\langle q_l^{c_l}\rangle)$$
as R-modules.

Hence all the factors in the invariant factor decomposition arising from modules of the form $R/\langle a_i\rangle$ with $a_i \in R^*$ can be decomposed into the desired form and the result is proved. □

There is a uniqueness statement for the primary form of the structure theorem but we do not pursue that here.

We have an immediate corollary.

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

Theorem 16.2 (Structure of finitely generated commutative groups)**.** *Suppose that G is a finitely generated commutative group. Then there are unique (non-zero) natural numbers $1 \neq d_r \mid d_{r-1} \mid \cdots \mid d_1$ and $s \in \mathbb{N}_0$ such that*

$$G \cong \mathbb{Z}^s \oplus (\mathbb{Z}/d_1\mathbb{Z}) \oplus \cdots \oplus (\mathbb{Z}/d_r\mathbb{Z}).$$

Proof. G is a $\mathbb{Z}$-module, so we may apply Theorem 16.1 to get the desired structure, writing $\mathbb{Z}/N\mathbb{Z}$ for $\mathbb{Z}/\langle N\rangle$ and $\mathbb{Z}^s$ for the s copies of $\mathbb{Z}/\langle 0\rangle$ in the given decomposition. Then uniqueness follows from the fact that $U(\mathbb{Z}) = \{-1, 1\}$. □

This result tells us a lot, for example if N is square-free then there is exactly one commutative group of order N – the cyclic group of order N. To see this, suppose that G has order N, then by the above $G \cong \mathbb{Z}^s \oplus (\mathbb{Z}/d_1\mathbb{Z}) \oplus \cdots \oplus (\mathbb{Z}/d_r\mathbb{Z})$ for some $s \in \mathbb{N}_0$ and $d_r \mid \cdots \mid d_1$. Since G is finite $s = 0$ and $N = d_r \cdots d_1$. Thus if $r > 1$ then $d_r^2 \mid N$ and hence $d_r = 1$, a contradiction. Hence $r = 1$ and G is cyclic as claimed.

Theorem 16.3 (Jordan Normal Form)**.** *Suppose that V is a finite-dimensional vector space over $\mathbb{C}$ and $T : V \to V$ is linear. Then there is a basis for V such that the matrix for T in this basis is*

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

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

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

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

Proof. We regard V as a $\mathbb{C}[X]$-module via the map $\mathbb{C}[X] \to \mathrm{End}(V); p \mapsto (v \mapsto p(T)v)$. Since $\mathbb{C}[X]$ is a PID we may apply the primary form of the structure theorem to V. We get irreducible polynomials $p_1, \ldots, p_t \in \mathbb{C}[X]$ and natural numbers $n_1, \ldots, n_t$ such that

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

is a $\mathbb{C}[X]$-linear bijection. In particular, ϕ is a $\mathbb{C}$-linear bijection but V is finite-dimensional and $\mathbb{C}[X]$ is infinite dimensional so $s = 0$. By the Fundamental Theorem of Arithmetic, every non-constant polynomial in $\mathbb{C}[X]$ has a root in $\mathbb{C}$, and so every non-constant polynomial has a degree 1 factor and so the only irreducible polynomials in $\mathbb{C}[X]$ have degree 1. Thus there are $\lambda_1, \dots, \lambda_t \in \mathbb{C}$ such that $\langle p_i(X)^{n_i} \rangle = \langle (X - \lambda_i)^{n_i} \rangle$; write $M_i := \mathbb{C}[X]/\langle (X-\lambda_i)^{n_i} \rangle$ for the ith $\mathbb{C}[X]$-module above. For each $1 \leqslant i \leqslant t$ let $(e_{i,j})_{j=1}^{n_i}$ be such that

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

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

Now we have

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

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

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

⚠ The λ_is in the theorem need not be distinct.

The fact that $\mathbb{C}$ is algebraically closed *i.e.* every polynomial with coefficients in $\mathbb{C}$ has a root in $\mathbb{C}$ is vital to the Jordan normal form, but there is another simple form available more generally.

Theorem 16.4 (Rational Canonical Form)**.** *Suppose that V is a finite-dimensional vector space over $\mathbb{F}$ and $T : V \to V$ is linear and not identically 0. Then there are monic polynomials $f_1 \mid \cdots \mid f_r$ of degree $n_t, \dots, n_1$ with f_1 non-constant, and a basis for V such that the matrix for T in this basis is*

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

where $0_{n\times m}$ is the the all zeros matrix in $M_{n\times m}(\mathbb{F})$, and $C(f)$ is[24] *the $n \times n$ matrix, called the* ***companion matrix*** *for the monic $f(X) = X^n + a_{n-1}X^{n-1} + \cdots + a_1 X + a_0$,*

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

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

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

Proof. We regard V as an $\mathbb{F}[X]$-module via the map $\mathbb{F}[X] \to \operatorname{End}(V); p \mapsto (v \mapsto p(T)v)$. Since $\mathbb{F}[X]$ is a PID we may apply the invariant factor form of the structure theorem to V. Then we get polynomials $f_1 \mid \cdots \mid f_r$ with $f_1 \not\sim 1$ and

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

an $\mathbb{F}[X]$-linear bijection. First none of the f_is is 0 since then $\mathbb{F}[X]/\langle f_i\rangle \cong \mathbb{F}[X]$ and this is an infinite dimensional vector space while V is not, and ϕ is an $\mathbb{F}$-linear bijection. Thus we may put $n_i := \deg f_i$ and may suppose that each f_i is monic (since multiplying by a unit does not change the ideal).

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

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

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

Now, we have

$$\phi(Te_{i,j}) = X\phi(e_{i,j}) = \begin{cases} \phi(e_{i,j+1}) & \text{if } j < n_i \\ -a_0\phi(e_{i,1}) - a_1\phi(e_{i,2}) - \cdots - a_{n_i-1}\phi(e_{i,n_i}) & \text{if } j = n_i \end{cases}.$$

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

For the last part we first show that for a monic polynomial f the minimal polynomial of $C(f)$ is f: By design $f(C(f)) = 0$ and so the minimal polynomial divides f and we shall be done if we can show the minimal polynomial has degree n. For $0 \leqslant r \leqslant n-1$ the first column of $C(f)^r$ is $(0, \dots, 0, 1, 0, \dots, 0)^t$ where the 1 is in the $(r+1)$th position, thus the matrices $I, C(f), \dots, C(f)^{n-1}$ are linearly independent over $\mathbb{F}$ and hence the degree of the minimal polynomial is at least n.

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

The characteristic polynomial is invariant under change of basis, and the characteristic polynomial of $C(f)$ is f (it is degree n and divisible by the minimal polynomial), hence the characteristic polynomial is the product of the characteristic polynomials of the companion matrices in the rational canonical form. It follows that it is $\prod_{i=1}^r f_i$ as required. □

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

As an example, suppose that $n \geqslant 2$ and T is the $n \times n$ all 1s matrix:

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

The image of T is one-dimensional and so by the rank-nullity theorem the kernel has dimension $n-1$. On the other hand n is an eigenvalue with eigenvector $(1, \dots, 1)^t$ and so T has a basis of eigenvectors and is diagonalisable hence the minimal polynomial is $X(X-n)$ and the characteristic polynomial is $X^{n-1}(X-n)$.

Suppose that $f_1 \mid \cdots \mid f_r$ are monic polynomials with f_1 non-constant *i.e.* of degree at least 1. Then $f_1 \cdots f_r = X^{n-1}(X-n)$ and $f_1 = X(X-n)$. Thus $f_1 \cdots f_{r-1} = X^{n-2}$ and so primality of X means that each f_i is a non-zero power of X. It follows that $f_i = X$ for all $1 \leqslant i < r$, and hence $r = n-1$ so that T is similar to

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

17. Presentations

Suppose that R is a commutative unital ring and M is a module over R. M is finitely generated if and only if there is a $k \in \mathbb{N}$ and an R-linear surjection $R^k \to M$. In particular, if $x_1, \dots, x_k$ is a generating set for M then there is a unique R-linear surjection $\psi : R^k \to M$ such that $\psi(e_i) = x_i$ for $1 \leqslant i \leqslant k$ and $e_i = (0, \dots, 0, 1, 0, \dots, 0)$ with a 1 in the ith position. We shall say that ψ is *the* R-linear surjection corresponding to the generating set $x_1, \dots, x_k$.

Given a generating set $x_1, \dots, x_k$ for M, by the First Isomorphism Theorem we have $M \cong R^k / \ker \psi$ where ψ is the corresponding R-linear surjection. If $\ker \psi$ is itself finitely generated then we say that M is **finitely presented**.

⚠This definition seems to depend on the particular generating set $x_1, \dots, x_k$ chosen rather than just on the module M, but we shall see in Proposition 17.1 that this dependence is illusory.

⚠There are finitely generated modules that are not finitely presented: Suppose that $R := \mathbb{F}[X_1, X_2, \dots]$, the ring of polynomials with countably many different variables[25]. Then $I := \langle X_1, X_2, \dots \rangle$ is an ideal in R and R/I is a finitely generated R-module, but it turns out it is not finitely presented.

While not every finitely generated module is finitely presented, if R is a PID then Theorem 16.1 tells us that there is an R-linear isomorphism

$$\psi : (R/\langle a_1 \rangle) \oplus \cdots \oplus (R/\langle a_r \rangle) \to M$$

and so putting

$$x_i := \psi(0_R + \langle a_1 \rangle, \dots, 0_R + \langle a_{i-1} \rangle, 1_R + \langle a_i \rangle, 0_R + \langle a_{i+1} \rangle, \dots, 0_R + \langle a_r \rangle)$$

for $1 \leqslant i \leqslant r$ we have that $x_1, \dots, x_r$ generates M and, moreover, the kernel of the corresponding R-linear surjection is $\langle a_1 \rangle \oplus \cdots \oplus \langle a_r \rangle$ which is generated by $(a_1, 0_R, \dots, 0_R)$, $\dots$, $(0_R, \dots, 0_R, a_r)$. Thus *every* finitely generated module over a PID is finitely presented.

[25] We did not formally define this at the start of the course, but it behaves in a fairly natural way.

Proposition 17.1. *Suppose that R is a commutative unital ring, M is a finitely presented R-module, and $\phi : R^m \to M$ is a surjective R-linear map. Then* $\ker\phi$ *is finitely generated.*

Proof. Since M is finitely presented there is an R-linear surjection $\psi : R^k \to M$ with $\ker\psi$ finitely generated. We first choose an R-linear q such that the following diagram commutes

$$\begin{tikzcd} R^k \arrow[r, "\psi"] \arrow[d, "q"] & M \\ R^m \arrow[ur, "\phi"] & \end{tikzcd}$$

To do this, note that since ϕ is surjective, for each $1 \leqslant i \leqslant k$ there is some $f_i \in R^m$ such that $\phi(f_i) = \psi(e_i)$ where $e_i = (0, \dots, 0, 1, 0, \dots, 0) \in R^k$ has the 1 in the ith position. Then put

$$q(\lambda_1 e_1 + \cdots + \lambda_k e_k) := \lambda_1 f_1 + \cdots + \lambda_k f_k \text{ for } \lambda_1, \dots, \lambda_k \in R.$$

This is a well-defined R-linear map since $e_1, \dots, e_k$ is a basis for R^k, and it has the desired property that $\phi \circ q = \psi$.

Since $\phi \circ q = \psi$ we have $x \in \ker\psi$ if and only if $q(x) \in \ker\phi$. First this tells us that q induces an R-linear map $\widetilde{q} : \ker\psi \to \ker\phi; x \mapsto q(x)$. Now consider the map

$$\Psi : \ker\phi/\operatorname{Im}\widetilde{q} \to R^m/\operatorname{Im} q; y + \operatorname{Im}\widetilde{q} \mapsto y + \operatorname{Im} q.$$

Ψ is a well-defined injection: Suppose $y, y' \in \ker\phi$. Then $y + \operatorname{Im} q = y' + \operatorname{Im} q$ if and only if $y - y' = q(x)$ for some $x \in R^k$. Since $y - y' \in \ker\phi$ we have $y - y' = q(x)$ for some $x \in R^k$ if and only if $y - y' = q(x')$ for some $x' \in \ker\phi$. Finally, $y - y' = q(x')$ for some $x' \in \ker\phi$ if and only if $x - x' \in \operatorname{Im}\widetilde{q}$.

Ψ is also a surjection: Suppose that $y \in R^m$. Then there is some $x \in R^k$ such that $\phi(y) = \psi(x)$, and hence $\phi(y - q(x)) = 0$ and hence $y - q(x) \in \ker\phi$ and $\Psi(y - q(x) + \operatorname{Im}\widetilde{q}) = y - q(x) + \operatorname{Im} q = y + \operatorname{Im} q$.

Finally Ψ is R-linear, and hence an R-linear isomorphism and we have an R-linear surjection

$$R^m \to \ker\phi/\operatorname{Im}\widetilde{q}; y \mapsto \Psi^{-1}(y + \operatorname{Im} q),$$

and $\ker\phi/\operatorname{Im}\widetilde{q}$ is finitely generated by some set $z_1 + \operatorname{Im}\widetilde{q}, \dots, z_r + \operatorname{Im}\widetilde{q}$ (where $z_1, \dots, z_r \in \ker\phi$). On the other hand $\ker\psi$ is finitely generated by $w_1, \dots, w_l$ and so $\operatorname{Im}\widetilde{q}$ is generated by $q(w_1), \dots, q(w_l)$. Thus $\ker\phi$ is generated by $z_1, \dots, z_r, q(w_1), \dots, q(w_l)$ and the result is proved. □

This argument can also be cast in terms of the Snake Lemma which is an important result in future courses on commutative algebra.

18. Elementary operations and Smith normal form

Suppose that R is a commutative unital ring, and M is finitely presented over R. If $x_1, \dots, x_m$ is a generating set for M and $r_1, \dots, r_n$ is a generating set for the the kernel of the R-linear surjection corresponding to $x_1, \dots, x_m$ then there is a matrix $A \in M_{m\times n}(R)$, called the **presentation matrix**, defined by

$$r_i = (A_{1i}, \dots, A_{mi}) \text{ for all } 1 \leqslant i \leqslant n.$$

We have $M \cong R^m/AR^n$ by the first isomorphism theorem (since AR^n is the kernel of the surjection $R^m \to M$ defined by the generating set $x_1, \dots, x_m$), and this module can be understood through the matrix A. It is the purpose of this section to examine how we can put A in a particularly nice form.

Then we put $\mathrm{GL}_n(R) := U(M_n(R))$ and we say that two $m \times n$ matrices A and B are **equivalent** if there are $S \in \mathrm{GL}_m(R)$ and $T \in \mathrm{GL}_n(R)$ such that $A = SBT$. Note, in particular, if A and B are equivalent then they are presentation matrices for isomorphic modules.

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

(i) *(Transvections)* Given $1 \leqslant i, j \leqslant n$ with $i \neq j$ and $\lambda \in R$ put $P_n(i, j; \lambda) = I + \lambda E(i, j)$. We write
$$A \xrightarrow{c_j \mapsto c_j + \lambda c_i} AP_n(i, j; \lambda).$$
to mean that the matrix A after the column operation replacing c_j by $c_j + \lambda c_i$ is the matrix A post-multiplied by $P_n(i, j; \lambda)$. This can be checked by direct calculation.

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

(ii) *(Dilations)* Given $1 \leqslant i \leqslant n$ and $u \in U(R)$ let $D_n(i; u) := I_n + (u - 1)E_n(i, i)$ so that $D_n(i; u)$ is the matrix with 1s on the diagonal except for the ith element which is u, and 0s elsewhere. As above we write
$$A \xrightarrow{c_i \mapsto uc_i} AD_n(i; u) \text{ and } A \xrightarrow{r_i \mapsto ur_i} D_m(i; u)A$$
to mean the matrix A with column c_i replaced by uc_i *etc.*

(iii) *(Interchanges)* Given $1 \leqslant i, j \leqslant n$ let $S_n(i, j) = I_n + E_n(i, j) + E_n(j, i) - E_n(i, i) - E_n(j, j)$. By
$$A \xrightarrow{c_i \leftrightarrow c_j} AS_n(i, j) \text{ and } A \xrightarrow{r_i \leftrightarrow r_j} S_m(i, j)A$$
to mean the matrix A with c_i and c_j swapped *etc.*

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

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

The subgroup of $\mathrm{GL}_n(R)$ generated by the elementary row operations is denoted $\mathrm{GE}_n(R)$. Of course $\mathrm{GE}_n(R) \leqslant \mathrm{GL}_n(R)$, and for some rings it is a proper subgroup (Cohn in [Coh66] gives an example of a PID where these groups are different), however for Euclidean domains these two groups are the same. We shall not need this fact though it could be proved by the approach below.

We say that an $m \times n$ matrix A is in **Smith normal form** if there are elements $a_1 \mid a_2 \mid \cdots \mid a_{\min\{n,m\}}$ such that

$$A = \begin{pmatrix} a_1 & 0 & \cdots \\ 0 & a_2 & \ddots \\ \vdots & \ddots & \ddots \end{pmatrix}.$$

Note the divisibility condition so that, for example,

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

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

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

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

Suppose that R is a Euclidean domain with Euclidean function f and A is an $m \times n$ matrix with entries in R. Then there is an algorithm to find a matrix $\tilde{A}$ that is equivalent to A and which is in Smith normal form.

The main step is to show that A is equivalent to a matrix of the form

$$\left(\begin{array}{c|ccc} a_1 & 0 & \cdots & 0 \\ \hline 0 & & & \\ \vdots & & A' & \\ 0 & & & \end{array}\right) \tag{18.1}$$

where $\tilde{A}$ is an $(m-1) \times (n-1)$ matrix with $a_1 \mid \tilde{A}_{ij}$ for all $1 \leqslant i \leqslant m-1$ and $1 \leqslant j \leqslant n-1$. We can then proceed recursively since the application of any row and column operations to $\tilde{A}$ do not impact that first column or row of the matrix above.

Achieving the above is a two step process: first we establish the above without the divisibility conclusion.

Extend the Euclidean function by putting $f(0) = \infty$, so that $f(0) > f(x)$ for all $x \in R^*$, and write $f(A)$ for the smallest value of $f(A_{ij})$ for $1 \leqslant i \leqslant m$ and $1 \leqslant j \leqslant n$. Finally for a vector $x \in R^k$ write $z(x)$ for the number of indices i such that $x_i = 0$.

(i) Suppose $f(A_{11}) \neq f(A)$. Then we use interchanges to move the matrix entry with smallest weight in the Euclidean function to the $(1,1)$ position in the matrix.

Specifically, let (i,j) be such that $f(A_{ij}) < f(A_{11})$ and apply the row operation $r_i \leftrightarrow r_1$ and the column operation $c_j \leftrightarrow c_1$ to get an equivalent matrix $\widetilde{A}$ where $f(\tilde{A}_{11}) < f(A_{11})$ and $f(\tilde{A}) = f(A)$.

(ii) Suppose $f(A_{11}) = f(A)$ and the top row or first column of the matrix has a non-zero entry other than its first coordinate, say it is in column j (rows are similar). Then

(a) if $A_{11} \mid A_{1j}$ we apply the column operation $c_j \mapsto c_j - (A_{1j}/A_{11})c_1$ to get an equivalent matrix $\widetilde{A}$ with $f(\widetilde{A}_{11}) = f(A_{11})$, $f(\widetilde{A}) = f(A)$, and $z(\widetilde{c}_1) + z(\widetilde{r}_1) > z(r_1) + z(c_1)$;

(b) if $A_{11} \nmid A_{1j}$ then since $f(A_{11}) \leqslant f(A_{1j})$ there is some q such that $A_{1j} = qA_{11} + r$ where $f(r) < f(A_{11})$, and we apply the column operation $c_j \mapsto c_j - qc_1$ to get an equivalent matrix $\tilde{A}$ where $f(\tilde{A}) \leqslant f(\tilde{A}_{1j}) = f(r) < f(A_{11}) = f(A)$.

At each step of the iteration we produce an equivalent matrix $\widetilde{A}$ such that either $f(\widetilde{A}) < f(A)$; or $f(\widetilde{A}) = f(A)$ and $f(\widetilde{A}_{11}) < f(A_{11})$; or $f(\widetilde{A}) = f(A)$, $f(\widetilde{A}_{11}) = f(A_{11})$, $z(\widetilde{c}_1) + z(\widetilde{r}_1) > z(c_1) + z(r_1)$.

Since $z(c) + z(r) \leqslant n + m - 1$ we see that the above algorithm must terminate at some stage when A has the form (18.1) where $f(a_1) \leqslant f(A')$. Now, if there is some (i,j) such that $a_1 \nmid A_{ij}$ then $j \neq 1$ since $A_{i1} = 0$ for all $i > 1$ and so we can apply the column operation $c_1 \mapsto c_1 + c_j$. By the Euclidean property we have $A_{ij} = qa_1 + r$ for some r with $f(r) < f(a_1)$ hence we may apply the row operation $r_i \mapsto r_i - qr_1$ to get an equivalent matrix $\widetilde{A}$ with $f(\widetilde{A}) \leqslant f(r) < f(a_1) = f(A)$.

Again, this process must terminate since the natural numbers are bounded below. The resulting matrix has the from as described in (18.1). We can now repeat the algorithm on A'. Eventually this process of passing to smaller matrices terminates since the number of rows and columns decreases by 1 at each step.

The above shows that any matrix is equivalent to a matrix in Smith normal form. However, it may not be the most efficient route. We are, of course, free to apply elementary operations as we wish to put a matrix into Smith normal form – any sequence of applications leads to an equivalent matrix since all elementary operations are in $\mathrm{GL}_n(R)$.

This argument can be used to give a proof of the structure theorem for modules over EDs, and conversely the structure theorem can be used to give a non-constructive proof of the existence of Smith Normal form. It does not, however, give an algorithm and that is the benefit of the above.

18.1. **Describing the structure of a commutative group using the SNF.**

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

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

This group is isomorphic to $(\mathbb{Z}/\langle 10\rangle)\oplus\mathbb{Z}^2$, and to show this we use the Smith normal form. First we put the relation matrix, R, into Smith normal form:

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

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

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

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

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

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

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

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

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

$$\begin{pmatrix} 1 & -1 & 8 & 1 & -25 \\ 0 & 1 & -8 & -2 & 25 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \xrightarrow{c_3 \leftrightarrow c_4} \begin{pmatrix} 1 & -1 & 1 & 8 & -25 \\ 0 & 1 & -2 & -8 & 25 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \xrightarrow{\substack{c_4 \mapsto c_4 - 4c_3 \\ c_5 \mapsto c_5 + 10c_3}}$$

$$\begin{pmatrix} 1 & -1 & 1 & 4 & -15 \\ 0 & 1 & -2 & 0 & 5 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & -4 & 10 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Similarly we can compute P:

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \xrightarrow{r_1 \leftrightarrow r_2} \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \xrightarrow{\substack{r_2 \mapsto r_2 - 2r_1 \\ r_3 \mapsto r_3 - 5r_1}} \begin{pmatrix} 0 & 1 & 0 \\ 1 & -2 & 0 \\ 0 & -5 & 1 \end{pmatrix}$$

$$\xrightarrow{r_2 \mapsto r_2 + r_3} \begin{pmatrix} 0 & 1 & 0 \\ 1 & -7 & 1 \\ 0 & -5 & 1 \end{pmatrix} \xrightarrow{r_3 \mapsto r_3 - 5r_2} \begin{pmatrix} 0 & 1 & 0 \\ 1 & -7 & 1 \\ -5 & 30 & -4 \end{pmatrix}.$$

This gives us a well-defined isomorphism

$$\begin{aligned} \phi : G &\to (\mathbb{Z}/\langle 10 \rangle) \oplus \mathbb{Z}^2 \\ z_1g_1 + \cdots + z_5g_5 &\mapsto (z_1 - 2z_2 + z_4, 4z_1 + z_3 - 4z_4, -15z_1 + 5z_2 + 10z_4 + z_5). \end{aligned}$$

For a matrix A we write $\operatorname{RowSpan}(A)$ for the $\mathbb{Z}$-module generated by the rows of A. To see that ϕ is a well-defined injection note:

$$\begin{aligned} & z_1g_1 + \cdots + z_5g_5 = z_1'g_1 + \cdots + z_5'g_5 \\ \Leftrightarrow & (z_1 - z_1', \ldots, z_5 - z_5') \in \operatorname{RowSpan}(R) && \Big) \textit{ Definition of } G \\ \Leftrightarrow & (z_1 - z_1', \ldots, z_5 - z_5') \in \operatorname{RowSpan}(PR) && \Big) \textit{ Since } P \in \mathrm{GL}_3(\mathbb{Z}) \\ \Leftrightarrow & (z_1 - z_1', \ldots, z_5 - z_5')Q \in \operatorname{RowSpan}(PRQ) && \Big) \textit{ Since } Q \in \mathrm{GL}_5(\mathbb{Z}) \\ \Leftrightarrow & (z_1 - z_1', \ldots, z_5 - z_5')Q \in \{(u, -v, 10w, 0, 0) : u, v, w \in \mathbb{Z}\} && \Big) \textit{ Design of } PRQ \\ \Leftrightarrow & \phi((z_1 - z_1')g_1 + \cdots + (z_5 - z_5')g_5) = 0 && \Big) \textit{ Definition of } \phi \\ \Leftrightarrow & \phi(z_1g_1 + \cdots + z_5g_5) = \phi(z_1'g_1 + \cdots + z_5'g_5). \end{aligned}$$

The map ϕ is also certainly $\mathbb{Z}$-linear (in fact we have already used this to some extent above). Moreover, since ϕ is well-defined and $\phi(g_5) = (0, 0, 1)$, $\phi(g_3) = (0, 1, 0)$, and $\phi(g_1 - 4g_3 + 15g_5) = (1, 0, 0)$ we see that the image of ϕ contains a generating set for the codomain and hence ϕ is a surjection. The claim that ϕ is an isomorphism is complete.

18.2. **Computing the rational canonical form using the SNF.** Suppose we wish to compute the rational canonical form of the matrix

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

We begin by putting the matrix $XI - A$ in Smith normal form over the Euclidean domain $\mathbb{Q}[X]$:

$$\begin{pmatrix} X-1 & 1 & -1 \\ 0 & X & -1 \\ 0 & -1 & X \end{pmatrix} \xrightarrow{c_1 \leftrightarrow c_2} \begin{pmatrix} 1 & X-1 & -1 \\ X & 0 & -1 \\ -1 & 0 & X \end{pmatrix} \xrightarrow[c_3 \mapsto c_3 + c_1]{c_2 \mapsto c_2 - (X-1)c_1}$$

$$\begin{pmatrix} 1 & 0 & 0 \\ X & X-X^2 & X-1 \\ -1 & X-1 & X-1 \end{pmatrix} \xrightarrow[r_3 \mapsto r_3 + r_1]{r_2 \mapsto r_2 - Xr_1} \begin{pmatrix} 1 & 0 & 0 \\ 0 & X-X^2 & X-1 \\ 0 & X-1 & X-1 \end{pmatrix} \xrightarrow{c_2 \leftrightarrow c_3}$$

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & X-1 & X-X^2 \\ 0 & X-1 & X-1 \end{pmatrix} \xrightarrow{c_3 \mapsto c_3 + Xc_2} \begin{pmatrix} 1 & 0 & 0 \\ 0 & X-1 & 0 \\ 0 & X-1 & X^2-1 \end{pmatrix}$$

$$\xrightarrow{r_3 \mapsto r_3 - r_2} \begin{pmatrix} 1 & 0 & 0 \\ 0 & X-1 & 0 \\ 0 & 0 & X^2-1 \end{pmatrix}.$$

As above we can identify the matrices $P, Q \in \mathrm{GL}_2(\mathbb{Q}[X])$ such that

$$\begin{pmatrix} 1 & 0 & 0 \\ -X & 1 & 0 \\ X+1 & -1 & 1 \end{pmatrix} \begin{pmatrix} X-1 & 1 & -1 \\ 0 & X & -1 \\ 0 & -1 & X \end{pmatrix} \begin{pmatrix} 0 & 0 & 1 \\ 1 & 1 & 1 \\ 0 & 1 & X \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & X-1 & 0 \\ 0 & 0 & X^2-1 \end{pmatrix}.$$

This form can be used to identify the rational canonical form of A: the invariant polynomials are read off the diagonal as $X - 1$ and $X^2 - 1$ and A is similar to

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

References

[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.

[Coh00] P. M. Cohn. *Introduction to ring theory*. Springer Undergraduate Mathematics Series. Springer-Verlag London, Ltd., London, 2000. doi:10.1007/978-1-4471-0475-9.

[Con] K. Conrad. Cyclicity of $\mathbb{Z}/(p)^\times$. URL https://kconrad.math.uconn.edu/blurbs/grouptheory/cyclicmodp.pdf.

[Ear19] R. Earl. Rings and modules. Lecture notes, Oxford Part A, 2019. URL https://courses.maths.ox.ac.uk/node/5393/materials.

[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.

[Hod79] W. Hodges. Krull implies Zorn. *Journal of the London Mathematical Society*, s2-19(2):285–287, 1979. doi:10.1112/jlms/s2-19.2.285.

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

[PV08] V. Perić and M. Vuković. Some examples of principal ideal domain which are not Euclidean and some other counterexamples. *Novi Sad J. Math.*, 38(1):137–154, 2008.

[Sch60] E. Schenkman. The basis theorem for finitely generated abelian groups. *The American Mathematical Monthly*, 67(8):770–771, 1960. doi:10.2307/2308656.

[Zor35] M. Zorn. A remark on method in transfinite algebra. *Bull. Amer. Math. Soc.*, 41(10):667–670, 1935. doi:10.1090/S0002-9904-1935-06166-X.

Mathematical Institute, University of Oxford, Radcliffe Observatory Quarter, Woodstock Road, Oxford OX2 6GG, United Kingdom

Email address: `tom.sanders@maths.ox.ac.uk`