ALGEBRA II: RINGS AND MODULES.
LECTURE NOTES, HILARY 2016.

KEVIN MCGERTY.

1. INTRODUCTION.

These notes accompany the lecture course ”Algebra II: Rings and modules” as lectured in Hilary term of 2016. They are an edited version of the notes which were put online in four sections during the lectures, compiled into a single file. A number of non-examinable notes were also posted during the course, and these are included in the current document as appendices.

If you find any errors, typographical or otherwise, please report them to me at mcgerty@maths.ox.ac.uk. I will also post a note summarizing the main results of the course next term.

CONTENTS

Date: March, 2016.

2. Rings: Definition and examples.

The central characters of this course are algebraic objects known as rings. Informally, a ring is any mathematical structure with a notion of addition and multiplication (the precise definition will be given shortly). As such it is a very general notion. The most basic example is $\mathbb{Z}$, the set of integers, and in this course we will largely focus on a class of rings (known as principal ideal domains or PIDs) which are in some sense very similar to $\mathbb{Z}$. By seeing how many properties of the integers naturally extend to PIDs we will not only gain a better understanding of topics like factorization, but also of questions in linear algebra, obtaining for example a canonical form for matrices over an arbitrary field[1]. Moreover, factorization in PIDs has, amongst other things, interesting applications to studying when certain equations have integer solutions: for example we will be able to say for which primes $p \in \mathbb{N}$ there are integer solutions (x, y) to the equation $x^2 + y^2 = p$.

Definition 2.1. A ring is a datum $(R, +, \times, 0, 1)$ where R is a set, $1, 0 \in R$ and $+, \times$ are binary operations on R such that

(1) R is an abelian group under $+$ with identity element 0.
(2) The binary operation $\times$ is associative and $1 \times x = x \times 1 = x$ for all $x \in R$.[2]
(3) Multiplication distributes over addition:

$$\begin{aligned} x \times (y + z) &= (x \times y) + (x \times z), \\ (x + y) \times z &= (x \times z) + (y \times z), \quad \forall x, y, z \in R. \end{aligned}$$

Just as for multiplication of real numbers or integers, we will tend to suppress the symbol for the operation $\times$, and write "." or omit any notation at all. If the operation $\times$ is commutative (*i.e.* if $x.y = y.x$ for all $x, y \in R$) then we say R is a commutative ring[3]. Sometimes[4] people consider rings which do not have a multiplicative identity. We won't. It is also worth noting that some texts require an additional axiom asserting that $1 \neq 0$. In fact it's easy to see from the other axioms that if $1 = 0$ then the ring has only one element. We will refer to this ring as the "zero ring". While it is a somewhat degenerate object, it seems unnecessary to me to exclude it.

Example 2.2. *i*) The integers $\mathbb{Z}$ form the fundamental example of a ring. As mentioned before, in some sense much of the course will be about finding an interesting class of rings which behave a lot like $\mathbb{Z}$. Modular arithmetic gives another example: if $n \in \mathbb{Z}$ then $\mathbb{Z}/n\mathbb{Z}$, the integers modulo n, form a ring with the usual addition and multiplication.

[1] The Jordan form you learned last term only applies to fields like $\mathbb{C}$ which are algebraically closed.

[2] That is, R is a monoid under $\times$ with identity element 1 if you like collecting terminology.

[3] We will try and use the letter R as our default symbol for a ring, in some books the default letter is A. This is the fault of the French, as you can probably guess.

[4] In Algebra 1 last term, the definition of a ring did not demand a multiplicative identity, nevertheless in this course we will require it. For more on this see www-math.mit.edu/~poonen/papers/ring.pdf.

ii) The subset $\mathbb{Z}[i] = \{a + ib \in \mathbb{C} : a, b \in \mathbb{Z}\}$ is easily checked to be a ring under the normal operations of addition and multiplication of complex numbers. It is known as the *Gaussian integers*. We shall see later that it shares many of the properties with the ring $\mathbb{Z}$ of ordinary integers.

iii) Any field, *e.g.* $\mathbb{Q}, \mathbb{R}, \mathbb{C}$, is a ring – the only difference between the axioms for a field and for a ring is that in the case of a ring we do not require the existence of multiplicative inverses (and that, for fields one insists that $1 \neq 0$, so that the smallest field has two elements).

iv) If k is a field, and $n \in \mathbb{N}$, then the set $M_n(\mathsf{k})$ of $n \times n$ matrices with entries in k is a ring, with the usual addition and multiplication of matrices.

v) Saying the previous example in a slightly more abstract way, if V is a vector space over a field k then $\mathrm{End}(V)$ the space of linear maps from V to V, is a ring. In this case the multiplication is given by composition of linear maps, and hence is not commutative. We will mostly focus on commutative rings in this course.

vi) Example *iv*) also lets us construct new rings from old, in that there is no need to start with a field k. Given any ring R, the set $M_n(R)$ of $n \times n$ matrices with entries in R is again a ring.

vii) Polynomials in any number of indeterminates form a ring: if we have n variables $t_1, t_2, \ldots, t_n$ and k is a field then we write $\mathsf{k}[t_1, \ldots, t_n]$ for the ring of polynomials in the variables $t_1, \ldots, t_n$ with coefficients in k.

viii) Just as in *v*), there is no reason the coefficients of our polynomials have to be a field – if R is a ring, we can build a new ring $R[t]$ of polynomials in t with coefficients in R in the obvious way. What is important to note in both this and the previous example is that polynomials are no longer functions: given a polynomial $f \in R[t]$ we may evaluate it at an $r \in R$ and thus associate it to a function from R to R, but this function may not determine f. For example if $R = \mathbb{Z}/2\mathbb{Z}$ then clearly there are only finitely many functions from R to itself, but $R[t]$ still contains infinitely many polynomials. We will construct $R[t]$ rigorously shortly.

ix) If we have two rings R and S, then we can form the *direct sum* of the rings $R \oplus S$: this is the ring whose elements are pairs (r, s) where $r \in R$ and $s \in S$ with addition and multiplication given componentwise.

x) Another way to construct new rings from old is to consider, for a ring R, functions on some set X taking values in R. The set of all such functions $R^X = \{f : X \to R$ inherits a ring structure from R by defining addition and multiplication pointwise, *i.e.* $(f+g)(x) = f(x)+g(x), (f.g)(x) = f(x).g(x)$ for all $x \in X$ (exactly as we do for $\mathbb{R}$ and $\mathbb{C}$-valued functions). The simplest example of this is when $X = \{1, 2, \ldots, n\}$ when you get[5] $R^n = \{(a_1, \ldots, a_n) : a_i \in R\}$, where we add and multiply coordinatewise.

xi) To make the previous example more concrete, the set of all functions $f : \mathbb{R} \to \mathbb{R}$ is a ring. Moreover, the set of all continuous (or differentiable, infinitely

[5] Recall, for example, that sequences of real numbers are defined to be functions $a : \mathbb{N} \to \mathbb{R}$, we just tend to write a_n for the value of a at n (and refer to it as the n-th term) rather than $a(n)$.

differentiable,...) functions also forms a ring by standard algebra of limits results.

Definition 2.3. If R is a ring, a subset $S \subseteq R$ is said to be a *subring* if it inherits the structure of a ring from R, thus we must have $0, 1 \in S$ and moreover S is closed under the addition and multiplication operations in R. It is then straight-forward to check that $(S, +, \times, 0, 1)$ satisfies the axioms for a ring.

For example, the integers $\mathbb{Z}$ are a subring of $\mathbb{Q}$, the ring of differentiable functions from $\mathbb{R}$ to itself is a subring of the ring of all functions from $\mathbb{R}$ to itself. The ring of Gaussian integers is a subring of $\mathbb{C}$, as are $\mathbb{Q}$,$\mathbb{R}$ (the latter two being fields of course). Recall that for a group G containing a subset H, the *subgroup criterion* says that H is a subgroup if and only if it is nonempty and whenever $h_1, h_2 \in H$ we have $h_1 h_2^{-1} \in H$ (here I'm writing the group operation on G multiplicatively). We can use this to give a similar criterion for a subset of a ring to be a subring.

Lemma 2.4 (Subring criterion)**.** *Let R be a ring and S a subset of R, then S is a subring if and only if $1 \in S$ and for all $s_1, s_2 \in S$ we have $s_1 s_2, s_1 - s_2 \in S$.*

Proof. The condition that $s_1 - s_2 \in S$ for all $s_1, s_2 \in S$ implies that S is an additive subgroup by the subgroup test (note that as $1 \in S$ we know that S is nonempty). The other conditions for a subring hold directly. □

When studying any kind of algebraic object[6] it is natural to consider maps between those kind of objects which respect their structure. For example, for vector spaces the natural class of maps are linear maps, and for groups the natural class are the group homomorphisms. The natural class of maps to consider for rings are defined similarly:

Definition 2.5. A map $f\colon R \to S$ between rings R and S is said to be a (*ring*) *homomorphism* if

(1) $f(1_R) = 1_S$,
(2) $f(r_1 + r_2) = f(r_1) + f(r_2)$,
(3) $f(r_1.r_2) = f(r_1).f(r_2)$,

where strictly speaking we might have written $+_R$ and $+_S$ for the addition operation in the two different rings R and S, and similarly for the multiplication operation[7]. Partly because the meaning is clear from context and partly because otherwise the notation becomes hard to read, we will (as is conventional) use the same notation for the addition and multiplication in all rings. Note that it follows from (2) that $f(0) = 0$.

[6] Or more generally any mathematical structure: if you're taking Topology this term then continuous maps are the natural maps to consider between topological spaces, similarly in Integration you consider measurable functions: loosely speaking, you want to consider maps which play nicely with the structures your objects have, be that a topology, a vector space structure, a ring structure or a measure.

[7] though since I've already decided to suppress the notation for it, it's hard to distinguish the two when you suppress both...

It is worth checking which of our examples of rings above are subrings of another example, *e.g.* $\mathbb{R}$ and $\mathbb{Z}[i]$ are both subrings of $\mathbb{C}$.

If $f\colon R \to S$ is a ring homomorphism, it is easy to see that its image

$$\mathrm{im}(f) = f(R) = \{s \in S : \exists r \in R, f(r) = s\}$$

is a subring of S. If it is all of S we say that f is surjective. We say that $f\colon R \to S$ is an isomorphism if there is a homomorphism $g\colon S \to R$ such that $f \circ g = \mathrm{id}_S$ and $g \circ f = \mathrm{id}_R$. It is easy to check that f is an isomorphism if and only if it is a bijection (that is, to check that the set-theoretic inverse of f is automatically a ring homomorphism – you probably did a similar check for linear maps between vector spaces before.)

Example 2.6. *i*) For each positive integer n, there is a natural map from $\mathbb{Z}$ to $\mathbb{Z}/n\mathbb{Z}$ which just takes an integer to its equivalence class modulo n. The standard calculations which show that modular arithmetic is well-defined exactly show that this map is a ring homomorphism.

ii) Let V be a k-vector space and let $\alpha \in \mathrm{End}_{\mathsf{k}}(V)$. Then $\phi\colon \mathsf{k}[t] \to \mathrm{End}_{\mathsf{k}}(V)$ given by $\phi(\sum_{i=0}^n a_i t^i) = \sum_{i=0}^n a_i \alpha^i$ is a ring homomorphism. Ring homomorphisms of this type will reveal the connnection between the study of the ring $\mathsf{k}[t]$ and linear algebra. (In a sense you saw this last term when defining things like the minimal polynomial of a linear map, but we will explore this more fully in this course.)

iii) The inclusion map $i\colon S \to R$ of a subring S into a ring R is a ring homomorphism.

iv) Let $A = \{\begin{pmatrix} a & -b \\ b & a \end{pmatrix} : a, b \in \mathbb{R}\}$. It is easy to check this A is a subring of $\mathrm{Mat}_2(\mathbb{R})$. The map $\phi\colon \mathbb{C} \to A$ given by $a + ib \mapsto \begin{pmatrix} a & -b \\ b & a \end{pmatrix}$ is a ring isomorphism. (This homomorphism arises by sending a complex number z to the map of the plane to itself given by multiplication by z.)

The first of the above examples has an important generalisation which shows that any ring R in fact has a smallest subring: For $n \in \mathbb{Z}_{\geq 0}$ set $n_R = 1 + 1 + \ldots + 1$ (that is, 1, added to itself n times), and for n a negative integer $n_R = -(-n)_R$. The problem sheet asks you to check that $\{n_R : n \in \mathbb{Z}\}$ is a subring of R, and indeed that the map $n \mapsto n_R$ gives a ring homomorphism from $\phi\colon \mathbb{Z} \to R$. Since a ring homomorphism is in particular a homomorphism of the underlying abelian groups under addition, using the first isomorphism theorem for abelian groups we see that $\{n_R : n \in \mathbb{Z}\}$, as an abelian group, is isomorphic to $\mathbb{Z}/d\mathbb{Z}$ for some $d \in \mathbb{Z}_{\geq 0}$. Since any subring S of R contains 1, and hence, since it is closed under addition, n_R for all $n \in \mathbb{Z}$, we see that S contains the image of ϕ, so that the image is indeed the smallest subring of R.

Definition 2.7. The integer d defined above is called the *characteristic* of the ring R.

2.1. **Polynomial Rings.** The remark above that in general polynomials with coefficients in a ring cannot always be viewed as functions might have left you wondering what such a polynomial actually is. In other words, what do we mean when we say $\mathsf{k}[t]$ is a ring where "t is an indeterminate."

To make sense of this, suppose that R is a ring, and consider the set[8] $R^{\mathbb{N}} = \{f \colon \mathbb{N} \to R\}$. We already saw this set has the structure of a ring using "pointwise" addition and multiplication, but it also has a more interesting multiplication operation, $(f, g) \mapsto f * g$ where

$$(f * g)(n) = \sum_{k+l=n} f(k).g(l) = \sum_{k=0}^{n} f(k)g(n-k), \quad n \in \mathbb{N}.$$

Since $k, l \in \mathbb{N}$, the above sum is finite, so we get a well defined function $f * g$ (*c.f.* convolution of real-valued functions.). It follows directly from the definitions that $(R^{\mathbb{N}}, +, *, \mathbf{1}, \mathbf{0})$ is a ring, where $\mathbf{0}$ is the constant function taking value 0, while $\mathbf{1}(n) = 1$ if $n = 0$ and is zero otherwise. We will write $R[[t]]$ for this ring.

Let $R[t] = \{f \in R^{\mathbb{N}} : \exists N > 0, f(n) = 0 \quad \forall n > N\}$. $R[t]$ is a subring of $R[[t]]$ (check this) and it is the ring of polynomials with coefficients in R which we wanted to construct. To see why, let $t \in R^{\mathbb{N}}$ be the function such that $t(1) = 1$ and $t(n) = 0$ for $n \neq 1$. By induction it is easy to check that $t^k = t * \ldots * t$ (k times) is such that $t^k(n) = 1$ if $k = n$ and is zero otherwise. Now suppose that $f \in R[t]$ and say $f(n) = 0$ for all $n > N$. Then we see that $f = \sum_{n=0}^{N} f(n)t^n$. In fact in general for $f \in R[[t]]$ we have[9] $f = \sum_{n \geq 0} f(n)t^n$.

Remark 2.8. Our definition of $*$ makes sense on a bigger set than $R^{\mathbb{N}}$: If we take the set $S = \{f \colon \mathbb{Z} \to R : \exists N \in \mathbb{Z}, f(n) = 0 \quad \forall n < N\}$ then you can check that if $f, g \in S$ the function

$$(f * g)(n) = \sum_{k \in \mathbb{Z}} f(k)g(n-k)$$

is well-defined (in that only finitely many terms on the right are non-zero for any given integer n. This ring is denoted $R((t))$, and it turns out that if R is a field, so is $R((t))$.

Note also that we have a ring homomorphism $\iota_R \colon R \to R[t]$ given by $\iota_R(a) = a.\mathbf{1}$ which is injective, thus we can view R as a subring of $R[t]$.

The fundamental property of polynomial rings is that they have natural "evaluation" homomorphisms: to specify a homomorphism from a polynomial ring $R[t]$ to a ring S you only need to say what happens to the elements of R (the coefficients) and what happens to t. We formalise this in the following Lemma.

[8] Here $\mathbb{N} = \mathbb{Z}_{\geq 0}$ the set of non-negative integers. In some places (though hopefully not other parts of these notes) $\mathbb{N}$ denotes the strictly positive integers.

[9] At first sight the right hand side of this expression looks like it might not make sense because it is an infinite sum. However it does give a well-defined function on $\mathbb{N}$ because on any element of $\mathbb{N}$ only finitely many terms (in fact exactly one) in the infinite sum are nonzero.

Lemma 2.9. *(Evaluation homomorphisms.) Let R, S be rings and $\phi\colon R \to S$ a ring homomorphism. If $s \in S$ then there is an unique ring homomorphism $\Phi\colon R[t] \to S$ such that $\Phi \circ \iota_R = \phi$ (where $\iota_R\colon R \to R[t]$ is the inclusion of R into $R[t]$) and $\Phi(t) = s$.*

Proof. Any element of $R[t]$ has the form $\sum_{i=0}^{n} a_i t^i$, $(a_i \in R)$, hence if Θ is any homomorphism satisfying $\Theta \circ i = \phi$ and $\Theta(t) = s$ we see that

$$\Theta(\sum_{i=0}^{n} a_i t^i) = \sum_{i=0}^{n} \Theta(a_i t^i) = \sum_{i=0}^{n} \Theta(a_i)\Theta(t^i) = \sum_{i=0}^{n} \phi(a_i)s^i,$$

Hence Θ is uniquely determined. To check there is indeed such a homomorphism we just have to check that the function $\Phi(\sum_{i=0}^{n} a_i t^i) = \sum_{i=0}^{n} \phi(a_i)s^i$ is indeed a homomorphism, but this is straight-forward from the definitions. □

3. Basic properties.

From now on, unless we explicitly state otherwise, all rings will be assumed to be commutative.

Now that we have seen some examples of rings, we will discuss some basic properties of rings and their elements. Note that it is a routine exercise[10] in axiom grubbing to check that, for any ring R, we have $a.0 = 0$ for all $a \in R$. The next definition records the class of rings for which this is the only case in which the product of two elements is zero.

Definition 3.1. If R is a ring, then an element $a \in R\backslash\{0\}$ is said to be a *zero-divisor* if there is some $b \in R\backslash\{0\}$ such that $a.b = 0$. A ring which is not the zero ring and has no zero-divisors is called an *integral domain*. Thus if a ring is an integral domain and $a.b = 0$ then one of a or b is equal to zero.

Another way to express the fact that a ring is an integral domain is observe that it is exactly the condition which permits cancellation[11], that is, if $x.y = x.z$ then in an integral domain you can conclude that either $y = z$ or $x = 0$. This follows immediately from the definition of an integral domain and the fact that $x.y = x.z \iff x.(y - z) = 0$, which follows from the distributive axiom.

Example 3.2. If R is a ring, then R^2 is again a ring, and $(a, 0).(0, b) = (0, 0)$ so that $(a, 0)$ and $(0, b)$ are zero-divisors. The (noncommutative) ring of $n \times n$ matrices $M_n(\mathsf{k})$ for a field k also has lots of zero divisors (even though a field k has none). The integers modulo n have zero-divisors whenever n is not prime.

On the other hand, it is easy to see that a field has no zero-divisors. The integers $\mathbb{Z}$ are an integral domain (and *not* a field). Slightly more interestingly, if R is an

[10] It's a good idea to try and check that the axioms for a ring do indeed imply that you can perform the standard algebraic manipulations you are used to, so things like $0.x = 0$ hold in any ring. None of the checks you have to do are very exciting, so it's best to pick a few such statements. One operation you have to be careful about however, is cancellation (but then again you already should be aware of this issue from matrix algebra).

[11] Except for the assertion the ring is not the zero ring, the zero ring having cancellation vacuously.

integral domain, then $R[t]$ is again an integral domain. Moreover, the same is true of $R[[t]]$. (You are asked to check this in the problem sheet.)

Recall the characteristic of a ring defined in the last lecture.

Lemma 3.3. *Suppose that R is an integral domain. Then any subring S of R is also an integral domain. Moreover, char(R), the characteristic of R, is either zero or a prime $p \in \mathbb{Z}$.*

Proof. It is clear from the definition that a subring of an integral domain must again be an integral domain. Now from the definition of the characteristic of a ring, if $\text{char}(R) = n > 0$ then $\mathbb{Z}/n\mathbb{Z}$ is a subring of R. Clearly if $n = a.b$ where $a, b \in \mathbb{Z}$ are both greater than 1, then $a_R.b_R = 0$ in R with neither a_R nor b_R zero, thus both are zero divisors. It follows that if R is an integral domain then $\text{char}(R)$ is zero or a prime. □

Note that in particular, the characteristic of a field is always zero or a prime.

Recall that in a ring we do not require that nonzero elements have a multiplicative inverse[12]. Nevertheless, because the multiplication operation is associative and there is a multiplicative identity, the elements which happen to have multiplicative inverses form a group:

Definition 3.4. Let R be a ring. The subset

$$R^\times = \{r \in R : \exists s \in R, r.s = 1\},$$

is called the group of *units* in R – it is a group under the multiplication operation × with identity element 1.

Example 3.5. The units in $\mathbb{Z}$ form the group $\{\pm 1\}$. On the other hand, if k is a field, then the units $\mathsf{k}^\times = \mathsf{k}\backslash\{0\}$. If $R = M_n(\mathsf{k})$ then the group of units is $\text{GL}_n(\mathsf{k})$.

Remark 3.6. In our example of $\mathbb{Z}/n\mathbb{Z}$ notice that this ring either has zero-divisors (when n is composite) or is a field (when n is prime). In fact this is dichotomy holds more generally: a *finite* integral domain is always a field. (See the problem sheet for more details.)

3.1. **The field of fractions.** We first describe the construction of the rational numbers from the integers: A rational number is, of course, a ratio of two integers. To say formally what this means we start with the set $Q(\mathbb{Z}) = \{(a, b) \in \mathbb{Z}^2 : b \neq 0\}$. Then we define a relation $\sim$ on $Q(\mathbb{Z})$ by setting $(a, b) \sim (c, d)$ if $ad = bc$. (To see where this comes from, notice that it expresses the fact that $a/b = c/d$ without using division). It is clear that $\sim$ is reflexive and symmetric, and an easy calculation shows that it is transitive, so that it is an equivalence relation: Indeed suppose that $(a, b) \sim (c, d)$ and $(c, d) \sim (e, f)$. Then we have $ad = bc$ and $cf = de$ and need to check that $(a, b) \sim (e, f)$, that is, $af = be$. But since $d \neq 0$ we have

$$af - be = 0 \iff d.(af - be) = 0 \iff (ad).f - b.(de) = 0 \iff (bc).f - b.(cf) = 0,$$

[12] As noted above, the axioms for a ring imply that $0.x = 0$ for all $x \in R$, thus the additive identity cannot have a multiplicative inverse, hence the most we can ask for is that every element of $R\backslash\{0\}$ does – this is exactly what you demand in the axioms for a field.

as required

The set of equivalence classes[13] is denoted $\mathbb{Q}$, and we write $\frac{a}{b}$ for the equivalence class containing (a, b). You can then check that the formulas

$$\frac{a}{b} + \frac{c}{d} = \frac{ad+bc}{bd}, \quad \frac{a}{b}.\frac{c}{d} = \frac{ac}{bd},$$

give a well-defined addition and multiplication on the set of equivalence classes, forming the field $\mathbb{Q}$. The fact that it is a field and not just a ring follows because $\frac{a}{b} = 0$ exactly when $a = 0$, and thus if $\frac{a}{b} \neq 0$ it has inverse $\frac{b}{a}$. The details of verifying the operations are independent of the representatives you chose in the equivalence classes take some time to write down rigorously, but there are no surprises in the process.

The interesting thing to notice here is that this construction also makes sense for an arbitrary integral domain: given an integral domain R, the relation on $Q(R) = \{(a, b) \in R^2 : b \neq 0\}$ given by $(a, b) \sim_R (c, d)$ if $ad = bc$ is an equivalence relation and the same formulas give the set of equivalence classes $F(R)$ the structure of a field. At various points you need to use cancellation (for example in showing the relation $\sim$ is transitive) which is why the construction only works for integral domains, and not more general rings[14].

Definition 3.7. The field $F(R)$ is known as the *field of fractions* of R. The ring R embeds into $F(R)$ via the map $r \mapsto \frac{r}{1}$, thus an integral domain is naturally a subring of its field of fractions.

The rational numbers are the smallest field which contain the integers, in the sense that any field which contains $\mathbb{Z}$ automatically contains the rationals (essentially because if you are a field and contain $m, n \in \mathbb{Z}$ then you contain $\frac{1}{n}$ and so $\frac{m}{n}$). This is in fact the characterising property of the field of fractions, which can be formalised as follows:

Proposition 3.8. *Let* k *be a field and let* $\theta \colon R \to \mathsf{k}$ *be an embedding (that is, an injective homomorphism). Then there is a unique injective homomorphism* $\tilde{\theta} \colon F(R) \to \mathsf{k}$ *extending* θ *(in the sense that* $\tilde{\theta}_{|R} = \theta$ *where we view* R *as a subring of* $F(R)$ *via the above embedding).*

Proof. (*non-examinable*): Suppose that $f \colon F(R) \to \mathsf{k}$ was such a homomorphism. Then by assumption $f(\frac{a}{1}) = \theta(a)$, and since homomorphism of rings respect multiplicative inverses this forces $f(\frac{1}{a}) = \theta(a)^{-1}$. But then, again because f is supposed to be a homomorphism, we must have $f(\frac{a}{b}) = f(\frac{a}{1}.\frac{1}{b}) = f(\frac{a}{1}).f(\frac{1}{b}) = \theta(a).\theta(b)^{-1}$. Thus if f exists, it has to be given by this formula.

The rest of the proof consists of checking that this recipe indeed works: Given $(a, b) \in R \times R\backslash\{0\}$ first define $\Theta(a, b) = \theta(a).\theta(b)^{-1}$. Then it is easy to check that Θ is constant on the equivalence classes of $\sim$ the relation defining $F(R)$, so that it

[13] Notice that we have thus *defined* a ratio to be the set of all pairs of integers (a, b) (with b non-zero) which are in that ratio.

[14] There is a more sophisticated construction which works for more general rings however, but we leave that for later courses.

induces a map $\tilde{\theta}\colon F(R) \to \mathsf{k}$. Finally it is straight-forward to see that this map is a homomorphism extending θ as required. □

Remark 3.9. Notice that this Proposition implies that any field k of characteristic zero contains a (unique) copy of the rationals. Indeed by definition of characteristic, the unique homomorphism from $\mathbb{Z}$ to k is an embedding, and the above theorem shows that it therefore extends uniquely to an embedding of $\mathbb{Q}$ into k as claimed.

4. Ideals and Quotients.

From now on we will assume all our rings are commutative. In this section we study the basic properties of ring homomorphisms, and establish an analogue of the "first isomorphism theorem" which you have seen already for groups. Just as for homomorphisms of groups, homomorphisms of rings have kernels and images.

Definition 4.1. Let $f\colon R \to S$ be a ring homomorphism. The *kernel* of f is

$$\ker(f) = \{r \in R : f(r) = 0\},$$

and the *image* of f is

$$\mathrm{im}(f) = \{s \in S : \exists r \in R, f(r) = s\}.$$

Just as for groups, the image of a homomorphism is a subring of the target ring. For kernels the situation is a little different. In the case of groups, kernels of homomorphisms are subgroups, but not any subgroup is a kernel – the kernels are characterised intrinsically by the property of being normal (*i.e.* perserved by the conjugation action of the group). We will show that the kernels of ring homomorphisms can similarly be characterised intrinsically, but the situation, because we have two binary operations, is slightly different: a kernel is both more and less than a subring. Indeed since homomorphisms are required to send 1 to 1, the kernel never contains 1 unless it is the entire ring, thus a kernel is *not* a subring unless the target of the homomorphism is the zero ring[15]. However, it is closed under addition and mulitplication (as is straight-forward to check) and because $0.x = 0$ for any x, it in fact obeys a stronger kind of closure with respect to multiplication[16]: If $x \in \ker(f)$ and $r \in R$ is any element of R, then $f(x.r) = f(x).f(r) = 0.f(r) = 0$ so that $x.r \in \ker(f)$. This motivates the following definition:

Definition 4.2. Let R be a ring. A subset $I \subseteq R$ is called an *ideal* if it is a subgroup of $(R, +)$ and moreover for any $a \in I$ and $r \in R$ we have $a.r \in I$. We write $I \lhd R$ to denote the fact that I is an ideal in R.

Lemma 4.3. *If $f\colon R \to S$ is a homomorphism, then ker(f) is an ideal. Moreover if $I \subseteq R$ then I is an ideal if and only if it is nonempty, closed under addition, and closed under multiplication by arbitrary elements of R.*

[15] It's also worth noticing that if R is the zero ring, and $\theta\colon R \to S$ is a ring homomorphism we must have $S = R$, since we insist that a ring homomorphism preserves the additive and multiplicative identities.

[16] This is analogous to the fact that kernels of group homomorphisms, being normal, are loosely speaking "more closed" than arbitrary subgroups.

Proof. This is immediate from the definitions. For the moreover part, we just need to check that I is closed under taking additive inverses. But this follows from the fact that it is closed under multiplication by any element of R since $-x = (-1).x$ for any $x \in R$. □

Note that if I is an ideal of R which contains 1 then $I = R$. We will shortly see that in fact any ideal is the kernel of a homomorphism. First let us note a few basic properties of ideals: First we need some notation: if X, Y are any subsets of R define

$$X + Y = \{x + y : x \in X, y \in Y\}, \quad X.Y = \{\sum_{k=1}^{n} x_k.y_k : n \in \mathbb{Z}_{\geq 0}, x_k \in X, y_k \in Y, 1 \leq k \leq n\}.$$

Note that $X.Y$ is closed under addition. By convention we will take $X.Y = \{0\}$ if either X or Y is empty.

Lemma 4.4. *Let R be a ring, and I, J ideals in R and X any subset of R. Then $I + J$, $I \cap J$ and IX are ideals. Moreover we have $IJ \subseteq I \cap J$ and $I, J \subseteq I + J$.*

Proof. For $I + J$ it is clear that this is an abelian subgroup of R, while if $i \in I$, $j \in J$ and $r \in R$, then $r(i + j) = (r.i) + (r.j) \in I + J$ as both I and J are ideals, hence $I + J$ is an ideal. Checking $I \cap J$ is an ideal is similar but easier. To see that IX is an ideal, note that it is clear that the sum of two elements of IX is of the same form, and if $\sum_{k=1}^{n} i_k x_k \in IX$ then

$$r.\sum_{k=1}^{n} i_k x_k = \sum_{k=1}^{n} (r.i_k).x_k \in IX.$$

Thus by the moreover part of Lemma 4.3, IX is an ideal[17]. The containments are all clear once you note that if $i \in I$ and $j \in J$ then ij in in $I \cap J$ because both I and J are ideals. □

In fact given a collection of ideals $\{I_\alpha : \alpha \in A\}$ in a ring R, their intersection $\bigcap_{\alpha \in A} I_\alpha$ is easily seen to again be an ideal. This easy fact is very useful for the following reason:

Definition 4.5. Given *any* subset T of R, one can define

$$\langle T \rangle = \bigcap_{T \subseteq I} I$$

(where I is an ideal) the ideal *generated* by T. We can also give a more explicit "from the ground up" description of the ideal generated by a subset X:

Lemma 4.6. *Let $T \subseteq R$ be a nonempty subset. Then we have*

$$\langle T \rangle = R.T.$$

[17] This is one reason for the convention that $X.Y = \{0\}$ if either of X or Y is empty – it ensures $I.X$ an ideal even when X is empty

Proof. We have already seen that $R.T$ is an ideal (since R itself is an ideal). We first check that $R.T$ is contained in any ideal I which contains T. But if $\{x_1, \ldots, x_k\} \subseteq T \subseteq J$ and $r_1, \ldots, r_k \in R$, then since J is an ideal certainly $r_k x_k \in J$ and hence $\sum_{k=1}^{n} r_k x_k \in J$. Since the x_k, r_k and $n \in \mathbb{N}$ were arbitrary it follows that $R.T \subseteq J$.

It follows that $R.T \subseteq \bigcap_{I \triangleleft R, T \subseteq R} I$, but since $R.T$ is itself an ideal containing T, clearly the intersection lies in $R.T$ also, so we have the desired equality.

□

This is completely analogous to the notion of the "span" of a subset in a vector space. If I and J are ideals, it is easy to see that $I + J = \langle I \cup J \rangle$. In the case where $T = \{a\}$ consists of a single element, we often write aR or[18] Ra for $\langle a \rangle$.

Remark 4.7. Note that in the above, just as for span in a vector space, there is no need for the set X to be finite.

Remark 4.8. Note that if $T \subset R$ is a subset of a ring R we can also consider the subring which it generates: the intersection of subrings is again a subring[19], so we may set

$$\langle T \rangle_s = \bigcap_{T \subseteq S} S,$$

where the subscript "s" is supposed to denote subring. I leave it as an exercise to find a "ground up" description of $\langle T \rangle_s$.

Definition 4.9. If an ideal is generated by a single element we say it is *principal*. Two elements $a, b \in R$ are said to be *associates* if there is a unit $u \in R^{\times}$ such that $a = u.b$. (This is an equivalence relation on the elements of R).

If $I = \langle a \rangle$ then just knowing I does not quite determine a, but it almost does, at least if R is an integral domain. The notion of associate elements lets us make this precise.

Lemma 4.10. *Let R be a domain. Then if I is a principal ideal, the generators*[20] *of I are associates, and any associate of a generator is again a generator. Thus the generators of a principal ideal form a single equivalence class of associate elements of R.*

Proof. If $I = \{0\} = \langle 0 \rangle$ the claim is immediate, so assume $I \neq \{0\}$ and hence any generator is nonzero. Let $a, b \in R$ be generators of I, so $I = \langle a \rangle = \langle b \rangle$. Since $a \in \langle b \rangle$, there is some $r \in R$ with $a = r.b$, and similarly as $b \in \langle a \rangle$ there is some s with $b = s.a$. It follows that $a = r.b = r(s.a) = (r.s)a$, hence $a(1 - r.s) = 0$, and so since $a \neq 0$ and R is an integral domain, $r.s = 1$, that is, r and s are units.

[18] Since R is commutative $Ra = \{r.a : r \in R\} = \{a.r : r \in R\} = aR$.

[19] Note also that this is a pretty general way of defining the widget "generated" by a subset of a given object: whatever a widget is, provided the intersection of widgets is again a widget, then if S is some subset of your object, the widget it "generates" is the intersection of all widgets which contain S – the stability under taking intersections ensures this intersection is still a widget, and it is thus the smallest widget containing S. The closure of sets in topological spaces, the ideal generated by a set in a ring and the subring generated by a set in a ring are all defined in this way.

[20] *i.e.* the elements $a \in R$ such that $I = \langle a \rangle$.

Finally if $I = \langle a \rangle$ and $b = u.a$ where $u \in R^{\times}$, then certainly $b \in \langle a \rangle = I$ so that $\langle b \rangle \subseteq I$, but also if $x \in I$, then $x = r.a$ for some $r \in R$ and hence $x = r.(u^{-1}.b) = (r.u^{-1}).b$ so that $x \in \langle b \rangle$, and hence $I \subseteq \langle b \rangle$. It follows $I = \langle b \rangle$ as required. □

For example in $\mathbb{Z}$ we will see that the ideals are all of the form $\langle n \rangle$ and the integer n is determined up to sign by the ideal $\langle n \rangle$ (the units in $\mathbb{Z}$ being exactly $\{\pm 1\}$).

4.1. **The quotient construction.** In order to show that ideals and kernels of ring homorphisms are the same thing, we now study a notion of quotient for rings, similar to the quotients of groups and vector spaces which you have already seen. This is one of the most important constructions in the course. The notion of a quotient object is a fundamental one (not just in algebra, but also in topology and many other subjects). It is a subtle thing that usually takes some time to become accustomed to, but it is absolutely essential.

Suppose that R is a ring and that I is an ideal in R. Then since $(I, +)$ is a subgroup of the abelian group $(R, +)$, we may form the quotient group $(R/I, +)$. We briefly recall the construction of the group $(R/I, +)$: The relations $r \sim s$ if $r - s \in I$ is easily checked to be an equivalence relation[21], and the equivalence classes are the cosets $\{r + I : r \in R\}$. We want to endow R/I with the structure of an abelian group in such a way that the map $q\colon \ \to R/I$ sending an element $r \in R$ to its equivalence class $r + I \in R/I$ is a group homomorphism. This condition forces us to require

$$(r_1 + I) + (r_2 + I) = (r_1 + r_2) + I,$$

so the only thing to do is to check that this formula makes sense, that is, to check that it is independent of the choice of representatives $r_1.r_2$. To see this, suppose that $s_j \in r_j + I$ for $j = 1, 2$ are two other choices of representatives. Then there are elements $i_1, i_2 \in I$ such that $s_j = r_j + i_j$ $(j = 1, 2)$, and hence $s_1 + s_2 = (r_1 + r_2) + (i_1 + i_2) \in (r_1 + r_2) + I$ since $i_1 + i_2 \in I$.

Now that we have shown R/I inherits a binary operation from R, it is easy to see that that operation makes R/I into an abelian group with identity $0 + I$. We now want to show that R/I has the structure of a ring where the multiplication is again induced from that on R, so that the map $q\colon R \to R/I$ is actually a ring homomorphism. Again there is at most one possibility for the multiplication on R/I which could satisfy this condition: we must have

$$(r_1 + I) \times (r_2 + I) = r_1.r_2 + I,$$

or said in terms of the quotient map $q\colon R \to R/I$ we must have $q(r_1).q(r_2) = q(r_1.r_2)$. Again the issue is whether this is indeed a well-defined operation, *i.e.* independent of the choice of representatives r_1, r_2. Thus as before take $s_j = r_j + i_j$ some other representatives for the cosets $r_j + I$, $(j = 1, 2)$. Then we have

$$s_1.s_2 = (r_1 + i_1).(r_2 + i_2) = r_1.r_2 + (i_1 r_2 + r_1 i_2 + i_1 i_2) \in r_1.r_2 + I$$

since $i_1 r_2, r_1 i_2, i_1 i_2$ all lie in I since I is an ideal. It follows that we have a well-defined binary operation on R/I coming from the multiplication in R also.

[21] And you have done similar checks for groups and vector spaces before.

Theorem 4.11. *The datum $(R/I, +, \times, 0 + I, 1 + I)$ defines a ring structure on R/I and moreover the map $q\colon R \to R/I$ given by $q(r) = r + I$ is a surjective ring homomorphism. Moreover the kernel of q is the ideal I.*

Proof. Checking each axiom is an easy consequence of the fact that the binary operations $+, \times$ on R/I are defined by picking arbitrary representatives of the cosets, computing up in the ring R and then taking the coset of the answer (the important part of the definitions being that this last step is well-defined). Thus for example, to check $\times$ is associative, let C_1, C_2, C_3 be elements of R/I and choose $r_1, r_2, r_3 \in R$ such that $C_i = q(r_i) = r_i + I$ for $i = 1, 2, 3$. Then

$$\begin{aligned} C_1 \times (C_2 \times C_3) &= q(r_1) \times (q(r_2) \times q(r_3)) \\ &= q(r_1) \times q(r_2 r_3) = q(r_1.(r_2 r_3)) \\ &= q((r_1 r_2).r_3)) = q(r_1 r_2) \times q(r_3) \\ &= (q(r_1) \times q(r_2)) \times q(r_3) = (C_1 \times C_2) \times C_3. \end{aligned}$$

where in going from the second to the third line we use the associativity of multiplication in R. Checking the other axioms is similarly straight-forward. Finally, the map $q\colon R \to R/I$ is clearly surjective, and that it is a homomorphism is also immediate from the definitions. Clearly $q(r) = 0 \in R/I$ precisely when $q(r) = r + I = 0 + I$, that is precisely when $r \in I$. Thus $\ker(q) = I$ as required. □

The map $q\colon R \to R/I$ is called the *quotient homomorphism* (or quotient map). The next corollary establishes the answer to the question we started this section with: what are the subsets of R which are kernels of homomorphisms from R? We already noted that any kernel is an ideal, but the construction of quotients now gives us the converse:

Corollary 4.12. *The ideals in R are exactly the kernels of the set of homomorphisms with domain R.*

Proof. We have already seen that the kernel of a ring homomorphism is always an ideal so it only remains to show that any ideal is the kernel of some homomorphism. But this is exactly what the previous theorem shows: If I is an ideal and $q\colon R \to R/I$ is the quotient map then q is a ring homomorphism and $\ker(q) = I$. □

For our next result about quotient rings, it may be helpful to compare with the following result from last term's linear algebra about quotient vector spaces: If $T\colon V \to W$ is a linear map, and $U < V$ is a subspace, then T induces a linear map $\bar{T}\colon V/U \to W$ on the quotient space V/U if and only if $U \subseteq \ker(T)$.

Theorem 4.13. *(Universal Property of Quotients.) Suppose that R is a ring, I is an ideal of R, and $q\colon R \to R/I$ the quotient homomorphism. If $\phi\colon R \to S$ is a ring homomorphism such that $I \subseteq \ker(\phi)$, then there is a unique ring homomorphism $\bar{\phi}\colon R/I \to S$ such that $\bar{\phi} \circ q = \phi$. That is, the following diagram commutes:*

$$\begin{tikzcd} R \arrow[rr, "\phi"] \arrow[dr, "q"'] & & S \\ & R/I \arrow[ur, "\bar{\phi}"'] & \end{tikzcd}$$

Moreover $ker(\bar{\phi})$ *is the ideal* $ker(\phi)/I = \{m + I : m \in ker(\phi)\}$.

Proof. Since q is surjective, the formula $\bar{\phi}(q(r)) = \phi(r)$ ($r \in R$) uniquely determines the values of $\bar{\phi}$, so that $\bar{\phi}$ is unique if it exists. But if $r - r' \in I$ then since $I \subseteq \ker(\phi)$ it follows that $0 = \phi(r - r') = \phi(r) - \phi(r')$ and hence ϕ is constant on the I-cosets, and therefore induces a map $\bar{\phi}(m + I) = \phi(m)$. The fact that $\bar{\phi}$ is a homomorphism then follows directly from the definition of the ring structure on the quotient R/I: For example, to see that $\bar{\phi}$ respects multiplication note that if $C_1, C_2 \in R/I$ then picking r_1, r_2 such that $C_1 = q(r_1), C_2 = q(r_2)$ we have

$$\begin{aligned} \bar{\phi}(C_1.C_2) &= \bar{\phi}(q(r_1)q(r_2)) = \bar{\phi}(q(r_1r_2)) = \phi(r_1r_2) = \phi(r_1).\phi(r_2) \\ &= \bar{\phi}(q(r_1))\bar{\phi}(q(r_2)) = \bar{\phi}(C_1).\bar{\phi}(C_2), \end{aligned}$$

where in the above equalities we just use the defining property $\bar{\phi} \circ q = \phi$ of $\bar{\phi}$ and the fact that q and ϕ are homomorphisms. To see what the kernel of $\bar{\phi}$ is, note that $\bar{\phi}(r + I) = \phi(r) = 0$ if and only if $r \in \ker(\phi)$, and hence $r + I \in \ker(\phi)/I$ as required. □

Remark 4.14. The name "universal property" is perhaps overly grand, but you should think of it as analogous to the fact that the ideal generated by a set is characterized as the smallest ideal containing that set: The quotient R/I is the *largest* quotient of R which sends all of I to 0, in the strong sense that if $\phi: R \to S$ is any surjective homomorphism such that $\phi(I) = 0$, then R/I surjects onto S (and thus is "at least as large" as S).

This theorem has important corollaries[22] which are collectively known as the "Isomorphism theorems" for rings.

Corollary 4.15. *We have the following isomorphisms:*

1. *(First isomorphism theorem.) If* $\phi: R \to S$ *is a homomorphism then* ϕ *induces an isomorphism* $\bar{\phi}: R/ker(\phi) \to im(\phi)$.
2. *(Second isomorphism theorem.) If R is a ring, A a subring of R, and I an ideal of R, then*
$$(A + I)/I \cong A/(A \cap I),$$
3. *(Third isomorphism theorem.) Suppose that* $I_1 \subseteq I_2$ *are ideals in R. Then we have*
$$(R/I_1)/(I_2/I_1) \cong R/I_2.$$

Proof. For the first isomorphism theorem, apply the universal property to $I = \ker(\phi)$. Since in this case $\ker(\bar{\phi}) = \ker(\phi)/\ker(\phi) = 0$ it follows $\bar{\phi}$ is injective and hence induces an isomorphism onto its image which from the equation $\bar{\phi} \circ q = \phi$ must be exactly $\mathrm{im}(\phi)$.

[22] Though as I said in lecture, they are somewhat over-rated – the crucial thing to understand is the quotient construction itself, and the universal property.

For the second isomorphism theorem, note first that if A is a subring and I is an ideal, it is easy to check[23] that $A + I$ is again a subring of R which contains I as an ideal. Let $q\colon R \to R/I$ be the quotient map. It restricts to a homomorphism p from A to R/I, whose image is clearly $(A + I)/I$, so by the first isomorphism theorem it is enough to check that the kernel of p is $A \cap I$. But this is clear: if $a \in A$ has $p(a) = 0$ then $a + I = 0 + I$ so that $a \in I$, and so $a \in A \cap I$. (Note this argument automatically shows that $A \cap I$ is an ideal of A since it is the kernel of the homomorphism p).

For the third isomorphism theorem, let $q_i\colon R \to R/I_j$ for $j = 1, 2$. By the universal property for q_2 we see that there is a homomorphism $\bar{q}_2\colon R/I_1 \to R/I_2$ induced by the map $q_2\colon R \to R/I_2$, with kernel $\ker(q_2)/I_1 = I_2/I_1$ and $\bar{q}_2 \circ q_1 = q_2$. Thus $\bar{q}_2$ is surjective (since q_2 is) and hence the result follows by the first isomorphism theorem. □

Example 4.16. Suppose that V is a k-vector space and $\alpha \in \mathrm{End}(V)$. Then we saw before that $\phi\colon \mathsf{k}[t] \to \mathrm{End}(V)$ given by $\phi(f) = f(\alpha)$. It is easy to see that this map is a homomorphism, and hence we see that $\mathrm{im}(\phi)$ is isomorphic to $\mathsf{k}[t]/I$ where $I = \ker(f)$ is a principal ideal. The monic polynomial generating I is the minimal polynomial of α as studied in Algebra I.

Another useful application of these results is a general version[24] of the "Chinese Remainder Theorem". To state it recall from Example 2.2 *ix*) the direct sums construction for rings: if R and S are rings, then $R \oplus S$ is defined to be the ring of ordered pairs (r, s) where $r \in R$, $s \in S$, with addition and multiplication done componentwise.

Theorem 4.17. *Let R be a ring, and I, J ideals of R such that $I + J = R$. Then*

$$R/I \cap J \cong R/I \oplus R/J.$$

Proof. We have quotient maps $q_1\colon R \to R/I$ and $q_2\colon R \to R/J$. Define $q\colon R \to R/I \oplus R/J$ by $q(r) = (q_1(r), q_2(r))$. By the first isomorphism theorem, it is enough to show that q is surjective and that $\ker(q) = I \cap J$. The latter is immediate: if $q(r) = 0$ then $q_1(r) = 0$ and $q_2(r) = 0$, whence $r \in I$ and $r \in J$, that is, $r \in I \cap J$. To see that q is surjective, suppose $(r + I, s + J) \in R/I \oplus R/J$. Then since $R = I + J$ we may write $r = i_1 + j_1$ and $s = i_2 + j_2$, where $i_1, i_2 \in I$, $j_1, j_2 \in J$. But then $r + I = j_1 + I$ and $s + J = i_2 + J$, so that $q(j_1 + i_2) = (r + I, s + J)$. □

Remark 4.18. Suppose that $R = I + J$ where I and J are ideals as above and moreover that $I \cap J = \{0\}$. Then each $r \in R$ can be written *uniquely* in the form $i + j$ where $i \in I$ and $j \in J$ (the proof is exactly the same as it is for subspaces in a vector space). In this situation we write[25] $R = I \oplus J$. Note that since $I.J \subseteq I \cap J$ it follows that

[23] See the problem set.

[24] The classical Chinese Remainder Theorem shows that if $m, n \in \mathbb{Z}$ are coprime then for any $a, b \in \mathbb{Z}$ there is a solution to the pair of equations $x = a \mod m$ and $x = b \mod n$, moreover this solution is unique modulo $m.n$. Check you see why the general version stated above implies this.

[25] The notation is compatible with the direct sum notation used in the first lecture – see the next paragraph.

$i.j = 0$ for any $i \in I, j \in I$, thus if $i_1, i_2 \in I$ and $j_1, j_2 \in J$ we see $(i_1 + j_1).(i_2 + j_2) = i_1 i_2 + j_1 j_2$. Writing $1 = e_1 + e_2$ where $e_1 \in I$ and $e_2 \in J$ if follows $(I, +, \times, 0, e_1)$ is a ring as is $(J, +, \times, 0, e_2)$, and it is easy to see that these rings are isomorphic to R/J and R/I respectively. This gives a more explicit description of the isomorphism $R \cong R/I \oplus R/J$ provided by the Chinese Remainder Theorem in this case.

Note also that if we start with two rings S_1, S_2, and define $R = S_1 \oplus S_2$ as in Example 2.2 *ix*), then the copies S_1^R, S_2^R of S_1 and S_2 inside R (that is, the elements $\{(s, 0) : s \in S_1\}$ and $\{(0, t) : t \in S_2\}$ respectively) are ideals in R (not subrings because they do not contain the multiplicative identity element $(1, 1)$) and clearly their intersection is $\{(0, 0)\}$, so that $R = S_1^R \oplus S_2^R$, thus the "external" notion of direct sum we saw in lecture 1 is compatible with the "internal" direct sum notation we used above (that is, when we write $R = I \oplus J$ to denote that I, J are ideals in R with $I + J = R$ and $I \cap J = \{0\}$).

When $R = \mathbb{Z}$ and $I = n\mathbb{Z} = \{nd : d \in \mathbb{Z}\}$, $J = m\mathbb{Z}$, then you can check that $I + J = \mathbb{Z}$ precisely when n and m are coprime, and then it also follows that $I \cap J = (n.m)\mathbb{Z}$ (the problem sheet asks you to work out the details of this), and so we recover the classical "Chinese Remainder Theorem": if m, n are coprime integers, then $\mathbb{Z}/(nm)\mathbb{Z} \cong (\mathbb{Z}/n\mathbb{Z}) \oplus (\mathbb{Z}/m\mathbb{Z})$. For example, if $R = \mathbb{Z}/6\mathbb{Z}$ then $R = \bar{3}R \oplus \bar{4}R$ (writing $\bar{n}$ for $n + 6\mathbb{Z}$ *etc.*) and this gives the identification $R = \mathbb{Z}/2\mathbb{Z} \oplus \mathbb{Z}/3\mathbb{Z}$.

4.2. **Images and preimages of ideals.** Next we want to compare ideals in a quotient ring with ideals in the original ring.

Lemma 4.19. *Let $\phi\colon R \to S$ be a surjective homomorphism of rings. If $I \lhd R$ then*

$$\phi(I) = \{s \in S : \exists i \in I, s = \phi(i)\}$$

is an ideal in S. Similarly if $J \lhd S$ then $\phi^{-1}(J) = \{r \in R : \phi(r) \in J\}$ is an ideal in R. Thus ϕ induces a pair of maps:

$$\begin{tikzcd} \{\text{ Ideals in } R\ \} \arrow[r, bend left, "\phi"] & \{\text{ Ideals in } S\ \} \arrow[l, bend left, "\phi^{-1}"] \end{tikzcd}$$

Proof. Let $I \lhd R$. Then $\phi(I)$ is certainly an additive subgroup of S since ϕ is a homomorphism of additive groups, and if $s \in S$ and $j = \phi(i) \in \phi(I)$, since ϕ is surjective, we may find $r \in R$ such that $\phi(r) = s$. It follows that $s.j = \phi(r).\phi(i) = \phi(r.i) \in \phi(I)$ since $r.i \in I$ because I is an ideal in R.

If $J \lhd S$ we can consider the subset $\phi^{-1}(J) = \{x \in R : \phi(x) \in J\}$. We claim this is an ideal of R. Indeed if $x, y \in \phi^{-1}(J)$ then $\phi(x + y) = \phi(x) + \phi(y) \in J$ since J is an additive subgroup, so that $\phi^{-1}(J)$ is an additive subgroup, and if $r \in R, x \in \phi^{-1}(J)$, then $\phi(r.x) = \phi(r).\phi(x) \in J$ since $\phi(x) \in J$ and J is an ideal in S. □

The next proposition shows that these maps can be used to identify the ideals of S with the subset of the ideals of R consisting of those ideals which contain the kernel of the homomorphism ϕ.

Proposition 4.20. *Let $\phi\colon R \to S$ be a surjective ring homomorphism and let $K = ker(\phi) \lhd R$. Then*

(1) *If $J \lhd S$ then we have $\phi(\phi^{-1}(J)) = J$;*
(2) *If $I \lhd R$ then we have $\phi^{-1}(\phi(I)) = I + K$.*

In particular the maps $J \mapsto \phi^{-1}(J)$ and $I \mapsto \phi(I)$ induce bijections between the set of ideals in S and the set of ideals in R which contain K:

$$\begin{tikzcd} \{ \textit{ Ideals in } R \textit{ containing } K \} \arrow[r, bend left, "\phi"] & \{ \textit{ Ideals in } S \} \arrow[l, bend left, "\phi^{-1}"] \end{tikzcd}$$

Proof. For the first part, note that if $f : X \to Y$ is any map of sets and $Z \subseteq Y$ then $f(f^{-1}(Z)) = Z \cap \mathrm{im}(f)$. Thus because ϕ is surjective we see that for any subset $J \subseteq S$ (and in particular for any ideal) $\phi(\phi^{-1}(J)) = J$.

For the second part, note that if $I \lhd R$ then $0 \in \phi(I)$, and so $K = \ker(\phi) = \phi^{-1}(0) \subseteq \phi^{-1}(\phi(I))$. Since $I \subseteq \phi^{-1}(\phi(I))$ also, it follows that $I+K$, the ideal generated by I and K, must lie in the ideal $\phi^{-1}(\phi(I))$. To see the reverse inclusion, note that if $x \in \phi^{-1}(\phi(I))$ then by definition there is some $i \in I$ with $\phi(x) = \phi(i)$, and hence $\phi(x - i) = 0$. But then $x = i + (x - i) \in I + K$, so that $\phi^{-1}(\phi(I)) \subseteq I + K$ as required.

Finally, to see the bijective correspondence, note we have already seen that for an ideal $J \lhd S$ we have $\phi(\phi^{-1}(J)) = J$, and since $K \subseteq \phi^{-1}(J)$ it follows that $J \mapsto \phi^{-1}(J)$ is an injective map whose image lands in the set of ideals of R which contain K. On the other hand, if $I \supseteq K$ is an ideal in R the $I + K = I$ and so $\phi^{-1}(\phi(I)) = I$, so that $I \mapsto \phi(I)$, when restricted to the set of ideals of R which contain K, is the inverse map to $J \mapsto \phi^{-1}(J)$ as required. □

In particular, if $I \lhd R$ and we take $\phi = q$ to be the canonical quotient homomorphism $q : R \to R/I$ we get the following:

Corollary 4.21. *Let R be a ring, I an ideal in R and $q : R \to R/I$ the quotient map. If J is an ideal then $q(J)$ is an ideal in R/I, and if K is an ideal in R/I then $q^{-1}(K) = \{r \in R : q(r) \in K\}$ is an ideal in R which contains I. Moreover, these correspondences give a bijection between the ideals in R/I and the ideals in R which contain I.*

5. Prime and maximal ideals, Euclidean domains and PIDs.

The quotient construction gives us a powerful way to build new rings and fields. The properties of the rings we obtain as quotients depend on the properties of the ideals we quotient by, and this leads us to the study of certain classes of ideals. In this section we begin studying two important such classes.

Definition 5.1. Let R be a ring, and I an ideal of R. We say that I is a maximal ideal if it is not strictly contained in any proper ideal of R. We say that I is a *prime* ideal if $I \neq R$ and for all $a, b \in R$, whenever $a.b \in I$ then either $a \in I$ or $b \in I$. If a prime I is principal any generator of I is said to be a *prime* element.

Lemma 5.2. *An ideal I in a ring R is prime if and only if R/I is an integral domain*[26]*. It is maximal if and only if R/I is a field. In particular, a maximal ideal is prime.*

[26] Note that this is "why" one wants to exclude R from being a prime ideal – I defined an integral domain to be a ring which was not the zero ring and had no zero divisors.

Proof. Suppose that $a, b \in R$. Note that $(a+I)(b+I) = 0+I$ if and only if $a.b \in I$. Thus if R/I is an integral domain, $(a+I)(b+I) = 0$ forces either $a+I = 0$ or $b+I$ is zero, that is, a or b lies in I, which shows I is prime. The converse is similar.

For the second part, note that a field is a ring which has no nontrivial ideals (*check this*!). The claim then follows immediately from the correspondence between ideals in the quotient ring and the original ring given in Lemma 4.21. Since fields are obviously integral domains, the "in particular" claim follows immediately. □

Remark 5.3. You can also give a direct proof that a maximal ideal is prime. Indeed if I is maximal and $a.b \in I$, and suppose that $b \notin I$. Then the ideal $J = I + bR$ generated by I and b is strictly larger than I, and so since I is maximal it must be all of R. But then $1 = i + br$ for some $i \in I$ and $r \in R$, and hence $a = a.1 = a.i + (a.b)r \in I$ since $i, a.b \in I$ as required.

Example 5.4. Let $R = \mathbb{Z}$. Since an ideal I in $\mathbb{Z}$ is in particular an subgroup of the abelian group $\mathbb{Z}$, we know it must be cyclic, that is $I = d\mathbb{Z}$ for some integer d. Thus every ideal in $\mathbb{Z}$ is principal. An ideal $d\mathbb{Z}$ is prime exactly when d is prime, and since in that case $\mathbb{Z}/d\mathbb{Z}$ is a field provided $d \neq 0$ it follows the maximal ideals are exactly the nonzero prime ideals.

We now consider a more substantial example, that of polynomials in one variable over a field. Although the case of field coefficients is the only one we really need for the moment, the following lemma captures, for polynomials with coefficients in a general ring, when you can do "long division with remainders" in polynomial rings. For this we first need to recall the notion of the degree of a nonzero polynomial:

Definition 5.5. If R is a ring and $f \in R[t]$ is nonzero, then we may write $f = \sum_{i=0}^{n} a_i t^i$, where $a_n \neq 0$. We set the *degree* $\deg(f)$ of f to be n, and say a_n is a the *leading coefficient* of f. If R is an integral domain, then for any $f, g \in R[t]$ you can check that $\deg(f.g) = \deg(f) + \deg(g)$ (and so in particular this implies $R[t]$ is also an integral domain).

Lemma 5.6. *(Division Algorithm). Let R be a ring and $f = \sum_{i=0}^{n} a_i t^i \in R[t]$, where $a_n \in R^\times$. Then if $g \in R[t]$ is any polynomial, there are unique polynomials $q, r \in R[t]$ such that either $r = 0$ or $\deg(r) < \deg(f)$ and $g = q.f + r$.*

Proof. This is straight-forward to prove by induction on $\deg(g)$. Since the $a_n \in R^\times$, if $h \in R[t]\backslash\{0\}$ it is easy to see[27] that $\deg(f.h) = \deg(f) + \deg(h)$. It follows that if $\deg(g) < \deg(f)$ we must take $q = 0$ and thus $r = g$. Now suppose that $g = \sum_{j=0}^{m} b_j t^j$ where $b_m \neq 0$ and $m = \deg(g) \geq n = \deg(f)$. Then since $a_n^{-1} b_m t^{m-n}.f$ has leading term $b_m t^m$ the polynomial

$$h = g - a_n^{-1} b_m t^{m-n}.f,$$

[27] The key here is that a unit is never a zero-divisor: if $a.b = 0$ and a is a unit, then $b = (a^{-1}.a).b = a^{-1}.(a.b) = a^{-1}.0 = 0$.

has $\deg(h) < \deg(g)$. It follows by induction that there are unique q', r' with $h = q'.f + r'$. Setting $q = a_n^{-1} b_n t^{m-n} + q'$ and $r = r'$ it follows $g = q.f + r$. Since q and r are clearly uniquely determined by q' and r' they are also unique as required. □

It follows from the previous lemma that if k is a field, then we have the division algorithm for all non-zero polynomials. This allows us to prove that all ideals in $\mathsf{k}[t]$ are principal.

Lemma 5.7. *Let I be a nonzero ideal in $\mathsf{k}[t]$. Then there is a unique monic polynomial f such that $I = \langle f \rangle$. In particular, all ideals in $\mathsf{k}[t]$ are principal.*

Proof. Since I is nonzero we may pick an $f \in I$ of minimal degree, and rescale it if necessary to make it monic. We claim $I = \langle f \rangle$. Indeed if $g \in I$, then using the division algorithm, we may write $g = q.f + r$ where either $r = 0$ or $\deg(r) < \deg(f)$. But then $r = g - q.f \in I$, and thus by the minimality of the degree of $f \in I$ we must have $r = 0$ and so $g = q.f$ as required. The uniqueness follows[28] from the fact that if $I = \langle f \rangle$ and $I = \langle f' \rangle$ then we would have $f = a.f'$ and $f' = b.f$, for some polynomials $a, b \in \mathsf{k}[t]$. But then $f = a.f' = (ab).f$ so that a and b must have degree zero, that is, $a, b \in \mathsf{k}$. Since we required f and f' to be monic, it follows that $a = b = 1$ and so $f = f'$ as required. □

The division algorithm also allows to give a reasonably explicit description of the rings we obtain quotient of a polynomial ring $\mathsf{k}[t]$: We have just seen that any nonzero ideal I is of the form $\langle f \rangle$ for a monic polynomial f. By the division algorithm, any polynomial g can be written *uniquely* as $g = q.f + r$ where $\deg(r) < \deg(f)$. Thus the polynomials of degree strictly less that $d = \deg(f)$ form a complete set of representatives for the I-cosets: every coset contains a unique representative r of degree strictly less than $\deg(f)$. Since $\{1, t, \dots, t^{\deg(f)-1}\}$ form a basis of the k-vector space of polynomials of degree less than $\deg(f)$ this means that if we let $q\colon \mathsf{k}[t] \to \mathsf{k}[t]/I$ be the quotient map, and $\alpha = q(t)$, then $\{1, \alpha, \dots, \alpha^{d-1}\}$ form a k-basis for $\mathsf{k}[t]/I$, and we multiply in $\mathsf{k}[t]/I$ using the rule $\alpha^d = -a_0 - a_1\alpha - \dots - a_{d-1}\alpha^d$, where $f(t) = t^d + \sum_{i=0}^{d-1} a_i t^i$. In particular, $\mathsf{k}[t]/\langle f \rangle$ is a k-vector space of dimension $\deg(f)$. We can therefore interpret the quotient construction $\mathsf{k}[t]/\langle f \rangle$ as a way of building a new ring out of k and an additional element α which satisfies the relation $f(\alpha) = 0$, or rather, the quotient construction gives us a rigorous way of doing this. The following example shows how one can use this to give a new construction of the complex numbers.

Example 5.8. When $\mathsf{k} = \mathbb{R}$, intuitively we build $\mathbb{C}$ out of $\mathbb{R}$ and an element "i" which satisfied $i^2 + 1 = 0$. The quotient construction lets us make this intuition rigorous: we simply define $\mathbb{C}$ to be the quotient ring $\mathbb{R}[t]/\langle t^2 + 1 \rangle$. Indeed this is a field because $t^2 + 1$ is irreducible[29] in $\mathbb{R}[t]$ (see Lemma 5.17 below for more on this) and if we let i

[28] This also follows from the fact that generators of a principal ideal are all associates, and the fact (which you proved in the first problem sheet) that the units in $\mathsf{k}[t]$ are exactly $\mathsf{k}^{\times}$.

[29] In general it is not so easy to decide if a polynomial $f \in \mathsf{k}[t]$ is irreducible, but in the case where $\deg(f) \leq 3$, f is reducible if and only if it has a root in k, which can (sometimes) be easy to check.

denote the image of t under the quotient map from $\mathbb{R}[t]$ to $\mathbb{C}$, then $\mathbb{C} = \mathbb{R}[t]/\langle t^2 + 1\rangle$ is a two-dimensional $\mathbb{R}$-vector space with basis $\{1, i\}$ and i satisfies $i^2 + 1 = 0$.

Remark 5.9. In fact with a little more care[30] it is straight-forward to check that if R is any ring and $f \in R[t]$ is a monic polynomial of degree d, and we let $Q = R[t]/\langle f\rangle$ and $\alpha = q(t)$ (where $q\colon R[t] \to R[t]/\langle f\rangle$ is the quotient map as before) then any element of Q can be written uniquely in the form $r_0 + r_1\alpha + \ldots + r_{d-1}\alpha^{d-1}$, where the multiplication in Q is given by the same rule as above. Of course for a general ring, not all ideals in $R[t]$ will necessarily be principal, and even if $I = \langle f\rangle$, if the leading coefficient of f is not a unit, we cannot apply the division algorithm.

Notice that the argument we used in the proof of Lemma 5.7 runs exactly the same way as the proof that every subgroup of $(\mathbb{Z}, +)$ is cyclic (or that any ideal in $\mathbb{Z}$ is principal). This suggests it might be useful to abstract the division algorithm for a general integral domain.

Definition 5.10. Let R be an integral domain and let $N\colon R\backslash\{0\} \to \mathbb{N}$ be a function. We say that R is a Euclidean domain if given any $a, b \in R$ with $b \neq 0$ there are $q, r \in R$ such that $a = b.q + r$ and either $r = 0$ or $N(r) < N(b)$.

Remark 5.11. Some texts require that the norm N satisfies additional properties, and in practice these additional properties are often very useful. For example sometimes the norm satisfies $N(a.b) = N(a).N(b)$ (in which case the norm is said to be *multiplicative*) or $N(a.b) = N(a) + N(b)$. The most general additional property one often asks for is that $N(a) \leq N(a.b)$ for all $a, b \in R\backslash\{0\}$. You can check that if R is a Euclidean domain satisfying this last property then the group of units $R^\times$ is precisely the set $\{a \in R : N(a) = N(1)\}$. However, if one just wants to know the ring is a PID the only condition one needs is the division algorithm.

Both $\mathbb{Z}$ and $\mathsf{k}[t]$, for any field k, are Euclidean domains with the norm given by the absolute value and the degree function respectively. We now show that the Gaussian integers $\mathbb{Z}[i]$ gives another example:

Lemma 5.12. *Let $R = \mathbb{Z}[i]$ and let $N\colon R \to \mathbb{N}$ be the function $N(z) = a^2 + b^2$, where $z = a + ib \in \mathbb{Z}[i], a, b \in \mathbb{Z}$. Then (R, N) is an Euclidean Domain.*

Proof. Note that N is the restriction of the square of the modulus function on $\mathbb{C}$, so in particular $N(z.w) = N(z).N(w)$. We write $|z|^2$ instead of $N(z)$ when $z \in \mathbb{C}\backslash\mathbb{Z}[i]$. Suppose that $s, t \in \mathbb{Z}[i]$ and $t \neq 0$. Then $s/t \in \mathbb{C}$, and writing $s/t = u + iv$ where $u, v \in \mathbb{Q}$ we can clearly take $a, b \in \mathbb{Z}$ such that $|u - a|, |v - b| \leq 1/2$ and so $q = a + ib$ we have $|s/t - q|^2 \leq \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$, and so $N(s - qt) \leq \frac{1}{2}N(t)$ (since $N(z_1z_2) = N(z_1).N(z_2)$) and hence if $r = s - qt \in \mathbb{Z}[i]$ we see that either $r = 0$ or $N(r) \leq \frac{1}{2}N(t) < N(t)$ as required. Note that r is not necessarily unique in this case. □

Lemma 5.13. *Let (R, N) be an Euclidean domain. Then any ideal in R is principal.*

[30] In particular, one needs to use the general statement of the division algorithm as given in Lemma 5.6.

Proof. The proof that any ideal is principal is exactly the same as for $\mathsf{k}[t]$: If I is a nonzero ideal, take $d \in I$ such that $N(d)$ is minimal. Then if $m \in I$ we may write $m = q.d + r$, where $r = 0$ or $N(r) < N(d)$. But $r = m - q.d \in I$ so that the minimality of $N(d)$ forces $r = 0$ and so $m = q.d$. It follows that $I \subseteq R.d$, and since $d \in I$ clearly $Rd \subseteq I$, hence $I = Rd$ as required. □

Definition 5.14. An integral domain in which every ideal is principal, that is, generated by a single element, is called a *Principal Ideal Domain*. This is usually abbreviated to PID. The previous Lemma shows that any Euclidean Domain is a PID.

Remark 5.15. It is also possible to consider rings in which every ideal is principal but which are not necessarily integral domains[31]. Such rings are called Principal Ideal Rings. As we mostly focus on integral domains, we will not however use this term in this course.

We would like to calculate which ideals in a Euclidean domain are prime and which are maximal. In fact we can give an answer for any PID not just any Euclidean domain.

Definition 5.16. Let R be an integral domain. A nonzero element $r \in R$ is said to be irreducible if whenever $r = a.b$ then exactly one of a or b is a unit (so that in particular r is not a unit). We will say an element $R \in R\backslash(\{0\} \cup R^{\times})$ is *reducible* if it is not irreducible[32].

Lemma 5.17. *Let R be a PID, and let $d \in R\backslash\{0\}$. Then the following are equivalent:*

(1) *$R.d = \langle d \rangle$ is a prime ideal.*
(2) *d is irreducible in R.*
(3) *$R.d$ is a maximal ideal in R.*

Proof. 1) implies 2): If $d = a.b$ then as $d \in R.d$ is prime we must have $a \in R.d$ or $b \in R.d$. By symmetry we may assume $a \in R.d$ (and hence, since $R.d$ is a proper ideal and $R.a \subseteq R.d$ we see that a is not a unit[33]). But then there is some $r \in R$ with $a = r.d$, and so $d = a.b = (r.b).d$ and hence $(1 - r.b).d = 0$ and so since R is an integral domain and $d \neq 0$ we must have $r.b = 1$, that is $b \in R^{\times}$.

2) implies 3): Suppose that d is irreducible, and that $R.d \subseteq I \lhd R$. Since R is a PID, we must have $I = R.a$ for some $a \in R$, and $R.d \subseteq R.a$ shows that $d = a.b$ for some $b \in R$. But then as d is irreducible we must have one of a or b a unit. But if a is a unit, then $R.a = R$, while if b is a unit d and a are associates and so generate the same ideal, that is $R.d = I$. It follows $R.d$ is a maximal ideal as claimed.

3) implies 1): We have already seen that in any ring a maximal ideal must be prime. □

[31] As an exercise, you might try to find an example of such a ring.

[32] On the one hand, since units have no prime factors, it seems reasonable to consider them not to be reducible, but on the other hand, we do not want them to be irreducible: we will show any nonzero element of a PID is a product of irreducibles in an essentially unique way, and this uniqueness would not make sense if we allow units to be irreducible.

[33] Check you see that an element r of a ring R is a unit if and only if $R.r = R$.

Remark 5.18. Note that the implication "1) implies 2)" holds in any integral domain, while "3) implies 1)" holds in any commutative ring. In a general ring $d \in R$ irreducible is equivalent to the ideal $R.d$ being maximal *amongst principal ideals* in R.

It is also worth pointing out that the Lemma reduces the problem classifying prime and maximal ideals in a PID R to the problem of finding irreducible elements in R. When R is say $\mathbb{C}[t]$, this is easy: by the fundamental theorem of algebra a monic polynomial $p \in \mathbb{C}[t]$ is irreducible if and only if $p = t - \lambda$ for some $\lambda \in \mathbb{C}$. On the other hand if $R = \mathbb{Q}[t]$ then it is in general very difficult to decide if a polynomial $p \in \mathbb{Q}[t]$ is irreducible. For the ring $R = \mathbb{Z}[i]$ it is possible to give a fairly complete description of the irreducibles: see the problem sheet.

6. An introduction to fields.

In the previous section we saw how to construct $\mathbb{C}$ from $\mathbb{R}$ as the quotient $\mathbb{C} \cong \mathbb{R}[t]/\langle t^2 + 1\rangle$. This example generalises substantially, and in this section we use the quotientswe have developed to construct some examples of fields, and develop a little of their basic properties. Our main tool is Lemma 5.17, which shows that if $f \in \mathsf{k}[t]$ is any irreducible polynomial then $\mathsf{k}[t]/\langle f\rangle$ is a field, and moreover by the above discussion it is clearly a k-vector space of dimension $\deg(f)$.

Example 6.1. Suppose that E be a finite field (*i.e.* a field with finitely many elements). Then E has characteristic p for some prime $p \in \mathbb{N}$ (since otherwise E contains a copy of $\mathbb{Z}$ and is hence infinite). Thus E contains the subfield $\mathbb{F}_p \cong \mathbb{Z}/p\mathbb{Z}$. In particular we can view it as an $\mathbb{F}_p$-vector space, and since it is finite, it must certainly be finite-dimensional. But then if $d = \dim_{\mathbb{F}_p}(E)$, clearly there are p^d elements in E. Thus we see that a finite field must have prime-power order.

Let's see an explicit example: Take for example $p = 3$. Then it is easy to check that $t^2 + 1$ is irreducible in $\mathbb{F}_3[t]$ (you just need to check it does not have a root in $\mathbb{F}_3$, and there are only 3 possibilities!). But then by our discussion above $E = \mathbb{F}_3[t]/\langle t^2 + 1\rangle$ is field of dimension 2 over $\mathbb{F}_3$, and hence E is a finite field with 9 elements.

More generally, if we can find an irreducible polynomial f of degree d in $\mathbb{F}_p[t]$ the quotient $\mathbb{F}_p[t]/\langle f\rangle$ will be a finite field of order p^d. In the Problem sheets we will show that for each d there is an irreducible polynomial of degree d in $\mathbb{F}_p[t]$, hence showing that finite fields of any prime-power order exist. In fact there is only one field (up to isomorphism) of any fixed prime-power order, but we will not prove that in this course.

Definition 6.2. If E, F are any fields and $F \subseteq E$ we call E a *field extension* of F and write E/F. The inclusion of F into E gives E the structure of an F-vector space. If E is finite dimensional as an F-vector space, we write $[E : F] = \dim_F(E)$ for this dimension and call it the *degree* of the field extension E/F.

Although it probably seems a very crude notion, as it forgets alot of the structure of E, the degree of a field extension is nevertheless very useful. One reason for this is the following Lemma:

Lemma 6.3. *Let E/F be a field extension and let $d = [E : F] < \infty$. Then if V is an E-vector space, we may view V as an F-vector space, and V is finite dimensional as an F-vector space if and only if it is as an E vector space, and moreover* $\dim_F(V) = [E : F]\dim_E(V)$.

Proof. Certainly if V is an E-vector space then by restricting the scalar multiplication map to the subfield F it follows that V is an F-vector space. Moreover, if V is finite dimensional as an F-vector space it is so as an E-vector space (a finite F-spanning set will certainly be a finite E-spanning set). Conversely, suppose that V is a finite dimensional E-vector space. Let $\{x_1, x_2, \ldots, x_d\}$ be an F-basis of E, and let $\{e_1, \ldots, e_n\}$ be an E-basis of V. To finish the proof it is enough to check that $\{x_ie_j : 1 \leq i \leq d, 1 \leq j \leq n\}$ is an F-basis of V: Indeed if $v \in V$, then since $\{e_1, \ldots, e_n\}$ is an E-basis of V there are $\lambda_i \in E$ $(1 \leq i \leq n)$ such that $v = \sum_{i=1}^n \lambda_ie_i$. Moreover, since $\{x_1, \ldots, x_d\}$ is an F-basis of E then for each λ_i there are elements μ^i_j $(1 \leq j \leq d)$ such that $\lambda_i = \sum_{j=1}^d \mu^i_jx_j$. Thus we have

$$v = \sum_{i=1}^n \lambda_ie_i = \sum_{i=1}^n (\sum_{j=1}^d \mu^i_jx_j)e_i = \sum_{1\leq i\leq n, 1\leq j\leq d} \mu^i_j(x_je_i),$$

whence the set $\{x_je_i : 1 \leq i \leq n, 1 \leq j \leq d\}$ spans V as an F-vector space (and in particular we have already established that V is finite dimensional as an F-vector space). To see that this set is linearly independent, and hence establish the dimension formula, just notice that in the above equation, $v = 0$ if and only if each $\lambda_i = 0$ by the linear independence of the vectors $\{e_1, \ldots, e_n\}$, and $\lambda_i = 0$ if and only if each $\mu^j_i = 0$ for $1 \leq j \leq d$ by the linear independence of the x_js. □

Example 6.4. Let V be a $\mathbb{C}$ vector space with basis $\{e_1, \ldots, e_n\}$. Then since $\{1, i\}$ is an $\mathbb{R}$-basis of $\mathbb{C}$, it follows $\{e_1, \ldots, e_n, ie_1, \ldots, ie_n\}$ is an $\mathbb{R}$-basis of V.

We record a particularly useful case of the above Lemma:

Corollary 6.5. *(Tower Law) Let $F \subset E \subset K$ be fields, then $[K : F]$ is finite if and ony if both degrees $[E : F], [K : E]$ are, and when they are finite we have $[K : F] = [E : F][K : E]$.*

Proof. Apply the previous Lemma to the E-vector space K. □

We now use these tools to study finite extensions of $\mathbb{Q}$ inside the field of complex numbers. The problem sheets also study finite fields, that is, finite extensions of $\mathbb{Z}/p\mathbb{Z}$.

Definition 6.6. Let $\alpha \in \mathbb{C}$. We say that α is *algebraic* over $\mathbb{Q}$ if there is a field E which is a finite extension of $\mathbb{Q}$ containing α. Otherwise we say that α is *transcendental*. Notice that since the intersection of subfields is again a subfield[34], given any set $T \subseteq \mathbb{C}$ there is always a smallest subfield which contains it. This is called the field *generated by T*, and is denoted $\mathbb{Q}(T)$ (recall that any subfield of $\mathbb{C}$ contains $\mathbb{Q}$, since it contains $\mathbb{Z}$ and hence $\mathbb{Q}$ because it is the field of fractions of $\mathbb{Z}$). In the case where X has just a single element α we write $\mathbb{Q}(\alpha)$ rather than $\mathbb{Q}(\{\alpha\})$ and we say the field

[34] Just as for subspace of vector space, subrings of a ring, ideals in a ring *etc.*

extension is *simple*. Note that an element $\alpha \in \mathbb{C}$ is algebraic if and only if $\mathbb{Q}(\alpha)$ is a finite extension of $\mathbb{Q}$. Slightly more generally, if F is any subfield of $\mathbb{C}$ and $\alpha \in \mathbb{C}$ we let $F(\alpha) = \mathbb{Q}(F \cup \{\alpha\})$ be the smallest subfield of $\mathbb{C}$ containing both F and α, and one says α is algebraic over F if $F(\alpha)/F$ is a finite extension.

The next Lemma shows that simple extensions are exactly the kind of fields our quotient construction builds.

Lemma 6.7. *Suppose that E/F is a finite extension of fields (both say subfields of $\mathbb{C}$) and let $\alpha \in E$. Then there is a unique monic irreducible polynomial $f \in F[t]$ such that the evaluation homomorphism $\phi\colon F[t] \to E$ given by sending t to α induces an isomorphism $F(\alpha) \cong F[t]/\langle f \rangle$.*

Proof. The field $K = F(\alpha)$ is a finite extension of F since it is a subfield of E (and hence a sub-F-vector space of the finite dimensional F-vector space E). Let $d = [K : F] = \dim_F(K)$. Since the set $\{1, \alpha, \alpha^2, \dots, \alpha^d\}$ has $d + 1$ elements, it must therefore be linearly dependent, and so that there exist $\lambda_i \in F$ $(0 \leq i \leq d)$, not all zero, such that $\sum_{i=0}^{d} \lambda_i \alpha^i = 0$. But then if $g = \sum_{i=0}^{d} \lambda_i t^i \in F[t]\backslash\{0\}$, we see that $g(\alpha) = 0$. It follows that the kernel I of the homomorphism $\phi\colon F[t] \to E$ given by $\phi(\sum_{j=0}^{m} c_j t^j) = \sum_{j=0}^{m} c_j \alpha^j$ is nonzero. Now any nonzero ideal in $F[t]$ is generated by a unique monic polynomial, thus we have $I = \langle f \rangle$, where f is monic and f is uniquely determined by ϕ (and so by α). By the first isomorphism theorem, the image S of ϕ is isomorphic to $F[t]/I$. Now S is a subring of a field, so certainly an integral domain, hence $\langle f \rangle$ must be a prime ideal, and by our description of prime ideals in $F[t]$ it must therefore in fact be maximal, so that S is actually a field. Finally, any subfield of $\mathbb{C}$ containing F and α must clearly contain S (as the elements of S are F-linear combinations of powers of α) so it follows $S = F(\alpha)$. □

Definition 6.8. Given $\alpha \in \mathbb{C}$, the polynomial f associated to α by the previous Lemma, that is, the irreducible polynomial for which $\mathbb{Q}(\alpha) \cong \mathbb{Q}[t]/\langle f \rangle$, is called the *minimal polynomial* of α over $\mathbb{Q}$. Note that our description of the quotient $\mathbb{Q}[t]/\langle f \rangle$ shows that $[\mathbb{Q}(\alpha) : \mathbb{Q}] = \deg(f)$, hence the degree of the simple field extension $\mathbb{Q}(\alpha)$ is just the degree of the minimal polynomial of α.

Remark 6.9. (*Non-examinable*) For simplicity let's suppose that all our fields are subfields of $\mathbb{C}$. It is in fact the case that any finite extension E/F is simple, that is $E = F(\alpha)$ for some $\alpha \in E$ (this is known as the *primitive element theorem*, which is proved in next year's Galois theory course). Moreover it turns out that given any finite extension E/F of a field F there are in fact only finitely many fields K between E and F. Neither statement is obvious, but you should think about how the two facts are clearly closely related: if you accept the statement about finitely many subfields between E and F then it is not hard to believe the primitive element theorem – you should just pick an element of E which does not lie in any proper subfield, and to see such an element exists one just has to show that the union of finitely many proper subfields of E cannot be the whole field E. On the other hand, if E/F is a finite field extension and we know that $E = F(\alpha)$ for some $\alpha \in E$, then we have $E \cong F[t]/\langle f \rangle$ where $f \in F[t]$ is the minimal polynomial of α

over F. If K is a field with $F \subseteq K \subseteq E$, then certainly $E = K(\alpha)$ also, and it follows $E \cong K[t]/\langle g \rangle$, where $g \in K[t]$ is irreducible. But now you can check (using the tower law) that if $g = \sum_{i=0}^{k} c_i t^i \in K[t]$, then the c_is actually generate K over F, that is $K = F(\{c_i : 0 \leq i \leq k\})$, thus the the possible subfields of $F(\alpha)$ are all determined already by the roots of f (as the c_is are just polynomial functions of the roots).

Example 6.10. (1) Consider $\sqrt{3} \in \mathbb{C}$. There is a unique ring homomorphism $\phi\colon \mathbb{Q}[t] \to \mathbb{C}$ such that $\phi(t) = \sqrt{3}$. Clearly the ideal $\langle t^2 - 3 \rangle$ lies in $\ker(\phi)$, and since $t^2 - 3$ is irreducible in $\mathbb{Q}[t]$ so that $\langle t^2 - 3 \rangle$ is a maximal ideal, we see that $\ker \phi = \langle t^2 - 3 \rangle$, and hence $\text{im}(\phi) \cong \mathbb{Q}[t]/\langle t^2 - 3 \rangle$. Now the quotient $\mathbb{Q}[t]/\langle t^2 - 3 \rangle$ is field, hence $\text{im}(\phi)$ is also. Moreover, any subfield of $\mathbb{C}$ which contains $\sqrt{3}$ clearly contains $\text{im}(\phi)$, so we see that $\text{im}\phi = \mathbb{Q}(\sqrt{3})$. In particular, since the images of $\{1, t\}$ form a basis of the quotient $\mathbb{Q}[t]/\langle t^2 - 3 \rangle$ by our description of quotients of polynomial rings in the previous section, and under the isomorphism induced by ϕ these map to 1 and $\sqrt{3}$ respectively, we see that $\mathbb{Q}(\sqrt{3}) = \{a + b\sqrt{3} : a, b \in \mathbb{Q}\}$, a degree two extension of $\mathbb{Q}$. (Note that one can also just directly check that the right-hand side of this equality is a field – I didn't do that because I wanted to point out the existence of the isomorphism with $\mathbb{Q}[t]/(t^2 - 3)$.)

(2) Exactly the same strategy[35] shows that $\mathbb{Q}(2^{1/3})$ is isomorphic to $\mathbb{Q}[t]/\langle t^3 - 2 \rangle$, and hence $\mathbb{Q}(2^{1/3})$ is a 3-dimensional $\mathbb{Q}$-vectors space with basis $\{1, 2^{1/3}, 2^{2/3}\}$, again given by the image of the standard basis we defined in the quotient $\mathbb{Q}[t]/\langle t^3 - 2 \rangle$. Note that while its relatively easy to check directly that $\{a + b\sqrt{3} : a, b \in \mathbb{Q}\}$ is a subfield of $\mathbb{C}$, it's already noticeably harder to see directly that $\{a + b2^{1/3} + c2^{2/3} : a, b, c \in \mathbb{Q}\}$ is a subfield of $\mathbb{C}$: one needs to show that for any $a, b, c \in \mathbb{Q}$ not all zero, the reciprocal $(a + b2^{1/3} + c2^{2/3})^{-1}$ can be written as a $\mathbb{Q}$-linear combination of $\{1, 2^{1/3}, 2^{2/3}\}$.

Example 6.11. Now let $T = \{\sqrt{3}, 2^{1/3}\}$. Let us figure out what $E = \mathbb{Q}(T)$ looks like. Certainly it contains the subfields $E_1 = \mathbb{Q}(\sqrt{3})$ and $E_2 = \mathbb{Q}(2^{1/3})$. Using the tower law and the above examples, we see that $[E : \mathbb{Q}] = [E : E_1].[E_1 : \mathbb{Q}] = 2[E : E_1]$, and similarly $[E : \mathbb{Q}] = 3[E : E_2]$. It follows that $6 = l.c.m.\{2, 3\}$ divides $[E : \mathbb{Q}]$. On the other hand, consider E/E_2. If $\sqrt{3} \in E_2$, then clearly $E = E_2$, which would mean $[E : \mathbb{Q}] = 3$, which is not divisible by 6, so that we must have $\sqrt{3} \notin E_2$. But then arguing exactly as we did above, there is a unique homomorphism $E_2[t] \to \mathbb{C}$ sending t to $\sqrt{2}$, with kernel $\langle t^2 - 3 \rangle$, a maximal ideal since $\sqrt{3} \notin E_2$, so that $\text{im}(\phi)$ must be the field generated by $\sqrt{2}$ and $2^{1/3}$, and thus $[E : E_2] = \dim_{E_2}(E_2[t]/\langle t^2 - 3 \rangle) = 2$, and so $[E : \mathbb{Q}] = [E : E_2][E_2 : \mathbb{Q}] = 2.3 = 6$. Moreover, using the proof of the tower law, you can check that the arguments above even show that E has a $\mathbb{Q}$-basis given by $\{2^{a/3}3^{b/2} : 0 \leq a \leq 2, 0 \leq b \leq 1\}$.

[35] We just need to check that $t^3 - 2 \in \mathbb{Q}[t]$ is irreducible, but this follows because it does not have a root in $\mathbb{Q}$.

7. Unique factorisation.

Throughout this section unless otherwise explicitly stated all rings are integral domains.

For the integers $\mathbb{Z}$, any integer can be written as a product of prime numbers in an essentially unique way. (See the Appendix for a direct proof of this, which may be useful to review before reading this section.) We will show in this section that this property holds for any Principal Ideal Domain.

Definition 7.1. Let R be an integral domain. If $a, b \in R$ we say that *a divides b*, or a is a *factor* of b, and write $a|b$, if there is some $c \in R$ such that $b = a.c$. Note that we can also write this in terms of the ideals a and b generate: in fact $a|b$ if and only if $bR \subseteq aR$, as you can see immediately from the definitions.

It also makes sense to talk about least common multiples and highest common factors in any integral domain:

Definition 7.2. Let R be an integral domain. We say $c \in R$ is a *common factor* of $a, b \in R$ if $c|a$ and $c|b$, and that c is the *highest common factor*, and write $c = h.c.f.(a, b)$, if whenever d is a common factor of a and b we have $d|c$. In the same way, we can define the least common multiple of $a, b \in R$: a common multiple is an element $k \in R$ such that $a|k$ and $b|k$, and the *least common multiple* is a common multiple which is a factor of every common multiple.

Note that these definitions can be rephrased in terms of principal ideals: c is a common factor of a, b if and only if $\{a, b\} \subseteq cR$. An element g is the highest common factor of $\{a, b\}$ if and only if gR is minimal among principal ideals containing $\{a, b\}$, that is, if $\{a, b\} \subseteq cR$ then $gR \subseteq cR$. Similarly the l is the least common multiple of $\{a, b\}$ if it lR is maximal among principal ideals which lie in $aR \cap bR$.

Lemma 7.3. *If $a, b \in R$ where R is an integral domain, then if a highest common factor h.c.f$\{a, b\}$ exists, it is unique up to units. Similarly when it exists, the least common multiple is also unique up to units. Moreover if R is a PID then the highest common factor and least common multiple alway exist.*

Proof. This is immediate from our description of the highest common factor in terms of ideals. Indeed if g_1, g_2 are two highest common factors, then we must have $g_1R \subseteq g_2R$ (since g_1 is a highest common factor and g_2 is a common factor) and symmetrically $g_2R \subseteq g_1R$. But then $g_1R = g_2R$, and so since R is an integral domain this implies g_1, g_2 are associates, *i.e.* they differ by a unit. The proof for least common multiples is analogous.

If R is a PID then the ideal $\langle a, b\rangle$ is principal, and so is clearly the minimal principal idea containing a, b, and so any generator of it is a highest common factor. Similarly $Ra \cap Rb$ is principal and any generator of it will be a least common multiple. □

Recall that an nonzero element c in an integral domain R is *irreducible* if whenever $c = a.b$ exactly one of a or b is a unit[36], and that a nonzero element $p \in R$ is prime if

[36] We also say an element a of an integral domain R is *reducible* if it is non-zero, not a unit, and not irreducible, *i.e.* if we can write it as a product $a = b.c$ where $b, c \in R\backslash\{0\}$ and neither b or c is a unit.

the ideal Rp is prime. More explicitly $p \in R$ is *prime* if whenever $p|a.b$ either $p|a$ or $p|b$ (or possibly both). Note that it follows by induction on m that if p is prime and $p|a_1.a_2 \dots a_m$ then $p|a_j$ for some j $(1 \leq j \leq m)$. Moreover, by Lemma 5.17 and the remark immediately following it, if R is an integral domain then prime elements are always irreducible, and for elements of a PID the two notions are equivalent.

We now want to study factorisation in an integral domain, and in particular the question of when one can uniquely factor elements into a product of irreducibles. We formalise this with the following definition.

Definition 7.4. An integral domain R is said to be an *unique factorisation domain* (or UFD) if every element of $R\backslash\{0\}$ is either a unit, or can be written as a product of irreducible elements, and moreover the factorization into irreducibles is unique up to reordering and units. More explicitly, if R is a UFD and $r \in R$ is nonzero and not a unit, then there are irreducible elements $p_1, \dots, p_k$ such that $r = p_1 p_2 \dots p_k$ and whenever $r = q_1 q_2 \dots q_l$ is another such factorization for r, then $k = l$ and the q_js can be reordered so that $q_j = u_j p_j$, where $u_j \in R$ is a unit. If, as is normal, we interpret an empty product in a ring to be 1, then we can rephrase this too include the units in the assertion so that any nonzero element can be expressed as a product of irreducibles uniquely up to order and units.

Lemma 7.5. *Suppose that R is an integral domain. Then the following are equivalent:*

(1) *R is a UFD.*
(2) *Both of the following hold:*
  i) Every irreducible element is prime,
  ii) Every nonzero non-unit $a \in R$ can be written as a product of irreducibles.
(3) *Every nonzero non-unit $a \in R$ can be written as a product of prime elements.*

Proof. We first show R is a UFD if and only if *i*) and *ii*) of (2) holds: Suppose that R is a UFD and p is an irreducible. If p divides $a.b$, where $a, b \in R$, then if either a or b is zero or a unit we are done. Otherwise by assumption they can be written as a product irreducibles, say $a = q_1 \dots q_k$ and $b = r_1 \dots r_l$ for some $k, l \geq 1$. But we have $a.b = p.d$ by definition, and writing $d = s_1 \dots s_m$ as a product of irreducibles, by uniqueness of the factorization of $a.b$ into irreducibles we see that that up to units p must be one of the q_is or r_js, and hence p divides a or b as required.

For the converse, we use induction on the minimal number $M(a)$ of irreducibles (or equivalently, primes) in a factorization of a into irreducibles. If $M(a) = 1$ then a irreducible and uniqueness is clear by the definition of an irreducible element[37]. Now suppose that $M = M(a) > 1$ and $a = p_1 p_2 \dots p_M = q_1 q_2 \dots q_k$ for irreducibles p_i, q_j and $k \geq M$. Now it follows that $p_1 | q_1 \dots q_k$, and so since p_1 is prime there is some q_j with $p_1 | q_j$. Since q_j is irreducible, this implies that $q_j = u_1 p_1$ for some unit $u_1 \in R$. Reordering the q_ls if needed we can assume $j = 1$, and so we see that $(u_1^{-1} p_2) \dots p_M = q_2 q_2 \dots q_k$, and by induction it follows that $k - 1 = M - 1$, *i.e.* $k = M$, and moreover the irreducibles occuring are equal up to reordering and units as required.

[37] Or if you prefer, including units, we can start with $M(a) = 0$, so that a is already a unit.

To see that condition (2) is equivalent to condition (3), note that since prime elements are always irreducible, we need only check that irreducibles are prime. But if $a \in R$ is irreducible and a is a product of primes, say $a = p_1 p_2 \dots p_k$, then by the definition of irreducibility we must have $k = 1$ and hence a is prime as required. □

We are now going to show that unique factorisation holds in any PID. By the above, since we already know that irreducibles are prime in a PID, it is enough to show that any element has *some* factorization into irreducibles. At first sight this seems like it should be completely obvious: if an element $a \in R$ is irreducible, then we're done, otherwise it has a factorisation $a = b.c.$ where b, c are proper factors (that is, $b|a$ and $c|a$ and neither are associates of a). If either of b or c is not irreducible then we can find a proper factorisation of them and keep going until we reach a factorisation of a into irreducibles. The trouble with this argument is that we need to show the process we describe stops after finitely many steps. Again intuitively this seems clear, because the proper factors of a should be "getting smaller", but again *a priori* they might just keep getting "smaller and smaller". The key to showing that this cannot happen is to rephrase things in terms of ideals: Recall that $b|a$ if and only if $aR \subseteq bR$ and b is a proper factor of a (*i.e.* b divides a and is not an associate of a) if and only if $aR \subsetneq bR$, that is, aR is strictly contained in bR. Thus if R, our PID, contained an element which could be factored into smaller and smaller factors this would translate this into a nested sequence of ideals each of which strictly contained the previous ideal. The next Proposition shows that this cannot happen in a PID.

Proposition 7.6. *Let R be a PID and suppose that $\{I_n : n \in \mathbb{N}\}$ is a sequence of ideals such that $I_n \subseteq I_{n+1}$. Then the union $I = \bigcup_{n\geq 0} I_n$ is an ideal and there exists an $N \in \mathbb{N}$ such that $I_n = I_N = I$ for all $n \geq N$.*

Proof. Let $I = \bigcup_{n\geq 1} I_n$. Given any two elements $p, q \in I$, we may find $k, l \in \mathbb{N}$ such that $p \in I_k$ and $q \in I_l$. It follows that for any $r \in R$ we have $r.p \in I_k \subset I$, and taking $n = \max\{k, l\}$ we see that $r, s \in I_n$ so that $r + s \in I_n \subset I$. It follows that I is an ideal. Since R is a PID, we have $I = \langle c \rangle$ for some $c \in R$. But then there must be some N such that $c \in I_N$, and hence $I = \langle c \rangle \subseteq I_N \subseteq I$, so that $I = I_N = I_n$ for all $n \geq N$ as required. □

Remark 7.7. A ring which satisfies the condition that any nested ascending chain of ideals stabilizes is called a *Noetherian* ring. The condition is a very important "finiteness" condition in ring theory. (Note that the proof that the chain of ideals stabilizes generalises readily if you just know every ideal is generated by finitely many elements, rather than a single element.) Polynomial rings in any number of indeterminates have this property by a theorem know as Hilbert's Basis Theorem, which you can learn more about in the Commutative Algebra course in Part B.

Theorem 7.8. *Let R be a PID. Then R is a UFD.*

Proof. As discussed above, it follows from the fact that irreducibles are prime in a PID and Lemma 7.5 that we need only show any element can be factored as a product of irreducible elements. Thus suppose for the sake of a contradiction that there is some $a = a_1 \in R$ which is not a product of irreducible elements. Clearly a cannot be irreducible, so we may write it as $a = b.c$ where neither b nor c is a unit. If both b and c can be written as a product of prime elements, then multiplying these expressions together we see that a is also, hence at least one of b or c cannot be written as a product of prime elements. Pick one, and denote it a_2. Note that if we set $I_k = \langle a_k \rangle$ (for $k = 1, 2$) then $I_1 \subsetneq I_2$. As before a_2 cannot be irreducible, so we may find an a_3 such that $I_2 = \langle a_2 \rangle \subsetneq \langle a_3 \rangle = I_3$. Continuing in this fashion we get a nested sequence of ideals I_k each strictly bigger than the previous one. But by Proposition 7.6 this cannot happen if R is a PID, thus no such a exists.

□

Remark 7.9. (*Non-examinable*). The annoying "up to units" qualification for prime factorisation in a PID vanishes if you are willing to live with ideals rather than elements: in a PID any proper ideal I can be written as a product of nonzero prime ideals $I = P_1P_2 \ldots P_k$ where the prime ideals occuring in this factorisation are unique up to reordering. Indeed this is just the statement that two elements of an integral domain are associates if and only if they generate the same principal ideal. However, if you do Algebraic Number Theory next year you'll see this idea extended to rings where unique factorization of elements fails (in particular the rings are not PIDs!) but where nevertheless unique factorization of ideals continues to hold.

Remark 7.10. (*Again non-examinable, but perhaps illuminating.*) In special cases the proof that any element is a product of irreducibles can be simplified: more precisely, suppose that R is an Euclidean domain with a norm N which satisfies the condition that $N(a) \leq N(a.b)$ for all $a, b \in R\backslash\{0\}$. We will call[38] such a norm *weakly multiplicative*. (This holds for example if the norm satisfies something like $N(a.b) = N(a).N(b)$ or $N(a.b) = N(a) + N(b)$.) In this case we can replace the use of Proposition 7.6 with a more concrete inductive argument. In order to make the induction work however, we will need to know that when we factorise an element as a product of two proper factors (*i.e.* so neither factor is a unit) then the norms of the factors are strictly smaller than the norm of the element. Of course if have an explicit description of the norm (as we do say for $\mathsf{k}[t]$ or $\mathbb{Z}$) this may be easy to check directly, but it is in fact a consequence of the weakly multiplicative property. More precisely we have:
Claim: Let R be an ED with a weakly multiplicative norm. If $a, b \in R\backslash\{0\}$ satisfy $b|a$ and $N(a) = N(b)$ then a and b are associates.
Proof: To prove the claim, suppose that $N(a) = N(b)$ and $a = b.c$. We must show that c is a unit. By the division algorithm we have $b = q.a + r$ where $r = 0$ or $N(r) < N(a) = N(b)$. Substituting $a = b.c$ and rearranging we get $b(1 - q.c) = r$, and

[38] I don't know if there is a standard name for this property – "multiplicative" would suggest something like $N(a.b) = N(a).N(b)$. "Submultiplicative" might be another reasonable term, but it sounds pretty awful.

hence if $r \neq 0$ then $N(r) = N(b.(1-q.c)) \geq N(b) = N(a)$ which is a contradiction. Thus $r = 0$ and so since $b \neq 0$, $1 - q.c = 0$ and so c is a unit as required.

We now show how, in any Euclidean Domain R with a weakly multiplicative norm a nonunit $a \in R\backslash\{0\}$ is a product of irreducibles using induction on $N(a)$ the norm. Note that $N(1) \leq N(1.a) = N(a)$ for all $a \in R\backslash\{0\}$, so that the minimum value of N is $N(1)$. But by what we have just done, if $N(a) = N(1)$ then a is a unit (since 1 divides any $a \in R$). If $N(a) > N(1)$ then either a is an irreducible element, in which case we are done, or $a = b.c$, where neither b nor c is a unit. But then by the claim we must have $N(b), N(c) < N(a)$, and hence by induction they can be expressed as a product of irreducibles and so multiplying these expressions together we see so can a. It follows every $a \in R\backslash\{0\}$ is unit or a product of irreducibles as required.

A ring may be a UFD without being a PID: in fact we will now show that $\mathbb{Z}[t]$ is a UFD, even though it is not a PID. The idea is to use the fact that, since $\mathbb{Z}$ and $\mathbb{Q}[t]$ are PIDs, unique factorisation holds in each. We can then show $\mathbb{Z}[t]$ is a UFD by studying the inclusion of $\mathbb{Z}[t]$ into $\mathbb{Q}[t]$. The next definition and Lemma are the key to our understanding of factorisation in $\mathbb{Z}[t]$.

Definition 7.11. If $f \in \mathbb{Z}[t]$ then define the *content* $c(f)$ of f to be the highest common factor of the coeficients of f. That is, if $f = \sum_{i=0}^{n} a_i t^i$ then we set $c(f) = \text{h.c.f.}\{a_0, a_1, \ldots, a_n\}$. Note that in a general integral domain the highest common factor is only defined up to units, but in the case of $\mathbb{Z}$ if we insist $c(f) > 0$ then it is unique (since the units in $\mathbb{Z}$ are just $\{\pm 1\}$). In particular, given $f \in \mathbb{Z}[t]$ nonzero, $c(f)$ is the unique positive integer such that $f = c(f).f_1$ where f_1 has content 1, that is, its coefficients generate the whole ring $\mathbb{Z}$.

Lemma 7.12. *(Gauss). Let $f, g \in \mathbb{Z}[t]$. Then $c(f.g) = c(f).c(g)$.*

Proof. Suppose first $f, g \in \mathbb{Z}[t]$ have $c(f) = c(g) = 1$. Then let $p \in \mathbb{N}$ be a prime. We have for each such prime a homomorphism $\mathbb{Z}[t] \to \mathbb{F}_p[t]$ given by $\phi_p(\sum_{i=0}^{n} a_i t^i) = \sum_{i=0}^{n} \bar{a}_i t^i$, where $\bar{a}_i$ denotes $a_i + p\mathbb{Z} \in \mathbb{F}_p$. It is immediate that $\ker(\phi_p) = p\mathbb{Z}[t]$, so that we see $p|c(f)$ if and only if $\phi_p(f) = 0$. But since $\mathbb{F}_p$ is a field, $\mathbb{F}_p[t]$ is an integral domain, and so as ϕ_p is a homomorphism we see that

$$\begin{aligned} p|c(f.g) &\iff \phi_p(f.g) = 0 \iff \phi_p(f).\phi_p(g) = 0 \\ &\iff \phi_p(f) = 0 \text{ or } \phi_p(g) = 0 \iff p|c(f) \text{ or } p|c(g), \end{aligned}$$

whence it is clear that $c(f.g) = 1$ if $c(f) = c(g) = 1$.

Now let $f, g \in \mathbb{Z}[t]$, and write $f = a.f'$,$g = b.g'$ where $f', g' in \mathbb{Z}[t]$ have $c(f') = c(g') = 1$, (so that $c(f) = a, c(g) = b$). Then clearly $f.g = (a.b).(f'g')$ and since $c(f'g') = 1$ it follows that $c(f.g) = c(f).c(g)$ as required.

□

Alternative proof. If you found the above proof of the fact that $c(f.g) = 1$ if $c(f) = c(g) = 1$ a bit too slick, then a more explicit version of essentially the same argument goes as follows: Let $f = \sum_{i=0}^{n} a_i t^i$ and[39] $g = \sum_{i=0}^{n} b_i t^i$, and write $f.g = \sum_{k=0}^{2n} c_k t^k$.

[39] Note that so long as we do not assume that both b_a and a_n are nonzero we may take the same upper limit in the sums.

Suppose that d divides all the coefficients of $f.g$ and d is not a unit. Since $c(f) = 1$, there must be a smallest k such that d does not divide a_k and similarly since $c(g) = 1$ there is a smallest l such that d does not divide b_l. Consider

$$c_{k+l} = \sum_{i+j=k+l} a_i b_j,$$

Now d divides every term on the right-hand side except for $a_k b_l$, since every other term has one of $i < k$ or $j < l$, but then d does not divide the sum, contradicting the assumption that d divides c_{k+l}. Thus we have a contradiction and thus $c(f.g) = 1$ as required.

We can now extend the definition of content to arbitrary nonzero elements of $\mathbb{Q}[t]$.

Lemma 7.13. *Suppose $f \in \mathbb{Q}[t]$ is nonzero. Then there is an unique $\alpha \in \mathbb{Q}_{>0}$ such that $f = \alpha f'$ where $f' \in \mathbb{Z}[t]$ and $c(f') = 1$. We write $c(f) = \alpha$. Moreover, if $f, g \in \mathbb{Q}[t]$ then $c(f.g) = c(f).c(g)$.*

Proof. Let $f = \sum_{i=0}^{n} a_i t^i$ where $a_i = b_i/c_i$ for $b_i, c_i \in \mathbb{Z}$ and $h.c.f\{b_i, c_i\} = 1$ for all i, $0 \leq i \leq n$. Pick $d \in \mathbb{Z}_{>0}$ such that $da_i \in \mathbb{Z}$ for all i, $(1 \leq i \leq n)$ so that $df \in \mathbb{Z}[t]$ (for example you can take $d = \text{l.c.m.}\{c_i : 0 \leq i \leq n\}$ or $\prod_{i=0}^{n} c_i$). Set $c(f) = c(d.f)/d$ (where the righthand side is already defined because $d.f \in \mathbb{Z}[t]$). Then $f = c(f).f'$ where $f' = (d.f)/c(d.f)$ is clearly a polynomial in $\mathbb{Z}[t]$ with content one. To check $c(f)$ is well-defined we must show that if $f = \alpha_1.f_1 = \alpha_2.f_2$ where f_1, f_2 have content one and $\alpha_1, \alpha_2 \in \mathbb{Q}_{>0}$ then $\alpha_1 = \alpha_2$. But writing $\alpha_i = m_i/n_i$ for positive integers m_i, n_i, we find $n_2.(m_1.f_1) = n_1.(m_2 f_2) \in \mathbb{Z}[t]$. Taking content and using the trivial fact that if $n \in \mathbb{Z}\backslash\{0\} \subset \mathbb{Z}[t]$ then $c(d) = |d|$, we see that

$$c(n_2.(m_1 f_1)) = c(n_2).c(m_1 f_1) = n_2.m_1, \qquad c(n_1.(m_2 f_2)) = c(n_1).c(m_2 f_2) = n_1 m_2.$$

Equating it follows $\alpha_1 = m_1/n_1 = m_2/n_2 = \alpha_2$ as required.

It is now easy to check multiplicativity: if $f, g \in \mathbb{Q}[t]/\backslash\{0\}$ then if $d_1, d_2 \in \mathbb{Z}$ are such that $d_1.f, d_2 g \in \mathbb{Z}[t]$ then clearly $d_1 d_2.(f.g) \in \mathbb{Z}[t]$ so that

$$c(f.g) = \frac{c(d_1 d_2 f.g)}{d_1 d_2} = \frac{c((d_1 f).(d_2 g))}{d_1.d_2} = \frac{c(d_1 f)}{d_1}.\frac{c(d_2.g)}{d_2} = c(f).c(g),$$

as required.

□

Remark 7.14. Note in particular it follows immediately from the previous Lemma that if $f \in \mathbb{Q}[t]\backslash\{0\}$ then $f \in \mathbb{Z}[t]$ if and only if $c(f) \in \mathbb{Z}$.

We now relate factorization in $\mathbb{Z}[t]$ and $\mathbb{Q}[t]$, and obtain a description of some prime elements in $\mathbb{Z}[t]$.

Lemma 7.15.
(1) *Suppose that $f \in \mathbb{Z}[t] \subset \mathbb{Q}[t]$ is nonzero, and that $f = g.h$ where $g, h \in \mathbb{Q}[t]$. Then there exist $\alpha \in \mathbb{Q}$ such that $(\alpha.g), (\alpha^{-1}.h) \in \mathbb{Z}[t]$. Thus $f = (\alpha.g)(\alpha^{-1}h)$ is a factorisation of f in $\mathbb{Z}[t]$.*
(2) *Suppose that $f \in \mathbb{Q}[t]$ is irreducible and $c(f) = 1$. Then f is a prime element of $\mathbb{Z}[t]$.*

(3) *Let $p \in \mathbb{Z}$ be a prime number. Then p is a prime element in $\mathbb{Z}[t]$.*

Proof. For the first part, by Lemma 7.13 we may write $g = c(h).g_1$ and $h = c(h).h_1$ where $g_1, h_1 \in \mathbb{Z}[t]$ have content 1. Then $c(f) = c(g)c(h)$ so that as $f \in \mathbb{Z}[t]$ we have $c(g).c(h) \in \mathbb{Z}$. Setting $\alpha = c(h)$ we see that $f = (\alpha.g).(\alpha^{-1}.h)$ where $\alpha.g = (c(g).c(h)).g_1$ and $\alpha^{-1}h = h_1$ both lie in $\mathbb{Z}[t]$ as required.

For the second part, first note that if $f \in \mathbb{Q}[t]$ has $c(f) = 1$ then by definition f must lie in $\mathbb{Z}[t]$ (and has content 1). To see that such an f is prime, we need to show that if $g, h \in \mathbb{Z}[t]$ and $f|g.h$ in $\mathbb{Z}[t]$ then $f|g$ or $f|h$ in $\mathbb{Z}[t]$. Now if $f|g.h$ in $\mathbb{Z}[t]$, certainly it does so in $\mathbb{Q}[t]$. Since $\mathbb{Q}[t]$ is a PID, irreducibles are prime and so either $f|g$ or $f|h$ in $\mathbb{Q}[t]$. Suppose that $f|g$ (the argument being identical for h). Then we have $g = f.k$ for some $k \in \mathbb{Q}[t]$. Now by Lemma 7.13 we may write $k = c(k).k'$ where $k' \in \mathbb{Z}[t]$, and moreover by the same lemma, $c(g) = c(f).c(k) = c(k)$ since $c(f) = 1$. But $g \in \mathbb{Z}[t]$, hence $c(g) = c(k) \in \mathbb{Z}$ and hence $k \in \mathbb{Z}[t]$ so that f divides g in $\mathbb{Z}[t]$ as required.

For the final part, we have already seen that the homomorphism $\phi_p \colon \mathbb{Z}[t] \to \mathbb{F}_p[t]$ has kernel $p\mathbb{Z}[t]$, and so since $\mathbb{F}_p[t]$ is an integral domain, the ideal $p\mathbb{Z}[t]$ is prime, that is, p is a prime element of $\mathbb{Z}[t]$.

□

Theorem 7.16. *The ring $\mathbb{Z}[t]$ is a UFD.*

Proof. Since $\mathbb{Z}[t]$ is an integral domain (as $\mathbb{Z}$ is), by Lemma 7.5 it is enough to show that any element of $\mathbb{Z}[t]$ is a product of primes. Let $f \in \mathbb{Z}[t]$. We may write $f = a.f'$ where $c(f') = 1$, and since $\mathbb{Z}$ is a UFD we may factorise a into a product of prime elements of $\mathbb{Z}$ which we have just seen are prime in $\mathbb{Z}[t]$. Thus we may assume $c(f) = 1$. But then viewing f as an element of $\mathbb{Q}[t]$ we can write it as a product of prime elements in $\mathbb{Q}[t]$, say $f = p_1 p_2 \dots p_k$. Now using Lemma 7.13, each p_i can be written as $a_i q_i$ where $a_i \in \mathbb{Q}$ and $q_i \in \mathbb{Z}[t]$ and $c(q_i) = 1$. But then by the Lemma 7.15, q_i is prime in $\mathbb{Z}[t]$, and $f = (a_1 \dots a_k)q_1 \dots q_k$. Comparing contents we see that $(a_1 \dots a_k) = 1$ and so we are done. □

Remark 7.17. It is easy to see from this that in fact *all* primes in $\mathbb{Z}[t]$ are either primes in $\mathbb{Z}$ or primes (equivalently irreducibles) in $\mathbb{Q}[t]$ which have content 1.

Remark 7.18. In fact one can show directly (see the problem set) that if R is a UFD then highest common factors exist (that is, given elements $a_1, \dots, a_n \in R$ there is an element d such that $d|a_i$ for all i, $(1 \leq i \leq n)$ and if $c|a_i$ for all i also, then $c|d$). It follows that if R is a UFD then we can define the content of a nonzero element of $R[t]$ to be the highest common factor of its coefficients just as we did for $\mathbb{Z}[t]$. This observation implies the following theorem, whose proof is not examinable.

Theorem 7.19. *If R is a UFD then the polynomial ring $R[t]$ is also a UFD. More generally, if R is a UFD then $R[t_1, \dots, t_n]$ the ring of polynomials in n variables with coefficients in R, is a UFD.*

Proof. (*Nonexaminable.*) The proof follows the same strategy as for $\mathbb{Z}[t]$: if $f \in Rt]$ then as saw in the previous remark, the content of f makes sense (though now we cannot use positivity to make it unique, so it is only defined up to units). Let F be the field of fractions of R. Then $F[t]$ is a PID and hence a UFD, and the content lets us understand the relation of factorization in $R[t]$ to factorization in $F[t]$ using the analogue of Gauss's Lemma. You can then check that the primes in $R[t]$ are then the primes in R and the irreducibles in $F[t]$ with content 1, and the fact that any element of $R[t]$ has a prime factorization then follows exactly as for $\mathbb{Z}[t]$.

For the final part, since $R[t_1, \dots, t_n] = S[t_n]$ where $S = R[t_1, \dots, t_{n-1}]$ the result follows from the first part and induction on n. □

The previous theorem shows that, for example, $\mathbb{Q}[x, y]$ is a UFD. It is not hard to see that neither $\mathbb{Z}[t]$ nor $\mathbb{Q}[x, y]$ are PIDs[40], so the class of rings which are UFDs is strictly larger than the class of PIDs. In fact not every PID is a Euclidean domain either, so there are strict containments: EDs $\subsetneq$ PIDs $\subsetneq$ UFDs. Finding a PID which is not a Euclidean domain is a bit subtle, so we wont do it here, but see the Appendix.

7.1. **Irreducible polynomials.** In this section we develop some techniques for deciding when a polynomial $f \in \mathbb{Q}[t]$ is irreducible. By what we have done above, if $f \in \mathbb{Q}[t]$ is irreducible, we may write $f = c(f).g$ where $g \in \mathbb{Z}[t]$ has content 1 and is a prime in $\mathbb{Z}[t]$. Since f and g are associates in $\mathbb{Q}[t]$ it follows that to understand irreducible elements in $\mathbb{Q}[t]$ it is enough to understand the prime elements in $\mathbb{Z}[t]$ of positive degree (or equivalently, the irreducibles $f \in \mathbb{Q}[t]$ with $c(f) = 1$.)

This is useful for the following reason: Recall that for any prime $p \in \mathbb{Z}$ we have the homomorphism[41] $\phi_p \colon \mathbb{Z}[t] \to \mathbb{F}_p[t]$. This allows us to transport questions about factorisation in $\mathbb{Z}[t]$ to questions about factorisation in $\mathbb{F}_p[t]$: If $f \in \mathbb{Z}[t]$ with $c(f) = 1$, and $f = g.h$, then g and h have content 1. In particular, if $f = g.h$ is a factorization of f with neither of g, h a unit, then both g and h have positive degree. If we pick p a prime not dividing the leading coefficient of f, then $\deg(\phi_p(f)) = \deg(f)$ and so[42] we must have $\deg(\phi_p(g)) = \deg(g)$ and $\deg(\phi_p(h)) = \deg(h)$, and hence we obtain a proper factorization of $\phi_p(f)$. It follows that if $\phi_p(f)$ is irreducible in $\mathbb{F}_p[t]$, then f must be irreducible in $\mathbb{Z}[t]$. Since the rings $\mathbb{F}_p[t]$ are "smaller" than either $\mathbb{Z}[t]$ or $\mathbb{Q}[t]$ this can give us ways of testing irreducibility (indeed notice since the coefficient field is finite, it is in principle a finite check to see if a given element in $\mathbb{F}_p[t]$ is irreducible).

Example 7.20. Suppose that $f = t^3 - 349t + 19 \in \mathbb{Z}[t]$. If f is reducible in $\mathbb{Q}[t]$, it is reducible in $\mathbb{Z}[t]$ and hence its image under ϕ_p in $\mathbb{F}_p[t]$ will be reducible. But since f has degree 3 it follows it is reducible if and only if it has a degree 1 factor, and similarly for its image in $\mathbb{F}_p[t]$, which would therefore mean it has a root in $\mathbb{F}_p$. But taking $p = 2$ we see that $\phi_2(f) = \bar{f} = t^3 + t + 1 \in \mathbb{F}_2[t]$ and so it is easy to check that $\bar{f}(0) = \bar{f}(1) = 1 \in \mathbb{F}_2$, so $\bar{f}$ does not have a root, and hence f must be irreducible. Note on the other hand t^2+1 is irreducible in $\mathbb{Z}[t]$ but in $\mathbb{F}_2[t]$ we have $t^2+1 = (t+1)^2$, so $\phi_p(f)$ can be reducible even when f is irreducible.

[40] In fact for $\mathbb{Z}[t]$ this follows from Lemma 7.15 – do you see why?

[41] Note that there is *no* homomorphism from $\mathbb{Q}[t]$ to $\mathbb{F}_p[t]$ for any prime p. This is why we have to pass through $\mathbb{Z}[t]$.

[42] Because the degree of a product of polynomials is the sum of the degrees whenever the coefficient ring is an integral domain.

Lemma 7.21. *(Eisenstein's criterion.) Suppose that $f \in \mathbb{Z}[t]$ has $c(f) = 1$, and $f = a_n t^n + a_{n-1}t^{n-1} + \ldots a_1 t + a_0$. Then if there is a prime $p \in \mathbb{Z}$ such that $p|a_i$ for all i, $0 \leq i \leq n-1$ but p does not divide a_n and p^2 does not divide a_0 then f is irreducible in $\mathbb{Z}[t]$ and $\mathbb{Q}[t]$.*

Proof. Since $c(f) = 1$, we have already seen that irreducibility in $\mathbb{Z}[t]$ and $\mathbb{Q}[t]$ are equivalent. Let $\phi_p \colon \mathbb{Z}[t] \to \mathbb{F}_p[t]$ be the quotient map. Suppose that $f = g.h$ was a factorisation of f in $\mathbb{Z}[t]$ where say $0 < \deg(g) = k < n$. Then we have $\phi_p(f) = \phi_p(g).\phi_p(h)$. By assumption $\phi_p(f) = \bar{a}_n t^n$ (where for $m \in \mathbb{Z}$ we write $\bar{m}$ for $m + p\mathbb{Z}$, the image of m in $\mathbb{F}_p$). Since $\mathbb{F}_p[t]$ is a UFD, t is irreducible, and $\bar{a}_n$ is a unit, it follows that up to units we must have $\phi_p(g) = t^k, \phi_p(h) = t^{n-k}$. But then (since k and $n-k$ are both positive) the constant terms of both g and h must be divisible by p, and hence a_0 must be divisible by p^2, contradicting our assumption. □

Example 7.22. This gives an easy way to see that $2^{1/3} \notin \mathbb{Q}$: if it was $t^3 - 2$ would be reducible, but we see this is not the case by applying Eisenstein's criterion with $p = 2$. (It also gives a proof that $\sqrt{2}$, $\sqrt{3}$ are irrational).

One can also use Eisenstein's Criterion in more cunning ways. For example, it might be that the Criterion does not apply to $f(t)$ but it does to $f(t+1)$, as the next example shows:

Example 7.23. Suppose that $p \in \mathbb{N}$ is prime, and $f = 1 + t + \ldots + t^{p-1} \in \mathbb{Z}[t]$. Then we claim f is irreducible. Let $g = f(t+1)$. Then if g was reducible, say $g = h_1.h_2$ it would follow that $f(t) = g(t-1) = h_1(t-1)h_2(t-1)$ is reducible, and similarly if g is irreducible so is f. Thus f is irreducible if and only if g is. But as $f = \frac{t^p-1}{t-1}$ we see that

$$g = t^{-1}((t+1)^p - 1) = \sum_{i=0}^{p-1} \binom{p}{i+1} t^i,$$

But it is well know that p divides $\binom{p}{i+1}$ for any i, $0 \leq i \leq p-2$, while the constant term $\binom{p}{1} = p$ is not divisible by p^2, so Eisenstein's Criterion shows g and hence f is irreducible.

Remark 7.24. (*Non-examinable.*) You might be worried[43] about what "substituting $t+1$ for t" means for polynomials with coefficient in an arbitrary ring where we cannot think of them as functions. (In the case of $\mathbb{Z}[t]$ this is not a problem, since $\mathbb{Z}[t]$ embeds into $\mathbb{Q}[t]$ which can be viewed as a subring of the ring of functions from $\mathbb{Q}$ to itself since $\mathbb{Q}$ is infinite.) In fact we've done enough to make sense of this already: Recall that in Lemma 2.9 we showed that for a polynomial ring $R[t]$, if we are given a homomorphism $\phi \colon R \to S$ and an element of $s \in S$, then there is a unique homomorphism from $\psi \colon R[t] \to S$ taking t to s and $r \in R$ to $\phi(r)$. It follows that there is a unique homomorphism $\psi \colon R[t] \to R[t]$ which is the identity[44] on R

[43] You might also not be worried, I don't know which group is better off in life in general.

[44] In the case $R = \mathbb{Z}$, the identity is the only ring homomorphism from $\mathbb{Z}$ to itself so in that case you don't need to explicitly require this.

and sends t to $t + 1$. Since the homomorphism given by sending $t \mapsto t - 1$ is clearly an inverse to ψ we see that ψ is an isomorphism from $\mathbb{Z}[t]$ to itself. It follows that f is irreducible if and only if $\psi(f)$ is.

In fact the above trick could be generalized a little: the important point was that, as well as the homomorphisms $\phi_p \colon \mathbb{Z}[t] \to \mathbb{F}_p[t]$, we found an isomorphism from $\mathbb{Z}[t]$ to itself given by sending $t \mapsto t + 1$. If $\psi \colon \mathbb{Z}[t] \to \mathbb{Z}[t]$ was *any* isomorphism the we could similarly test the irreducibility of an element using the compositions $\phi_p \circ \psi$. However, this turns out to be not much of a generalization, as you can see if you find all isomorphisms of $\mathbb{Z}[t]$ with itself.

8. Modules: Definition and examples.

As usual, all rings are assumed to be commutative unless the contrary is explicitly stated.

In this section we begin the study of "linear algebra over rings". Recall a vector space is just an abelian group with an action of a field of "scalars" obeying some standard rules. The definition of a module is exactly the same, except now we allow our scalars to belong to an arbitrary ring, rather than insisting they belong to a field. Formally, we say the following:

Definition 8.1. Let R be a ring with identity 1_R. A *module* over R is an abelian group $(M, +)$ together with a multiplication action $a \colon R \times M \to M$ of R on M written $(r, m) \mapsto r.m$ which satisfies:

(1) $1_R.m = m$, for all $m \in M$;
(2) $(r_1.r_2).m = r_1.(r_2.m)$, for all $r_1, r_2 \in R$, $m \in M$
(3) $(r_1 + r_2).m = r_1.m + r_2.m$ for all $r_1, r_2 \in R$ and $m \in M$;
(4) $r.(m_1 + m_2) = r.m_1 + r.m_2$ for all $r \in R$ and $m_1, m_2 \in M$.

Remark 8.2. Just as with vector spaces, we write the addition in the abelian group M and the addition in the ring R as the same symbol "+", and similarly the multiplication action of R on M is written in the same way as the multiplication in the ring R, since the axioms ensure that there is no ambiguity in doing so.

Remark 8.3. Note that the definition makes perfectly good sense for a noncommutative ring (when it would normally be described as a *left module* since the action of the ring is on the left). Next year's course on Representation Theory will study certain noncommutative rings called group algebras, and modules over them. In this course we will focus on modules over integral domains and all our main results will be for modules over a PID, though even then, in some cases we will only give proofs for the case where our ring is a Euclidean domain.

Lemma 8.4. *Let M be an abelian group and R a ring.*

i) The set $End(M) = Hom(M, M)$ of group homomorphisms from M to itself is naturally a (in general noncommutative) ring where addition is give pointwise and multiplication is given by composition.
ii) Giving M the structure of an R-module, that is an action $R \times M \to M$, is equivalent to giving a ring homomorphism $\phi \colon R \to End(M)$.

Proof. For the first part, since M is abelian, if $f_1, f_2 \in \mathrm{End}(M)$ then $f_1 + f_2$ is again a group homomorphism, and clearly $f_1 + f_2 = f_2 + f_1$ so addition in $\mathrm{End}(M)$ is commutative. Composition of functions gives $\mathrm{End}(M)$ a multiplication which it is easy to check distributes over addition, thus $\mathrm{End}(M)$ is a (not necessarily commutative) ring.

For the second part, if the action map $a \colon R \times M$ is denoted by $(r, m) \mapsto r.m$ as usual, then the equation $\phi(r)(m) = r.m$ defines ϕ in terms of a and conversely. It is routine to check the equivalences – property (4) of an action shows that $m \mapsto r.m$ is a homomorphism of the abelian group M, property (1) corresponds to requiring $\phi(1_R) = 1_{\mathrm{End}(M)}$, property (2) to the compatibility of ϕ with multiplication, and property (3) to the compatibility of ϕ with addition. □

Remark 8.5. You should compare this Lemma with the corresponding result from group actions: if G is a group and X is a set, then giving an action of G on X is the same as giving a group homomorphism from G to the group of permutations of the set X (*i.e.* the group of bijections from X to itself). You could define vector spaces this way, but in Prelims we tell you what a vector space is before we tell you what a group action is (or indeed what a ring is!)

Example 8.6. Let's give a few examples:

(1) As mentioned above, if R is a field, the definition is exactly that of a vector space over R, so modules over a field are just vector spaces over that field.

(2) At the other end of the spectrum in a sense, if A is an abelian group, then it has a natural structure of $\mathbb{Z}$-module: if n is a positive integer, then set $n.a = a + a + \ldots + a$ (n times) and if n is a negative integer, set $n.a = -(a + a + \ldots + a)$ (where this time we add a to itself $-n$ times). It's easy to check this makes A a $\mathbb{Z}$-module, and moreover, the conditions $(1), (2), (3), (4)$ in fact force this definition on us, so that this $\mathbb{Z}$-module structure is unique[45]. Thus we see that $\mathbb{Z}$-modules are just abelian groups.

(3) Suppose that R is a ring. Then R is a module over itself in the obvious way.

(4) If R is a ring and I is an ideal in R, then it follows directly from the definitions that I is an R-module.

(5) Again if I is an ideal in R then R/I is naturally an R-module where the multiplication action is given via the quotient homomorphism $q \colon R \to R/I$, that is, if $m \in R/I$ and $r \in R$ we set $r.m = q(r).m$ (the multiplication on the right-hand side being inside the ring R/I). Indeed the properties $(1), (2)$ (3) and (4) all follow immediately from the fact that q is a ring homomorphism.

(6) Generalising the previous example somewhat, if $\phi \colon R \to S$ is a homomorphism of rings, and M is an S-module, then we can give M the structure of an R-module by setting $r.m = \phi(r).m$ (where the action on the right-hand side comes from the S-module structure. Thus for example any if I is an ideal of

[45] Writing down all the details of a proof of this is very similar to the exercise in the problem sheets in which you showed that given any ring R there is a unique homomorphism from $\mathbb{Z}$ to R. The reason for this is because the module structure corresponds to the unique ring homomorphism $\mathbb{Z} \to \mathrm{End}(M)$.

R then any R/I-module automatically has the structure of an R-module via the quotient map $q\colon R \to R/I$.

(7) Generalising the example of R being an R-module over itself in a slightly different way, given our ring R and a positive integer n, we may consider the module $R^n = \{(r_1, r_2, \ldots, r_n) : r_i \in R\}$ of n-tuples of elements of R (written as row vectors or column vectors – different books prefer different conventions), where the addition and the multiplication by scalars is done componentwise. (This is exactly the way we define the vector space $\mathbb{R}^n$ for the field $\mathbb{R}$). Such a module is an example of a *free module* over R.

(8) To give a more substantial example, suppose that V is a vector space over a field k and $\phi\colon V \to V$ is a linear map. Then we can make V into a $\mathsf{k}[t]$-module by setting $p(t).v = p(\phi)(v)$ for any $v \in V$ and $p(t) \in \mathsf{k}[t]$ (that is just evaluate the polynomial p on the linear map ϕ). Indeed a homomorphism from $\mathsf{k}[t]$ to $\mathrm{End}_{\mathsf{k}}(V)$ is uniquely determined by its restriction to the scalars k and the image of t. Here we define ϕ by the conditions that it sends the complex number $\lambda \in \mathsf{k} \subseteq \mathsf{k}[t]$ to $\lambda.\mathrm{id}_V$, and t to ϕ. The fact that the assignment $f.v = \phi(f)(v)$ for $v \in V, f \in \mathsf{k}$ makes V into a $\mathsf{k}[t]$-module follows directly from the fact that ϕ is a homomorphism. Conversely, if we are given a $\mathsf{k}[t]$-module M, we can view it as a k-vector space where the multiplication by scalars is given to us by viewing the elements of k as degree zero polynomials. The action of multiplication by t is then a k-linear map from M to itself. Thus $\mathsf{k}[t]$-modules are just k-vector spaces equipped with an endomorphism.

8.1. **Submodules, generation and linear independence.**

Definition 8.7. If M is an R-module, a subset $N \subseteq M$ is called a *submodule*[46] if it is an abelian subgroup of M and whenever $r \in R$ and $n \in N$ then $r.n \in N$.

If $\{N_i : i \in I\}$ is a collection of submodules then their intersection $\bigcap_{i\in I} N_i$ is also a submodule. This allows us to define (just as we did for ideals, subrings, subfields *etc.*) for a set $X \subset M$ the submodule generated by X,

Definition 8.8. If X is any subset of an R-module M then the submodule *generated* or *spanned* by X is defined to be:

$$\langle X\rangle = \bigcap_{N \supseteq X} N,$$

where N runs over the submodules of M which contain X. Explicitly, it is the subset $R.X = \{\sum_{i=1}^{k} r_i x_i : r_i \in R, x_i \in X\}$ (where this is by convention understood to be $\{0\}$ if $X = \emptyset$). The proof is exactly the same[47] as the proof for ideals in a ring.

If N_1, N_2 are submodules then the submodule they generate is their sum $N_1 + N_2 = \{m + n : m \in N_1, n \in N_2\}$. To prove this one first checks that the righthand side is indeed a submodule and then that any submodule containing N_1 and N_2 must

[46] Note the definitions in this subsection are exactly the same as for the case of a vector space.

[47] As should come as no surprise given example (3) above: a subset of R viewed as an R-module is a submodule if and only if it is an ideal.

contain all the elements of $N_1 + N_2$ (and these two steps both follow directly from the definitions). Note that this generalises the fact which we have already seen that the ideal generated by the union of two ideals $I \cup J$ is just their sum $I + J$.

Definition 8.9. If M is a module over R, we say a set $S \subseteq M$ is *linearly independent* if whenever we have an equation $r_1 s_1 + r_2 s_2 + \ldots r_k s_k = 0$ for $r_i \in R, s_i \in S$ $(1 \leq i \leq k)$ we have $r_1 = r_2 = \ldots r_k = 0$. We say that a set S is a *basis* for a module M if and only if it is linearly independent and it spans M. Any module which has a basis is called a *free* module. Finally we say that a module is *finitely generated* if it is generated by some finite subset.

9. Quotient modules and the isomorphism theorems.

Just as for vector spaces, given a module together with a submodule there is a natural notion of a quotient module. (If you've understood quotients of rings and quotients of vectors space, everything here should look very familiar, as the constructions mimics those cases, in fact they are word for word the same as for quotient vector spaces).

Definition 9.1. If N is a submodule of M, then in particular it is a subgroup of an abelian group, so we can form the quotient M/N. The condition that N is a submodule then is precisely what is needed for the multiplication on M to induce a module structure on M/N: If $r \in R$ and $m + N \in M/N$ then define $r.(m + N) = r.m + N$. This is well defined because if $m_1 + N = m_2 + N$ we have $m_1 - m_2 \in N$, and so $r.(m_1 - m_2) \in N$, whence $r.m_1 + N = r.m_2 + N$. The module M/N is called the *quotient module* of M by N.

Definition 9.2. There is also a natural analogue of linear maps for modules: if M_1, M_2 are R-modules, we say that $\phi \colon M_1 \to M_2$ is an *R-module homomorphism* (or just homomorphism) if:

(1) $\phi(m_1 + m_2) = \phi(m_1) + \phi(m_2)$, for all $m_1, m_2 \in M_1$,
(2) $\phi(r.m) = r.\phi(m)$, for all $r \in R, m \in M_1$,

that is, ϕ respects the addition and multiplication by ring elements. An isomorphism of R-modules is a homomorphism which is a bijection (and you can check, just as for groups, that this implies the inverse map of sets is also a homomorphism of modules). Just as the kernel and image of a linear map between vector spaces are subspaces, it is easy to see that $\ker(\phi) = \{m \in M_1 : \phi(m) = 0\}$ and $\operatorname{im}(\phi) = \{\phi(m) : m \in M_1\}$ are submodules of M_1 and M_2 respectively.

Remark 9.3. It is easy to check that if M_1, M_2 are R-modules then $\phi \colon M_1 \to M_2$ is an R-module homomorphism if and only if $\phi(v + tw) = \phi(v) + t\phi(w)$ for all $v, w \in M_1$ and $t \in R$.

Example 9.4. When R is a field, module homomorphisms are exactly linear maps. When $R = \mathbb{Z}$, a $\mathbb{Z}$-module homomorphism is just a homomorphism of the abelian groups. As another important example, it is easy to see that if M is an R-module and N is a submodule of M then the definition of the module structure on M/N

ensures precisely that the map $q\colon M \to M/N$ given by $q(m) = m + N$ is a (surjective) module homomorphism.

Lemma 9.5. *(Submodule correspondence:) Let M be an R-module and N a submodule. Let $q\colon M \to M/N$ be the quotient map. If S is a submodule of M then $q(S)$ is a submodule of M/N, while if T is a submodule of M/N then $q^{-1}(T)$ is a submodule of M. Moreover the map $T \mapsto q^{-1}(T)$ gives an injective map from submodules of M/N to the submodules of M which contain N, thus submodules of M/N correspond bijectively to submodules of M which contain N.*

Proof. The proof works precisely the same way as the proof of the correspondence between ideals given by a surjective ring homomorphism $\phi\colon R \to S$. Indeed that result is a special case of this Lemma, since ϕ makes S into an R-module, and ideals in S are precisely the R-submodules of S since ϕ is surjective.

To check that $q(S)$ and $q^{-1}(T)$ are submodules of N and M respectively follows directly from the definitions. We give the argument for $q^{-1}(T)$, the argument for $q(S)$ follows exactly the same pattern. If $m_1, m_2 \in q^{-1}(T)$ then $q(m_1), q(m_2) \in T$ and it follows since T is a submodule that $q(m_1) + q(m_2) = q(m_1 + m_2) \in T$ which says precisely that $m_1 + m_2 \in q^{-1}(T)$. Similarly if $r \in R$ then $q(r.m_1) = r.q(m_1) \in T$ since $q(m_1) \in T$ and T is a submodule, so that $r.m_1 \in q^{-1}(T)$. Thus $q^{-1}(T)$ is a submodule of M as required.

Now if T is *any* subset of M/N we have $q(q^{-1}(T)) = T$ simply because q is surjective. Since we have just checked $q^{-1}(T)$ is always a submodule in M, this immediately implies that the map $S \mapsto q(S)$ is a surjective map from submodules in M to submodules in M/N and that $T \mapsto q^{-1}(T)$ is an injective map[48], and moreover since $q(N) = \{0\} \subseteq T$ for any submodule T of M/N we have $N \subseteq q^{-1}(T)$ so that the image of the map $T \mapsto q^{-1}(T)$ consists of submodules of M which contain N. Hence it only remains to check that the submodules of M of the form $q^{-1}(T)$ are precisely these submodules. To see this suppose that S is an arbitrary submodule of M, and consider $q^{-1}(q(S))$. By definiton this is

$$\begin{aligned} q^{-1}(q(S)) &= \{m \in M : q(m) \in q(S)\} \\ &= \{m \in M : \exists s \in S \text{ such that } m + N = s + N\} \\ &= \{m \in M : \exists s \in S \text{ such that } m \in s + N\} \\ &= S + N. \end{aligned}$$

But if S contains N then we have $S + N = S$ and hence $q^{-1}(q(S)) = S$ and so any submodule S which contains N is indeed the preimage of a submodule of M/N as required. □

Remark 9.6. If $N \subseteq M$ is a submodule and $q\colon M \to M/N$ is the quotient map, for a submodule Q of M containing N we will usually write Q/N for the submodule $q(Q)$ of M/N.

[48] Again this is just set theory: if $f\colon X \to Y$ and $g\colon Y \to X$ are functions such that $g \circ f = \mathrm{id}_X$ then f is injective and g is surjective.

Theorem 9.7. *(Universal property of quotients.) Suppose that $\phi\colon M \to N$ is a homomorphism of R-modules, and S is a submodule of M with $S \subseteq ker(\phi)$ and let $q\colon M \to M/S$ be the quotient homomorphism. Then there is a unique homomorphism $\bar{\phi}\colon M/S \to N$ such that $\phi = \bar{\phi} \circ q$, that is, such that the following diagram commutes:*

$$\begin{tikzcd}
M \arrow[rr, "\phi"] \arrow[dr, "q"'] & & N \\
& M/S \arrow[ur, "\bar{\phi}"'] &
\end{tikzcd}$$

Moreover $ker(\bar{\phi})$ is the submodule $ker(\phi)/S = \{m + S : m \in ker(\phi)\}$.

Proof. The proof exactly mirrors the case for rings. Since q is surjective, the formula $\bar{\phi}(q(m)) = \phi(m)$ uniquely determines the values of $\bar{\phi}$, so that $\bar{\phi}$ is unique if it exists. But if $m-m' \in S$ then since $S \subseteq \ker(\phi)$ it follows that $0 = \phi(m-m') = \phi(m)-\phi(m')$ and hence ϕ is constant on the S-cosets, and therefore induces a map $\bar{\phi}(m + S) = \phi(m)$. The fact that $\bar{\phi}$ is a homomorphism then follows directly from the definition of the module structure on the quotient M/S, and clearly $\phi = \bar{\phi} \circ q$ by definition. To see what the kernel of $\bar{\phi}$ is, note that $\bar{\phi}(m + S) = \phi(m) = 0$ if and only if $m \in \ker(\phi)$, and hence $m + S \in \ker(\phi)/S$ as required. □

Corollary 9.8. *Let M be an R-module. We have the following isomorphisms.*

i) (First isomorphism theorem.) If $\phi\colon M \to N$ is a homomorphism then ϕ induces an isomorphism $\bar{\phi}\colon M/ker(\phi) \to im(\phi)$.

ii) (Second isomorphism theorem.) If M is an R-module and N_1, N_2 are submodules of M then

$$(N_1 + N_2)/N_2 \cong N_1/N_1 \cap N_2,$$

iii) (Third isomorphism theorem.) Suppose that $N_1 \subseteq N_2$ are submodules of M. Then we have

$$(M/N_1)/(N_2/N_1) \cong M/N_2.$$

Proof. The proofs again are exactly the same as for rings. For the first isomorphism theorem, apply the universal property to $S = \ker(\phi)$. Since in this case $\ker(\bar{\phi}) = \ker(\phi)/\ker(\phi) = 0$ it follows $\bar{\phi}$ is injective and hence induces an isomorphism onto its image which from the equation $\bar{\phi} \circ q = \phi$ must be exactly $\mathrm{im}(\phi)$.

For the second isomorphism theorem, let $q\colon M \to M/N_2$ be the quotient map. It restricts to a homomorphism p from N_1 to M/N_2, whose image is clearly $(N_1+N_2)/N_2$, so by the first isomorphism theorem it is enough to check that the kernel of p is $N_1 \cap N_2$. But this is clear: if $n \in N_1$ has $p(n) = 0$ then $n + N_2 = 0 + N_2$ so that $m \in N_2$, and so $n \in N_1 \cap N_2$.

For the third isomorphism theorem, let $q_i\colon M \to M/N_i$ for $i = 1, 2$. By the universal property for q_2 with $S = N_1$ we see that there is a homomorphism $\bar{q}_2\colon M/N_1 \to M/N_2$ induced by the map $q_2\colon M \to M/N_2$, with kernel $\ker(q_2)/N_1 = N_2/N_1$ and $\bar{q}_2 \circ q_1 = q_2$. Thus $\bar{q}_2$ is surjective (since q_2 is) and hence the result follows by the first isomorphism theorem. □

10. Free, torsion and torsion-free modules.

All rings R in this section are integral domains unless otherwise stated.

The fact that the nonzero elements of a ring do not have to be invertible means that modules behave less uniformly than vector spaces do. For example, you know that any vector space has a basis, and hence in the above terminology it is free. However, over $\mathbb{Z}$ there are many modules which are not free: indeed a finite abelian group is certainly a $\mathbb{Z}$-module, but it cannot be free since a free module must contain infinitely many elements (if an element m is part of a basis, it is easy to check that the elements $n.m$ must all be distinct for $n \in \mathbb{Z}$). In fact if M is a finite abelian group, every element is of finite order by Lagrange's theorem, which means that for every element $m \in M$ there is an integer $n \in \mathbb{N}$ such that $n.m = 0$. This is one important way in which a module over a general ring can be different from the case of vector spaces: we may have a nonzero scalar r and a nonzero element m of our module M whose product $r.m$ is nevertheless equal to zero. This is similar to the fact that a general ring may contain zero-divisors.

Definition 10.1. Let M be an R-module and suppose that $m \in M$. Then the *annihilator* of m, denoted $\text{Ann}_R(m)$ is $\{r \in R : r.m = 0\}$. A direct check shows that $\text{Ann}_R(m)$ is an ideal in R. When $\text{Ann}_R(m)$ is nonzero we say that $m \in M$ is a *torsion* element.

We say that a module M is *torsion* if every $m \in M$ is a torsion element. On the other hand, if a module M has no nonzero torsion elements we say that M is *torsion-free*. Note that a ring is an integral domain if and only if it is torsion-free as a module over itself, *i.e.* torsion elements in the R-module R itself are exactly the zero-divisors in R.

Remark 10.2. If M is an R-module, and $m \in M$ then the submodule $R.m$ of M generated by m is isomorphic as an R-module to $R/Ann_R(m)$. Indeed the map $r \mapsto r.m$ defines an R-module homomorphism from R to M whose image is exactly $R.m$. Since the kernel of the map is evidently $\text{Ann}_R(m)$ the isomorphism follows from the first isomorphism theorem. (Note this also shows $\text{Ann}_R(m)$ is an ideal, though this is also completely straight-forward to see directly.)

Definition 10.3. A module which is generated by a single element is known as a *cyclic* module. It follows from what we have just said that any cyclic module is isomorphic to a module of the form R/I where I is an ideal of R (corresponding to the annihilator of a generator of the cyclic module).

Recall from above that we say a module M is free if it has a basis S. The case where S is finite is the one of most interest to us. Then, just as picking a basis of a vector space gives you coordinates[49] for the vector space, the basis S allows us to write down an isomorphism $\phi\colon M \to R^n$ where $n = |S|$. Indeed if $S = \{s_1, s_2, \dots, s_n\}$ and $m \in M$ then we may write $m = \sum_{i=1}^n r_i s_i$ for a unique n-tuple $(r_1, r_2, \dots, r_n) \in$

[49] That is, if V is an n-dimensional $\mathbb{R}$-vector space, the fact that a choice of basis for V gives you an isomorphism from V to $\mathbb{R}^n$ is just a formal way of saying that picking a basis gives you coordinates for V.

R^n, and we set $\phi(m) = (r_1, \ldots, r_n)$. It is straight-forward to check that ϕ is then an isomorphism of modules.

It is easy to see that when R is an integral domain a free module must be torsion free, but the converse need not be true in general, as the next example shows. On the other hand, for principal ideal domains, whose modules will be our main focus, we will shortly see that torsion-free modules are actually free.

Example 10.4. Let $R = \mathbb{C}[x, y]$ be the ring of polynomials in two variables. Then the ideal $I = \langle x, y \rangle$ is a module for R. It is torsion-free because R is an integral domain (and I is a submodule of R) but it is a good exercise[50] to see that it is not free. The study of modules over a polynomial ring with many variables is a basic ingredient in algebraic geometry, and the commutative algebra course in Part B focuses largely the study of these rings and their quotients.

Recall we also had the notion of a torsion element in a module.

Lemma 10.5. *Let M be an R-modules, and let $M^{tor} = \{m \in M : Ann_R(m) \neq \{0\}\}$ is a submodule of M. Moreover, the quotient module M/M^{tor} is a torsion-free module.*

Proof. Let $x, y \in M^{\text{tor}}$. Then there are nonzero $s, t \in R$ such that $s.x = t.y = 0$. But then $s.t \in R\backslash\{0\}$, since R is an integral domain, and $(s.t)(x + y) = t.(s.x) + s.(t.y) = 0$, and clearly if $r \in R$ then $s.(r.x) = r.(s.x) = 0$, so that it follows M^{tor} is a submodule of M as required.

To see the moreover part, suppose that $x + M^{\text{tor}}$ is a torsion element in M/M^{tor}. Then there is a nonzero $r \in R$ such that $r.(x + M^{\text{tor}}) = 0 + M^{\text{tor}}$, that is, $r.x \in M^{\text{tor}}$. But then by definition there is an $s \in R$ such that $s.(r.x) = 0$. But then $s.r \in R$ is nonzero (since R is an integral domain) and $(s.r).x = 0$ so that $x \in M^{\text{tor}}$ and hence $x + M^{\text{tor}} = 0 + M^{\text{tor}}$ so that M/M^{tor} is torsion free as required. □

We will study finitely generated modules for a PID via the study of free modules. The free modules are, in a sense, the ones whose behaviour is closest to that of vector spaces over a field. In particular we will be able to understand maps between free modules in terms of matrices just like we do in linear algebra.

We first show that there is an analogue of the notion of dimension for a free module: Just as for vector spaces, the size of a basis for a free module is uniquely determined (even though a free module may have many different bases, just as for vector spaces).

Lemma 10.6. *Let M be a finitely generated free R-module. Then the size of a basis for M is uniquely determined and is known as the rank $\text{rk}(M)$ of M.*

Proof. Let $X = \{x_1, \ldots, x_n\}$ be a basis of M. Pick a maximal ideal[51] I in R. Let IM be the submodule generated by the set $\{i.m : i \in I, m \in M\}$ and let $M_I = \{\sum_{i=1}^n r_i x_i : r_i \in I\}$. Since I is an ideal it is easy to check that M_I is a submodule of M. We claim that

[50] Which is on Problem Set 4!

[51] In a PID we know that maximal ideals exist – if R is a field then we take $I = 0$, otherwise we take aR for $a \in R$ an irreducible element. In a general ring maximal ideals also always exist if you assume the axiom of choice.

$M_I = IM$. In fact, since X generates M, any element of the from $r.m$ where $r \in I$ and $m \in M$ lies in M_I, so that $IM \subseteq M_I$. On the other hand, certainly IM contains $r_i x_i$ for any $r_i \in I$, $i \in \{1, 2, \ldots, n\}$, and so all sums of the form $\sum_{i=1}^n r_i x_i$, and so $M_I \subseteq IM$ and hence $M_I = IM$ as required. Notice that in particular this means the submodule $M_I = IM$ does not depend on the choice of a basis of X.

Let $q \colon M \to M/IM$ be the quotient map. The quotient module M/IM is module for not just R, but in fact[52] for the quotient field $\mathsf{k} = R/I$, via the action $(r + I).q(m) = q(r.m)$. Indeed we just need to check this definition does not depend on the choice of $r \in r + I$. But if $r - r' \in I$ then $r.m - r'.m = (r - r').m \in IM$ and so $q(r'.m) = q(r.m)$ as claimed.

We now claim that if X is a basis for M then $q(X)$ is a basis for the k-vector space M/IM. Note that if we assume the claim then $|X| = \dim_{\mathsf{k}}(M/IM)$ and the right-hand side is clearly independent of X (since we have checked that the submodule IM is) so this will finish the proof of the Lemma. To prove the claim first note that since X generates M and q is surjective it follows that $q(X)$ generates (*i.e.* spans) M/IM. Now suppose we have $\sum_{i=1}^n c_i q(x_i) = 0 \in M/IM$, where $c_i \in \mathsf{k}$. Picking any representatives $r_i \in R$ for the $c_i \in R/I$ we see that

$$0 = \sum_{i=1}^n c_i q(x_i) = \sum_{i=1}^n q(r_i x_i) = q(\sum_{i=1}^n r_i x_i)$$

where the second equality follows from the definition of the R/I-action, and the lasts from the fact that q is an R-module homomorphism. But then it follows that $y = \sum_{i=1}^k r_i x_i \in \ker(q) = IM$. But since $IM = M_I$ this means that $r_i \in I$ for each i, that is $c_i = 0$. It follows $\bar{X}$ is linearly independent and hence a k-basis of M/IM as required. □

We will shortly see that any finitely generated module is a quotient of a free module R^n for some n. It will therefore be important to understand submodules of free modules. If R is a PID (as we will from now on assume) then the submodules of free modules are particularly well behaved.

Proposition 10.7. *Let M be a finitely generated free module over R a PID, and let $X = \{e_1, \ldots, e_n\}$ be a basis. Then if N is a submodule of M, N is also free and has rank at most n elements.*

$N \subseteq M \Rightarrow rk(N) \leq rk(M)$

Proof. We prove this by induction on $n = |X|$. If $n = 1$, then if $\phi \colon R \to M$ is the homomorphism defined by $\phi(r) = re_1$, the first isomorphism theorem shows that $M \cong R$. Now a submodule of R is just an ideal, and hence the condition that R is a PID exactly ensures any submodules N of R is cyclic, say $N = Ra$. Since R is an integral domain, it follows that $\{a\}$ is a basis of N unless $a = 0$, so that N is free of rank 0 or 1 and the $n = 1$ case is thus established.

Now suppose that $n > 1$. Let $W = Re_1 + Re_2 + \ldots Re_{n-1}$, and let $N_1 = N \cap W$. Now N_1 is a submodule of the free module W which has rank $n - 1$, so that by induction

[52] This is exactly what the submodule IM is cooked up to do – if you like M/IM is the largest quotient of M on which R/I acts naturally.

N_1 is free of rank $k \leq n-1$. Let $\{v_1, \ldots, v_k\}$ be a basis of N_1. The second isomorphism theorem shows that $N/N_1 \cong (N+W)/W \subseteq M/W$. But M/W clearly has basis $\{e_n + W\}$, so that N/N_1 is either zero or free of rank 1. In the former case $N = N_1$ and we are done, while in the latter we may pick v_{k+1} so that $v_{k+1} + N_1$ is a basis of N/N_1. We claim $\{v_1, \ldots, v_{k+1}\}$ is a basis of N. If $m \in N$ then since $\{v_{k+1} + N_1\}$ is a basis of N/N_1, we may find $r_{k+1} \in R$ such that $m + N_1 = r_{k+1}v_{k+1} + N_1$. But then $m - r_{k+1}v_{k+1} \in N_1$, and so since N_1 has basis $\{v_1, \ldots, v_k\}$ there are $r_1, \ldots, r_k \in R$ such that $m - r_{k+1}v_{k+1} = \sum_{i=1}^{k} r_k v_k$, that is, $m = \sum_{i=1}^{k+1} r_i v_i$. To see $\{v_1, \ldots, v_{k+1}\}$ is linearly independent, suppose that $\sum_{i=1}^{k+1} s_i v_i = 0$. Then $0 + N_1 = \sum_{i=1}^{k+1} s_i v_i + N_1 = s_{k+1}v_{k+1} + N_1$, so that we must have $s_{k+1} = 0$ since $\{v_{k+1} + N_1\}$ is a basis of N/N_1. But then $\sum_{i=1}^{k} s_k v_k = 0$ and hence $s_i = 0$ for all i, $1 \leq i \leq k$ since $\{v_1, \ldots, v_k\}$ is a basis of N_1. It follows $\{v_1, \ldots, v_{k+1}\}$ is a basis of N and since $k + 1 \leq (n-1) + 1 = n$ we are done.

□

Remark 10.8. Although it is noted in the proof above, it is worth emphasising that if R is an integral domain, then the submodules of a free module of rank d are free of rank at most d *if and only if* R is a PID, because the case of a free module of rank 1 requires that ideals of R must be principal.

Example 10.9. Suppose that $N \subset \mathbb{Z}^3$ is the submodule

$$N = \{(a, b, c) \in \mathbb{Z}^3 : a + b + c \in 2\mathbb{Z}\}.$$

Proposition 10.7 tells us that N must be free of rank at most 3, but let's use the strategy of proof to actually find a basis. Let $\{e_1, e_2, e_3\}$ be the standard basis of $\mathbb{Z}^3$ and let $M_i = Re_1 + Re_2 + \ldots Re_i$, so that each of M_i/M_{i-1} is isomorphic to R via the map induced by projecting to the i-th coordinate. Similarly let $N_i = N \cap M_i$, so that $N_1 \subseteq N_2 \subseteq N_3 = N$. Now $N_1 = \{(2a, 0, 0)\}$, so that $(2, 0, 0)$ is obviously a generator. For $N_2 = \{(a, b, 0) : a + b \in 2\mathbb{Z}\}$ we have $N_2/N_1 \cong (N_2 + M_1)/M_1$ and this is clearly all of M_2/M_1 (since the map is given by $(a, b, c) \mapsto b$ and b is clearly arbitrary), so has a basis $e_2 + M_1$. We can lift this to an element of N_2 by taking $(99, 1, 0)$ say. Finally taking $N_3/N_2 = N/N_2$ we again see that it is all of M/M_2 so that we can pick $(0, 89, 1)$ as a generator, and so $\{(2, 0, 0), (99, 1, 0), (0, 89, 1)\}$ is a basis of N

10.1. **Homorphisms between free modules.** In this section we want to study "linear maps" (or module homomorphisms) between free modules. The advantage of working with free modules here is that by choosing bases, we can record such a map using a matrix, just like you do in linear algebra over a field. Indeed all the results of this section have exactly the same proofs as the corresponding results for vector spaces, which were discussed in Prelims Linear algebra I (see Section §6.5 of the online notes for that course for example, or the Part A Linear Algebra notes for Lectures 1 and 2). We begin with a little notation.

Definition 10.10. If M, N are R-modules, let $\text{Hom}_R(M, N)$ denote the set of module homomorphisms from M to N. It is an R-module: if $\psi, \phi \in \text{Hom}_R(M, N)$ then $\psi + \phi$ is a module homomorphism (where $(\psi + \phi)(m) = \psi(m) + \phi(m)$) and $r.\psi$ is defined by $(r.\psi)(m) = r.(\psi(m))$.

Notice that the scalar multiplication gives a module structure only when R is commutative.

Lemma 10.11. *Let M and N be R-modules, and let $\phi\colon M \to N$ be an R-module homomorphism*

i) Let X be a spanning set for M, then ϕ is uniquely determined by its restriction to X.

ii) If X is a basis of M then given any function $f\colon X \to N$ there is a unique R-module homomorphism $\phi_f\colon M \to N$.

Proof. If $v \in F$, then since X spans M, there are elements $x_1, \dots, x_n \in X$ and $r_1, \dots r_n \in R$ such that $v = \sum_{i=1}^{n} r_i x_i$. Since ϕ is an R-homomorphism it follows $\phi(v) = \sum_{i=1}^{n} r_i \phi(x_i)$, so $\phi(v)$ is uniquely determined by the $\{\phi(x_i) : 1 \leq i \leq n\}$.

If X is also a basis of M we can reverse this process: given $f\colon X \to N$ since the expression for $v \in M$ in terms of the elements of X is unique, we get a well-defined function $\phi_f\colon M \to N$ by setting, for $v = \sum_{i=1}^{n} r_i x_i$ (where $r_i \in R, x_i \in X$, $1 \leq i \leq n$), $\phi(v) = \sum_{i=1}^{n} r_i f(x_i)$. This function is R-linear again because of uniqueness: if $v = \sum_{i=1}^{n} r_i x_i$ and[53] $w = \sum_{i=1}^{n} s_i x_i$ then for $t \in R$ we have $v + tw = \sum_{i=1}^{n} (r_i + ts_i)x_i$, hence

$$\phi_f(v + tw) = \sum_{i=1}^{n} (r_i + ts_i)x_i = \sum_{i=1}^{n} r_i x_i + t \sum_{i=1}^{n} s_i x_i = \phi(v) + t\phi(w),$$

as required. □

Corollary 10.12. *Let $\phi\colon F_1 \to F_2$ be a homomorphism of free modules with bases $X_1 = \{e_1, \dots, e_m\}$ and $X_2 = \{f_1, \dots, f_n\}$ respectively. Then ψ is determined by the matrix $A = (a_{ij}) \in Mat_{n,m}(R)$ given by*

$$\phi(e_i) = \sum_{j=1}^{n} a_{ji} f_j. \tag{10.1}$$

Conversely given a matrix $A \in Mat_{n,m}(R)$ the above formula determines a unique R-homomorphism $\phi_A\colon F_1 \to F_2$.

Proof. This follows immediately from the above, since the matrix A records (once we know the bases X_1 and X_2) since the map ϕ completely as it records the values of ϕ on X_1. Similarly, if we are given a matrix A, we may define a function $f\colon X_1 \to N$ using Equation (10.1), which extends uniquely to an R-module homomorphism $\phi_A\colon F_1 \to F_2$. □

Exactly as in linear algebra, composition of R-module homomorphisms corresponds to matrix multiplication: As we will use change of bases matrices (at least the special ones corresponding to elementary row and column operations) we review briefly the details now, but the arguments are exactly the same as in Prelims Linear Algebra.

[53] We can assume v, w lie in the span of some finite subset $\{x_1, \dots, x_n\}$ of X – by definition each of them does and the union of two finite sets is finite!

Lemma 10.13. *Let F_1, F_2, F_3 be free modules with bases $X_1 = \{e_1, e_2, \dots, e_n\}$ and $X_2 = \{f_1, f_2, \dots, f_m\}$ and $X_3 = \{g_1, g_2, \dots, g_l\}$ respectively. If $\phi\colon F_1 \to F_2$ and $\psi\colon F_2 \to F_3$, the matrix of ϕ with respect to the bases X_1 and X_2 is A and the matrix of ψ with respect to the bases X_2 and X_3 is B, then the matrix of the homomorphism $\psi \circ \phi$ with respect to the bases X_1 and X_3 is BA.*

Proof. We just need to compute $\psi \circ \phi(e_i)$ in terms of the basis $\{g_1, \dots, g_l\}$. But we have

$$\begin{aligned}\psi \circ \phi(e_i) = \psi(\sum_{j=1}^{m} a_{ji} f_j) &= \sum_{j=1}^{m} a_{ji}\psi(f_j)\\ &= \sum_{j=1}^{m} a_{ji}\left(\sum_{k=1}^{l} b_{kj} g_k\right)\\ &= \sum_{k=1}^{l}\left(\sum_{j=1}^{m} b_{kj} a_{ji}\right) g_k,\end{aligned}$$

Thus the matrix of $\psi \circ \phi$ has (k, i) entry $\sum_{j=1}^{m} b_{kj} a_{ji}$ as required.

□

Corollary 10.14. *If F is a free module with basis $X = \{e_1, \dots, e_n\}$, then the set of isomorphisms $\psi\colon F \to F$ corresponds under the above map to $GL_n(R) = \{A \in Mat_n(R) : \exists B \in Mat_n(R), A.B = B.A = I_n\}$, that is, the group of units in the (noncommutative) ring $Mat_n(R)$. Moreover, given two bases $X = \{x_1, \dots, x_n\}$ and $Y = \{y_1, \dots, y_n\}$ there is a unique isomorphism $\psi\colon F \to F$ such that $\psi(x_i) = y_i$.*

Proof. The first statement follows from the fact that composition of morphisms corresponds to matrix multiplication, so that the map sending a homomorphism to the corresponding $n \times n$ matrix is a ring map. For the moreover, note that if X and Y are bases, there is a unique module homomorphism $\psi\colon F \to F$ such that $\psi(x_i) = y_i$ and a unique module homomorphism $\phi\colon F \to F$ such that $\phi(y_i) = x_i$. The composition $\phi \circ \psi$ satisfies $\phi \circ \psi(x_i) = x_i$, and hence (again by the uniqueness property) $\phi \circ \psi = \text{id}$. □

Exercise 10.15. Let $A \in \text{Mat}_n(R)$. The determinant function makes sense for square matrices with entries in any commutative ring. Characterize the group of invertible matrices $\text{GL}_n(R)$ in terms of the determinant function.

Remark 10.16. Lemma 10.13 makes it easy to see how changing the bases of the free modules effects the matrix we associate to a homomorphism. If F_1 and F_2 are free modules with bases X_1 and X_2 respectively, write ${}_{X_2}[\phi]_{X_1}$ for the matrix of the homomorphism ϕ with respect to the bases X_1, X_2. Let Y_1, Y_2 be another pair of bases for F_1 and F_2 respectively. If $A = {}_{X_2}[\phi]_{X_1}$ we would like to calculate ${}_{Y_2}[\phi]_{Y_1}$ in terms of A. To do this, let $Q = {}_{Y_1}[\text{id}_{F_1}]_{X_1}$, and let $P = {}_{X_2}[\text{id}_{F_2}]_{Y_2}$. Then it follows from Lemma 10.13 and the fact that $\phi = \text{id}_{F_2} \circ \phi \circ \text{id}_{F_1}$ that

$${}_{Y_2}[\phi]_{Y_1} = PAQ.$$

Definition 10.17. The matrices P and Q are called the *change of bases* matrices for the pairs of bases X_2, Y_2 and X_1, Y_1 respectively. They are readily computed: if F is a free module with two bases X and Y, the matrix ${}_Y[\mathrm{id}_F]_X$ has columns given by the "Y-coordinates" of the elements of the basis X: If $Y = \{f_1, \dots, f_n\}$ and $X = \{e_1, \dots, e_n\}$ then $e_j = \sum_{i=1}^n p_{ij} f_i$ where $P = (p_{ij})$ is the change of basis matrix. For example, if $F = R^n$ with standard basis $\{e_1, \dots, e_n\}$ (that is, $e_i = (0, \dots, 1, \dots 0)$ where the 1 is in position i) and $Y = \{f_1, \dots, f_n\}$ is any other basis, then the change of basis matrix from Y to the standard basis is just the matrix with columns the basis vectors f_i, and thus the change of basis matrix from the standard basis to the basis Y is given by the inverse of this matrix.

The above discussion shows that if $\phi\colon F_1 \to F_2$ is a homomorphism between free modules F_1 and F_2 of rank m and n with bases X_1, X_2 respectively, we may associate to ϕ a matrix A, and by picking other possible bases we obtain matrices of the form PAQ where $P \in \mathrm{GL}_n(R)$ and $Q \in \mathrm{GL}_m(R)$ are change of bases matrices. Thus the homomorphism ψ corresponds to the equivalence class of A in $\mathrm{Mat}_{n,m}(R)$ where X and Y are equivalent if there are invertible matrices P and Q such that $Y = PXQ$, that is, if they are in the same orbit of the natural action[54] of $\mathrm{GL}_n(R) \times \mathrm{GL}_m(R)$.

It follows that in order to find a "canonical form" for the homomorphism ϕ, that is, a matrix representing ϕ which is as simple as possible (and preferably that it be as unique as possible), we need to find a canonical element of each orbit of the action of $\mathrm{GL}_n \times \mathrm{GL}_m(R)$ on $\mathrm{Mat}_{n,m}(R)$. We will solve this problem in the next section when R is a Euclidean domain.

11. Canonical forms for matrices over a Euclidean Domain.

Before we begin finding this canonical form, it is worth recalling the situation for vector spaces: in this case the statement we want is essentially the rank-nullity theorem: if R is a field and $\psi\colon R^m \to R^n$, then you can pick a basis of $\ker(\psi) \subseteq R^m$ and extend it to a basis of R^m. The image of the vectors you add in to obtain a basis of R^m give a basis of $\mathrm{im}(\psi)$, which can in turn be extended to a basis of R^n. The matrix of ψ with respect to the resulting bases of R^m and R^n is then diagonal with 1s and 0s on the diagonal, where the rank of ψ is the number of 1s and the nullity the number of zeros. The key to this argument is the Steinitz Exchange Lemma, which is that allows you to show that in a vector space you can extend any linearly independent set to a basis. However, this lemma is obviously *false* for modules: for example $2\mathbb{Z}$ is free inside $\mathbb{Z}$, with basis $\{2\}$, but the only bases for $\mathbb{Z}$ are $\{1\}$ and $\{-1\}$, thus we cannot "extend" the basis $\{2\}$ to a basis of $\mathbb{Z}$. We will show in this section that, in a sense, this is the only thing that fails for modules over a Euclidean domain: that is, we will show that if N is a submodule of R^n there is always a basis of $\{e_1, \dots, e_n\}$ of R^n for which we can obtain a basis of N by taking appropriate multiples of a subset of the basis. Explicitly, perhaps after reordering the basis $\{e_1, \dots, e_n\}$, we will show

[54] Note that this action, if we want it to be a *left* action, should be $(P, Q).X = PXQ^{-1}$, but the inverse is not too important since we are only interested in the orbits of the action: A and B are in the same orbit if and only if there are invertible matrices P and Q such that $B = PAQ$.

that there are elements $c_1, \dots, c_k \in R$ and some $k \leq n$, such that $\{c_1e_1, \dots, c_ke_k\}$ is a basis for N.

The most explicit way to prove rank-nullity for linear maps between vector spaces is to use row and column operations (which correspond to particularly simple changes of the basis of the source and target of the linear map respectively). We will use the same idea for modules over a Euclidean domain.

Definition 11.1. Let $A \in M_{m,n}(R)$ be a matrix, and let $r_1, r_2, \dots, r_m$ be the rows of A, which are row vectors in R^n. An *elementary row operation* on a matrix $A \in M_{m,k}(R)$ is an operation of the form

(1) Swap two rows r_i and r_j.
(2) Replace one row, row i say, with a new row $r_i' = r_i + cr_j$ for some $c \in R$, and $j \neq i$.

In the same way, viewing A as a list of n column vectors, we define *elementary column operations*.

Note that the row operations correspond to multiplying A by elementary matrices on the left and the column operations correspond to multiplying A by elementary matrices on the right. Indeed if we let E_{ij} denote the matrix with (i, j)-th entry equal to 1 and all other entries zero, then the matrix corresponding to the first row operation is $S_{ij} = I_k - E_{ii} - E_{jj} + E_{ij} + E_{ji}$, while second elementary row operation is given by multiplying on the left by $X_{ij}(c) = I_k + cE_{ij}$. The column operations are given by multiplying on the right by these matrices.

$$S_{ij} = \begin{pmatrix} 1 & & & & & \\ & 0 & & & 1 & \\ & & 1 & & & \\ & & & \ddots & & \\ & 1 & & & 0 & \\ & & & & & 1 \end{pmatrix}$$

$$X_{ij}(c) = \begin{pmatrix} 1 & & & & & \\ & \ddots & & & & \\ & & 1 & & c & \\ & & & \ddots & & \\ & & & & 1 & \\ & & & & & 1 \end{pmatrix}$$

Definition 11.2. If $A, B \in \mathrm{Mat}_{n,m}(R)$ we say that A and B are *equivalent* if $B = PAQ$ where $P \in \mathrm{Mat}_{n,n}(R)$ and $Q \in \mathrm{Mat}_{m,m}(R)$ are invertible matrices. Thus two matrices are equivalent if and only if they lie in the same orbit of the $GL_n(R) \times GL_m(R)$ action defined above. We will say that A and B are *ERC equivalent* if one can be obtained from the other by a sequence of elementary row and column operations. Since row and column operations correspond to pre- and pos-multiplying a matrix by elementary matrices, it is clear that two ERC equivalent matrices are equivalent.

(In fact, if you also allow the elementary row and column operations which simply rescale a row or column by a unit then you can show the converse too, but we do not need that here.)

For the remainder of this section we assume that R is a Euclidean domain. Recall that we write $N\colon R\backslash\{0\} \to \mathbb{N}$ for the norm function of our Euclidean domain R.

Theorem 11.3. *Suppose that $A \in Mat_{n,m}(R)$ is a matrix. Then A is ERC equivalent (and hence equivalent) to a diagonal matrix D where if $k = min\{m, n\}$ then*[55]

$$D = \begin{pmatrix} d_1 & 0 & \dots & 0 \\ 0 & d_2 & \ddots & 0 \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & d_k \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 0 \end{pmatrix}$$

and each successive d_i divides the next (thus possibly $d_s = d_{s+1} = \dots d_k = 0$, for some s, $1 \leq s \leq k$). Moreover, the sequence of elements $(d_1, d_2, \dots, d_k)$ is unique up to units.

Proof. We will not prove the uniqueness statement (though see Problem Sheet 4 for how one can do this). We claim that by using row and column operations we can find a matrix equivalent to A which is of the from

$$B = \begin{pmatrix} b_{11} & 0 & \dots & 0 \\ 0 & b_{22} & \dots & b_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & b_{n2} & \dots & b_{nm} \end{pmatrix} \tag{11.1}$$

where b_{11} divides all the entries b_{ij} in the matrix. Factoring out b_{11} from each entry, we may then applying induction (on n say) to the submatrix $B' = (b_{ij}/b_{11})_{i,j\geq 2}$, to obtain the proposition. (Note that row and column operations on B' correspond to row and column operations on B because b_{11} is the only nonzero entry in the first row and column of B.)

We are thus reduced to proving the claim. For this we use induction on $N(A) = \min\{N(a_{ij}) : 1 \leq i \leq n, 1 \leq j \leq m\}$. Using row and column swaps we may assume $N(a_{11}) = N(A)$.

Step 1: If any a_{i1} or a_{1j} is not divisible by a_{11}, say a_{i1}, then $a_{i1} = q_{i1}a_{11} + r_{i1}$, so taking q_{i1} times row 1 from row i we get a new matrix A' with entry r_{i1} and $N(A_1) \leq N(r_{i1}) < N(a_{11}) = N(A)$, so we are done by induction.

Step 2: If all the a_{i1}s and a_{1j}s are divisible by a_{11} we my subtract appropriate multiples of the first row from the other rows to get a matrix A_2 with all entries in the first column below a_{11} equal to zero and similarly using column operations we can then get a matrix A_3 with all entries on the first row after a_{11} equal to zero.

[55] The displayed matrix shows the case where $k = m$, and so there are $(n - m)$ rows below d_k consist entirely of zeros. If $k = n$ then there are $(m - n)$ columns consisting entirely of zeros to the right of d_k.

Step 3: Thus A_3 has the form we require except perhaps it has an entry not divisible by a_{11}. Thus either we are done, or letting $(A_3) = (a^3_{ij})$ we have for some $i, j > 1$, the entry a^3_{ij} is not divisible by $a_{11} = a^3_{11}$. Then add row i of A_3 to row 1, to get a matrix A_4 where now and we see that we are back in the situation of step 1, and we are done by induction.

The claim and hence the theorem are thus proved. □

Example 11.4. The above proposition is really an algorithm, so lets use it in an example, taking $R = \mathbb{Z}$: Let

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

The entry of smallest norm is the $(3,2)$ entry, so we swap it to the $(1,1)$ entry (by swapping rows 1 and 3 and then columns 1 and 2 say) to get

$$A_1 = \begin{pmatrix} 1 & 3 & 0 \\ 6 & 8 & 4 \\ 5 & 2 & 3 \end{pmatrix}$$

Now since the $(1,1)$ entry is a unit, there will be no remainders when dividing so we get

$$A_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -10 & 4 \\ 0 & -13 & 3 \end{pmatrix}$$

Next we must swap the $(3,3)$-entry to the $(2,2)$-entry to get:

$$A_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 3 & -13 \\ 0 & 4 & -10 \end{pmatrix}$$

Dividing and repeating our row and column operations now on the second row and column (this time we do get remainders) gives:

$$A_4 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 3 & -13 \\ 0 & 1 & 3 \end{pmatrix} \sim A_5 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 3 & 2 \\ 0 & 1 & 8 \end{pmatrix}$$

(where $\sim$ is to denote ERC equivalence). Now moving the $(3,2)$ entry to the $(2,2)$-entry and dividing again gives:

$$A_6 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 8 \\ 0 & 3 & 2 \end{pmatrix} \sim A_7 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 3 & -22 \end{pmatrix} \sim A_8 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -22 \end{pmatrix}$$

which is in the required normal form.

12. Presentations and the canonical form for modules.

In this section all rings R are PIDs.

So far we have concentrated on the study of free modules. Aside from the fact that these are the simplest modules to study, we do this because we can understand any (finitely generated) module in terms of free modules. The key to this is the notion of a *presentation* of a module. This is a notion which is important in other parts of algebra also – if you take the Group Theory option next term for example you will see how it arises in that subject.

The goal is to describe a (finitely generated) module M in concrete terms: explicitly we will show that any finitely generated module can be described in terms of a finite generating set and a matrix recording the "relations" amongst the generators. To do this we will use matrix algebra we developed for homomorphisms of free modules in the previous section. Once we have this explicit description of M, we can use the canonical form theorem for the matrix describing M to obtain a canonical form for the module M.

Proposition 12.1. *i) Let M be a nonzero finitely generated module. Then there is an $n \in \mathbb{N}$ and a surjective morphism $\phi \colon R^n \to M$. In particular, $R^n/ker(\phi) \cong M$.*

ii) Let M and ϕ be as in i). There exists a free module R^m with $m \leq n$ and an injective homomorphism $\psi \colon R^m \to R^n$ such that $im(\psi) = ker(\phi)$. In particular, M is isomorphic to $R^n/im(\psi)$.

Proof. For the first part, given any finite subset $\{m_1, m_2, \dots, m_n\}$ of M, if $\{e_1, \dots, e_n\}$ is a basis of R^n (say the standard basis consisting of elements $e_i = (0, \dots, 1, \dots 0)$ all of whose coordinates are zero except for the i-the entry which is equal to 1) then the map $e_i \mapsto m_i$ $(1 \leq i \leq n)$ extends, by Lemma 10.11*ii*), to a homomorphism $\phi \colon R^n \to M$. Clearly the condition that $\{m_1, m_2, \dots, m_n\}$ is a generating set is then equivalent to the map ϕ being surjective, since both assert that any element of M can be written in the form $\sum_{i=1}^n r_i m_i = \phi(\sum_{i=1}^n r_i e_i)$ $(r_i \in R,\ 1 \leq i \leq n)$. The surjectivity of ϕ and the first isomorphism theorem then show that $R^n/\ker\phi \cong M$.

For the second part, note that since R is a PID, the submodule $\ker(\phi)$ is a free submodule of rank $m \leq n$. Pick a basis $\{x_1, \dots, x_m\}$ of $\ker(\phi)$, and define ψ by sending the standard basis of R^m to $\{x_1, \dots, x_m\}$. This map is then clearly injective and has image exactly $\ker(\phi)$ as required. □

Definition 12.2. Let M be a finitely generated R-module. The pair of maps ϕ, ψ of the previous Lemma, so that $\mathrm{im}(\phi) = M$ and $\psi \colon R^m \to R^n$ has image $\mathrm{im}(\psi) = \ker(\phi)$ is called a *presentation* of the finitely generated modules M. When the map ψ can be chosen to be injective, the presentation is called a *resolution* of the module M. It is a special feature of modules over a PID that, for a finitely generated module, there are presentations which are also resolutions. Future courses in commutative algebra and what is called homological algebra study what happens to these two notions for more general rings.

Remark 12.3. (*Non-examinable.*) The properties of the above homomorphisms ψ and ϕ can be captured by noticing that

$$\begin{tikzcd} 0 \arrow[r] & R^m \arrow[r, "\psi"] & R^n \arrow[r, "\phi"] & M \arrow[r] & 0 \end{tikzcd}$$

is what is called a *short exact sequence*: An *exact sequence* is a sequence of homomorphisms where the image of each map is the kernel the next map in the sequence. A *short* exact sequence is one with exactly five terms, the outermost two terms both being 0. Exact sequences play an important role in algebraic topology and homological algebra.

To see why, in more concrete terms, one calls this a presentation, lets make explicit what we have done. If $\{e_1, \ldots, e_m\}$ is the standard basis of R^m and $\{f_1, \ldots, f_n\}$ is the standard basis of R^n, then just as in linear algebra, we may write

$$\psi(e_j) = \sum_{i=1}^{n} a_{ij} f_i$$

for some $a_{ij} \in R$, and the resulting matrix $A = (a_{ij})_{1 \leq i \leq n, 1 \leq j \leq m}$ encodes the homomorphism ψ. Describing a module M as the quotient $R^n/\text{im}(\psi)$ says that M has generators $m_1, \ldots, m_n$ (the images of the elements $f_i + \text{im}(\phi) \in R^n/\text{im}(\psi)$ under the isomorphism from $R^n/\text{im}(\psi) \to M$ induced by ϕ) and the R-linear dependencies these generators satisfy are all consequences of the m equations:

$$\sum_{i=1}^{n} a_{ij} m_i = 0 \quad (j = 1, 2, \ldots, m).$$

Thus the map $\phi\colon R^n \to M$ picks out the generators we use for M and the map ψ records the relations, or linear dependencies, among these generators: that they are R-linear relations among the generators follows because $\phi \circ \psi = 0$, while the fact that all other relations are a consequence of these follows because the elements $(\sum_{i=1}^{n} a_{ij} f_i)_{j=1}^{m}$ are a basis for $\ker(\phi) = \text{im}(\psi)$. Indeed if we have a relation $\sum_{i=1}^{n} r_i m_i = 0$, then it follows that $\phi(\sum_{i=1}^{n} r_i f_i) = 0$, that is $\sum_{i=1}^{n} r_i f_i \in \text{im}(\psi)$, which means that there are are scalars $(s_i)_{i=1}^{m}$ such that $\sum_{i=1}^{n} r_i f_i = \sum_{i=1}^{m} s_i \psi(e_i) = \sum_{i=1}^{m} s_i (\sum_{i=1}^{n} a_{ij} f_i)$. In other words, the R-linear relation given by the scalars $(r_i)_{i=1}^{n}$ can be obtained as a linear combination (given by the $(s_j)_{j=1}^{m}$) of the relations $\sum_{i=1}^{n} a_{ij} m_i = 0$ $(1 \leq j \leq m)$. Up to isomorphism then, the structure of the module M is captured by the matrix of relations $A = (a_{ij})$. It is this which allows us to use our canonical form theorem for matrices over a Euclidean Domain to obtain a canonical form for finitely generated modules over such rings.

For the rest of this section, we will assume all rings are Euclidean Domains, although all results stated here actually also hold more generally for PIDs.

Theorem 12.4. *Suppose that M is a finitely generated module over a Euclidean domain R. Then there is an integer s and nonzero nonunits $c_1, c_2, \ldots, c_r \in R$ such that $c_1 | c_2 | \ldots | c_r$ such that:*

$$M \cong (\bigoplus_{i=1}^{r} R/c_i R) \oplus R^s.$$

Proof. Since R is a PID we may find a presentation for M, that is, an injection $\psi\colon R^m \to R^n$ (so that $m \leq n$) and a surjection $\phi\colon R^n \to M$ with $\ker(\phi) = \operatorname{im}(\psi)$, so that $M \cong R^n/\operatorname{im}(\psi)$. Now if A is the matrix of ψ with respect to the standard bases of R^m and R^n, by Theorem 11.3, which gives a normal form for matrices over a Euclidean domain, we know we can transform A into a diagonal matrix D with diagonal entries $d_1|d_2|\ldots d_m$ using elementary row and column operations. But since row and column operations correspond to pre- and post-multiplying A by invertible matrices, and these correspond to changing bases in R^n and R^m respectively, it follows that there are bases of R^n and R^m with respect to which ψ has matrix D. But then if $\{f_1 \ldots, f_n\}$ denotes the basis of R^n, we see that the image of ψ has basis $\{d_1 f_1, \ldots, d_m f_m\}$. Now define a map $\theta\colon R^n \to (\bigoplus_{i=1}^m R/d_iR) \oplus R^{n-m}$ by setting for any $m = \sum_{i=1}^n a_i f_i \in M$,

$$\theta(\sum_{i=1}^n a_i f_i) = (a_1 + d_1R, \ldots a_m + d_mR, a_{m+1}, \ldots, a_n).$$

It is the clear that θ is surjective and $\ker(\theta)$ is exactly the submodule generated by $\{d_i f_i : 1 \leq i \leq m\}$, that is, $\operatorname{im}(\psi)$. It follows by the first isomorphism theorem that $M \cong R^k/\operatorname{im}(\psi) \cong \bigoplus_{i=1}^m (R/d_iR) \oplus R^{k-m}$ as required.

Finally, since ψ is injective it follows that each of the d_i are nonzero. On the other hand if d_i is a unit (and so all d_j for $j \leq i$ are also) then $R/d_iR = 0$, so this summand can be omitted from the direct sum. The result now follows. □

Remark 12.5. The sequence of elements $\{c_1, c_2, \ldots, c_r\}$ are in fact unique up to units. We won't have time to show this here (the problem sheets asks you to show uniqueness for c_1 and $c_1 \ldots c_m$ at least given a presentation.). The integer s is also unique, which we now show as a consequence of the important corollary to the structure theorem which says that a finitely generated torsion-free R-module is free.

Corollary 12.6. *Let M be a finitely generated torsion-free module over R. Then M is free. In general if M is a finitely generated R-module, the rank s of the free part of M given in the structure theorem is $\operatorname{rk}(M/M^{\text{tor}})$ and hence it is unique.*

Proof. By the above structure theorem, M is isomorphic to a module of the form $R^s \oplus (\bigoplus_{i=1}^r R/c_iR)$, thus we can assume M is actually equal to a module of this form.

Let $F = R^s$ and $N = \bigoplus_{i=1}^r R/c_iR$, so that $M = F \oplus N$. We claim that $N = M^{\text{tor}}$. Certainly if $a \in R/c_iR$ then since $c_i|c_k$ we see that $c_k(a) = 0$. But then if $m \in N$, say $m = (a_1, \ldots, a_k)$ where $a_i \in R/c_iR$ it follows $c_k(a_1, \ldots, a_m) = (c_k a_1, \ldots, c_k a_k) = (0, \ldots, 0)$ so N is torsion. On the other hand if $m = (f, n)$ where $f \in F$ and $n \in N$ then $r(f, n) = (r.f, r.n) = (0, 0)$ we must have $f = 0$ since a free module is torsion-free. Thus $M^{\text{tor}} = N$ as claimed. It follows that M is torsion-free if and only if $M = F$ is free. Moreover, by the second isomorphism theorem $F \cong M/M^{\text{tor}}$ (or more directly, just by noting that the restriction of the quotient map $q\colon M \to M/N = M/M^{\text{tor}}$ to F is an isomorphism since it is readily seen to be injective and surjective) so that $s = \operatorname{rk}(F) = \operatorname{rk}(M/M^{\text{tor}})$. □

(*Note that Problem sheet* 4 *gives an alternative proof that a torsion-free module over a PID is free using just Proposition 10.7.*)

Just to make it explicit, notice that since an abelian group is just a $\mathbb{Z}$-module, our structure theorem gives us a classification theorem for finitely generated abelian groups.

Corollary 12.7. *(Structure theorem for finitely generated abelian groups) Let A be a finitely generated abelian group. Then there exist an integer $r \in \mathbb{Z}_{\geq 0}$ and integers $c_1, c_2, \ldots, c_k \in \mathbb{Z}$ greater than* 1 *such that $c_1|c_2|\ldots|c_k$ and*

$$A \cong \mathbb{Z}^r \oplus (\mathbb{Z}/c_1\mathbb{Z}) \oplus \ldots \oplus (\mathbb{Z}/c_k\mathbb{Z}).$$

Moreover the integers $s, c_1, \ldots, c_k$ are uniquely determined.

Proof. This is simply a restatement of the previous theorem, except that once we insist the c_i are positive the ambiguity caused by the unit group $\mathbb{Z}^\times = \{\pm 1\}$ is removed. □

We can give an alternative formulation of the canonical form theorem, known as the *primary decomposition* form, using the Chinese Remainder Theorem, or the following slight generalization of it.

Lemma 12.8. *Let $d_1, \ldots, d_k \in R$ be a set of pairwise coprime elements of a PID, that is h.c.f$\{d_i, d_j\} = 1$ if $i \neq j$. Then*

$$R/(d_1 d_2 \ldots . d_k)R = \bigoplus_{i=1}^{k} R/d_i R.$$

Proof. The condition that the d_is are pairwise coprime means that if we set, for $i < k$, $c_i = d_{i+1} \ldots d_k$ then for each i we have *h.c.f*$\{d_i, c_i\} = 1$ (indeed if p is a prime element dividing d_i and c_i then since p is prime it divides one of the factors $d_{i+1}, \ldots, d_k$ of c_i, say d_j where $j > i$. But then p divides *h.c.f*$\{d_i, d_j\}$ contradicting our assumption). Thus we see that for each i with $1 \leq i \leq k-1$ we have $d_i R + c_i R = R$ and $d_i R \cap c_i R = d_i c_i R$. Thus by the Chinese Remainder Theorem and induction on k we see that

$$R/(d_1 \ldots d_k)R = R/(d_1 c_1 R) \cong R/d_1 R \oplus R/(c_1)R \cong \bigoplus_{i=1}^{k} R/d_i R.$$

where in the last isomorphism we may use induction on the factor $R/c_1 R$ since $c_1 = d_2 \ldots d_k$ is a product of $k-1$ pairwise coprime factors. □

In particular if $c = p_1^{n_1} \ldots . p_k^{n_k}$ is the prime factorisation of c where the p_i are distinct primes we can apply the previous Lemma (with $d_i = p_i^{n_i}$) to see that

$$R/cR \cong \bigoplus_{i=1}^{r} R/p_i^{r_i} R. \tag{12.1}$$

This allows us to give an alternative statement of the structure theorem:

Theorem 12.9. *(Structure theorem in primary decomposition form): Let R be a Euclidean domain and suppose that M is a finitely generated R-module. Then there are irreducibles $p_1, \ldots, p_k \in R$ and integers s, r_i, $1 \leq i \leq k$, such that:*

$$M \cong \bigoplus_{i=1}^{k} (R/p_i^{r_i}R) \oplus R^s.$$

Moreover, the pairs (p_i, r_i) are uniquely determined up to units (where the units act on the p_i only). (Note however that the p_is are not necessarily distinct.)

Proof. This follows immediately using the decomposition (12.1) on each of the cyclic modules R/c_iR in the statement of our first structure theorem. □

Example 12.10. Suppose that $A \cong \mathbb{Z}/44\mathbb{Z} \oplus \mathbb{Z}/66\mathbb{Z}$. Then the first structure theorem would write A as:

$$A \cong \mathbb{Z}/22\mathbb{Z} \oplus \mathbb{Z}/132\mathbb{Z}.$$

Indeed the generators corresponding to the direct sum decomposition give a presentation of A as $\mathbb{Z}^2 \to \mathbb{Z}^2 \to A$ where the first map is given by the matrix

$$\begin{pmatrix} 44 & 0 \\ 0 & 66 \end{pmatrix}$$

and as $66 = 1.44 + 22$ we see that row and column operations allow us to show this matrix is equivalent to:

$$\begin{pmatrix} 44 & 0 \\ 0 & 66 \end{pmatrix} \sim \begin{pmatrix} 44 & 44 \\ 0 & 66 \end{pmatrix} \sim \begin{pmatrix} 44 & 44 \\ -44 & 22 \end{pmatrix} \sim \begin{pmatrix} 22 & -44 \\ 44 & 44 \end{pmatrix} \sim \begin{pmatrix} 22 & 0 \\ 0 & 132 \end{pmatrix}.$$

On the other hand, for the primary decomposition (since $44 = 2^2.11$ and $66 = 2.3.11$) we would write A as:

$$A \cong ((\mathbb{Z}/2\mathbb{Z}) \oplus (\mathbb{Z}/2^2\mathbb{Z})) \oplus (\mathbb{Z}/3\mathbb{Z}) \oplus (\mathbb{Z}/11\mathbb{Z})^{\oplus 2}$$

Notice that the prime 2 appears twice raised to two different powers. Intuitively you should think of the primary decomposition as decomposing a module into a direct sum of as many cyclic summands as possible, while the canonical form decomposes the module into a direct sum with as few cyclic summands as possible.

Remark 12.11. Note that the first structure theorem gives a canonical form which can be obtained algorithmically, while the second requires one to be able to factorise elements of the Euclidean domain, which for example in $\mathbb{C}[t]$ is *not* an automatically computable operation.

13. Application to rational and Jordan canonical forms.

The structure theorem also allows us to recover structure theorems for linear maps: If V is a k-vector space and $T\colon V \to V$ is a linear map, then we view V as a $\mathsf{k}[t]$-module by setting $t.v = T(v)$ for $v \in V$.

Lemma 13.1. *Let M be a finitely generated $\mathsf{k}[t]$-module. Then M is finite dimensional as a k-vector space if and only if M is a torsion $\mathsf{k}[t]$-module. Moreover, a subspace U of V is a $\mathsf{k}[t]$-submodule if and only if U is T-invariant, i.e. $T(U) \subseteq U$.*

Proof. Given M we can apply the structure theorem to see that:

$$M \cong \mathsf{k}[t]^s \oplus \mathsf{k}[t]/\langle c_1\rangle \oplus \ldots \oplus \mathsf{k}[t]/\langle c_k\rangle,$$

where $s \in \mathbb{Z}_{\geq 0}$ and the c_k are nonconstant[56] polynomials. Now $\mathsf{k}[t]$ is infinite dimensional as a k-vector space while $\mathsf{k}[t]/\langle f\rangle$ is $\deg(f)$-dimensional as a k-vector space, so if follows that M is torsion if and only if $s = 0$ if and only if M is finite dimensional as a k-vector space. For the final statement, notice that a subspace U of M is T-invariant if and only if it is $p(T)$-invariant for every $p \in \mathsf{k}[t]$. □

The Lemma shows that pairs (V, T) consisting of a finite dimensional k-vector space V and a linear map $T\colon V \to V$ correspond to finitely generated torsion $\mathsf{k}[t]$-modules under our correspondence above. We can use this to give structure theorems for endomorphisms[57] of a vector space. Note that the ambiguity about units in the statement of the canonical form theorem can be removed in the case of $\mathsf{k}[t]$-modules by insisting that the generators c_i of the annihilators of the cyclic factors $\mathsf{k}[t]/\langle c_i\rangle$ are taken to be monic.

Definition 13.2. For a monic polynomial $f = t^n + \sum_{i=0}^{n-1} a_i t^i \in \mathsf{k}[t]$ of degree $n \geq 1$, the $n \times n$ matrix[58]

$$C(f) = \begin{pmatrix} 0 & \ldots & \ldots & 0 & -a_0 \\ 1 & 0 & & \vdots & -a_1 \\ 0 & 1 & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & 0 & -a_{n-2} \\ 0 & \ldots & 0 & 1 & -a_{n-1} \end{pmatrix}.$$

is called the *companion matrix* of f.

If $\lambda \in \mathsf{k}$ and $n \in \mathbb{N}$, then let $J_n(\lambda)$ be the $n \times n$ matrix

$$J_n(\lambda) = \begin{pmatrix} \lambda & 1 & \ldots & \ldots & 0 \\ 0 & \lambda & 1 & \ldots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 1 \\ 0 & \ldots & & 0 & \lambda \end{pmatrix}$$

Lemma 13.3. *Let* k *be any field, and suppose that* $f \in \mathsf{k}[t]$ *is a monic polynomial and* $\lambda \in \mathsf{k}$.

1. *If* $f = t^n + \sum_{k=0}^{n-1} a_k t^k$ *then the* $\mathsf{k}[t]$*-module* $\mathsf{k}[t]/\langle f\rangle$ *has basis* $\{t^i + \langle f\rangle : 0 \leq i \leq n-1$ *and the matrix for the action of* t *with respect to this basis is given by* $C(f)$ *the companion matrix of* f.
2. *If* $g = (t-\lambda)^n \in \mathsf{k}[t]$ *for some* $\lambda \in \mathsf{k}$ *and* $n \in \mathbb{Z}_{>0}$, *then the* $\mathsf{k}[t]$*-module* $\mathsf{k}[t]/\langle g\rangle$ *has basis* $\{(t-\lambda)^k + \langle g\rangle : 0 \leq k \leq n-1\}$. *With respect to this basis, the action of* t *is given by the Jordan block matrix* $J_n(\lambda)$ *(where we order the basis by decreasing powers of* $(t-\lambda)$ *in order to get an upper triangular matrix).*

[56] *i.e.* nonzero nonunit elements if $\mathsf{k}[t]$.

[57] Recall that an endomorphism of a vector space V is just a linear map from V to itself.

[58] If $n = 1$ the matrix is $C(f) = (-a_0)$.

Proof. Recall that by the division algorithm for polynomials, each coset of $\langle f \rangle$ has a unique representative of degree strictly smaller than that of f. The assertion that the two sets in parts (1) and (2) are k-bases then follows because $(t-\lambda)^k : 0 \leq k \leq n-1\}$ is clearly a basis for the space of polynomials of degree at most $n-1$ for any $\lambda \in \mathsf{k}$. For the assertions about the action of t, note for (1) that $t.(t^i + \langle f \rangle) = t^{i+1} + \langle f \rangle$ for $i < n-1$, while if $i = n-1$, $t.t^{n-1} + \langle f \rangle = t^n + \langle f \rangle = -\sum_{k=0}^{n-1} a^k t^k + \langle f \rangle$. For (2) note that t acts with matrix $J_n(\lambda)$ if and only if $(t-\lambda)$ acts by $J_n(0)$, which is clear: as noted in the statement of the Lemma, in order to get an upper triangular, rather than a lower triangular, matrix, we need to order the basis by *decreasing* degree rather than increasing degree. □

Theorem 13.4. *(Rational Canonical Form.) Suppose that V is a nonzero finite dimensional k-vector space and $T\colon V \to V$ is a linear map. Then there are unique nonconstant monic polynomials $f_1, \ldots, f_k \in \mathsf{k}[t]$ such that $f_1 | f_2 | \ldots | f_k$ and a basis of V with respect to which T has matrix which is block diagonal with blocks $C(f_i)$:*

$$\begin{pmatrix} C(f_1) & 0 & \ldots & 0 \\ 0 & C(f_2) & 0 & \vdots \\ \vdots & 0 & \ddots & \vdots \\ 0 & \ldots & 0 & C(f_k) \end{pmatrix}$$

Proof. By the canonical form theorem and Lemma 13.1, there is an isomorphism $\theta\colon V \to \bigoplus_{i=1}^k \mathsf{k}[t]/\langle f_i \rangle$ of $\mathsf{k}[t]$-modules, where[59] $f_1 | f_2 | \ldots | f_k$ and the f_i are monic nonunits (hence nonconstant polynomials). The f_i are unique (rather than unique up to units) since we insist they are monic. Now the direct sum $\bigoplus_{i=1}^k \mathsf{k}[t]/\langle f_i \rangle$ has a basis B given by the union of the bases in Lemma 13.3, and the preimage $\theta^{-1}(B)$ is thus a basis of V. The matrix of T with respect to this basis is thus the same as the matrix of the action of t on the direct sum $\bigoplus_{i=1}^k \mathsf{k}[t]/\langle f_i \rangle$, and again by Lemma 13.3 this is clearly block diagonal with blocks $C(f_i)$ $(1 \leq i \leq k)$ as required.

□

This matrix form for a linear map given by the previous theorem is known as the *Rational Canonical Form* of T. Notice that this form, unlike the Jordan canonical form, makes sense for a linear map on a vector space over *any* field, not just an algebraically closed field like $\mathbb{C}$.

We can also recover the Jordan canonical form for linear maps of $\mathbb{C}$-vector spaces from the second, primary decomposition, version of our structure theorem, which expresses each module in terms of cyclic modules $\mathsf{k}[t]/\langle f^k \rangle$ where f is irreducible. The monic irreducibles over $\mathbb{C}$ are exactly the polynomials $t - \lambda$ for $\lambda \in \mathbb{C}$. Thus the second structure theorem tells us that, for V a finite dimensional complex vector space and $T\colon V \to V$, we may write $V = V_1 \oplus V_2 \oplus \ldots \oplus V_k$ where each V_i isomorphic to $\mathbb{C}[t]/\langle (t-\lambda)^r \rangle$ for some $\lambda \in \mathbb{C}$, and $r \in \mathbb{N}$. The Jordan canonical form now follows exactly as in the proof of the rational canonical form, replacing the use of the

[59] Note that $k \neq 0$ since we are assuming that V is not $\{0\}$.

canonical form for modules with the primary decomposition, and the use of part (1) of Lemma 13.3 with part (2) of the same Lemma.

13.1. **Remark on computing rational canonical form.** It is also worth considering how one can explicitly compute the decomposition that yields the rational canonical form: our proof of the existence of the canonical form is constructive, so if we can find a presentation of the $\mathsf{k}[t]$ module given by a linear map acting on a k-vector space V then we should be able to compute. The following proposition shows one way to do this.

Proposition 13.5. *Let V be an n-dimensional k-vector space and $\phi\colon V \to V$ a linear map. If ϕ has matrix $A \in Mat_n(\mathsf{k})$ with respect to a basis $\{e_1, \dots, e_n\}$ of V, then the $\mathsf{k}[t]$-module corresponding to (V, ϕ) has a presentation*

$$\begin{tikzcd} \mathsf{k}[t]^n \arrow[r, "r"] & \mathsf{k}[t]^n \arrow[r, "f"] & V \end{tikzcd}$$

where the homomorphism r between the free $\mathsf{k}[t]$-modules is given by the matrix $tI_n - A \in Mat_n(\mathsf{k}[t])$, and the map from $f\colon \mathsf{k}[t]^n \to V$ is given by $(f_1, \dots, f_n) \mapsto \sum_{i=1}^n f_i(A)(e_i)$.

Proof. Sketch: Since t acts by ϕ on V, and ϕ has matrix A, it follows that the image N of the map r lies in the kernel of the f. It thus suffices to check that this map is injective and its image is the whole kernel. To see that it is the whole kernel, let $F = \mathsf{k}^n \subset \mathsf{k}[t]^n$ be the copy of k^n embedded as the degree zero polynomials. It follows immediately from the definitions that f restricts to an k-linear isomorphism from F to V, and thus it is enough to show that $N + F = \mathsf{k}[t]^n$ and $N \cap F = \{0\}$ (where the former is the *vector space* sum). Both of these statements can be checked directly: the intersection is zero because f restricts to an isomorphism on F and $N \subseteq \ker(f)$. The sum $N + F$ must be all of $\mathsf{k}[t]^n$ since it is easy to check that it is a submodule and it contains F which is a generating set. Finally, since the quotient $\mathsf{k}[t]^n/N$ is torsion, N must have rank n and hence r does not have a kernel (since the kernel would have to be free of positive rank, and hence the image would have rank less than n.) □

Example 13.6. Let A be the matrix

$$A = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & -2 & -3 \end{pmatrix}$$

Then we have

$$tI_3 - A = \begin{pmatrix} t & -1 & 0 \\ 0 & t & -1 \\ 1 & 2 & 3+t \end{pmatrix} \sim \begin{pmatrix} 1 & 2 & 3+t \\ 0 & t & -1 \\ t & -1 & 0 \end{pmatrix} \sim \begin{pmatrix} 1 & 0 & 0 \\ 0 & t & -1 \\ 0 & -1-2t & -3t-t^2 \end{pmatrix}$$

$$\sim \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & t \\ 0 & -3t-t^2 & -1-2t \end{pmatrix} \sim \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -t^3-3t^2-2t-1 \end{pmatrix}$$

so that $(\mathbb{Q}^3, A)$ is isomorphic as a $\mathbb{Q}[t]$-module to $\mathbb{Q}[t]/\langle g \rangle$ where $g = t^3 + 3t^2 + 2t + 1$.

14. Appendix A: Polynomial rings and convolution.

In this appendix we discuss in somewhat more detail the construction of polynomials rings with coefficients in an arbitrary ring, and point out how the construction generalizes in a number of interesting ways.

Consider the set $\mathbb{C}[t]$ of polynomials with complex coefficients. This is a ring with the "obvious" addition and multiplication: if p, q are polynomials, then $p + q$ and $p.q$ are the polynomials given by pointwise addition and multiplication – that is, $p.q(z) = p(z).q(z)$ for all $z \in \mathbb{C}$ and similarly for addition. In other words, we realise the ring of polynomials with complex coefficients as a subring of the ring of all functions from $\mathbb{C}$ to itself. To check that polynomials do indeed form a subring, we need to check (amongst other things[60]) that $p.q$ is a polynomial if p and q are. But let $p = \sum_{k=0}^{N} a_k t^k$ and $q = \sum_{k=0}^{M} b_k t^k$. Then

$$\begin{aligned} p.q &= \left(\sum_{k=0}^{N} a_k t^k\right)\left(\sum_{l=0}^{M} b_l t^l\right) = \sum_{k,l} a_k b_k t^{k+l} \\ &= \sum_{n=0}^{N+M} \Big(\sum_{k+l=n} a_k b_k \Big) t^n, \end{aligned} \tag{14.1}$$

where we take $a_k = 0$ for $k > N$ and $b_l = 0$ for $l > M$, and the second line is evidently a polynomial function as required.

However, if we want to consider polynomials with coefficients in an arbitrary ring, we encounter the problem that a polynomial will *not* determined by its values on elements of the ring: for example if $R = \mathbb{Z}/2\mathbb{Z}$, then since R has two elements there are only four functions from R to itself in total, but we want two polynomials to be equal only if all their coefficients are the same and so we want infinitely many polynomials even when our coefficient ring is finite. Indeed, for example, we want $1, t, t^2, \ldots$ to all be distinct as polynomials, but as functions on $\mathbb{Z}/2\mathbb{Z}$ they are all equal!

The solution is much like what we do when we construct complex numbers – there we simply define a new multiplication on $\mathbb{R}^2$ and check it, along with vector addition, satisfy the axioms for a field. We start by viewing a polynomial as its sequence of coefficients, and define what we want the addition and multiplication to be, and again just check that the ring axioms are satisfied. This approach will also give us a new ring, called the ring of formal power series, simply by allowing all sequences in R, not just ones which are zero for large enough $n \in \mathbb{N}$.

Definition 14.1. Let R be a ring, and define $R[[t]] = \{a\colon \mathbb{N} \to R\}$ the set of sequences taking values in R. We define binary operations as follows: for sequences[61]

[60] Though closure under product is probably the most substantial thing to check.

[61] As is standard enough for sequences, we write a_n rather than $a(n)$ for the values of the sequence $a\colon \mathbb{N} \to R$.

$(a_n), (b_n) \in R[[t]]$ let

$$(a_n) + (b_n) = (a_n + b_n); \quad (a_n) \star (b_n) = (c_n), \text{ where } c_n = \sum_{k+l=n} a_k b_l$$

Thus the addition just comes from pointwise addition of functions from $\mathbb{N}$ to R, but the multiplication comes from formula we got in Equation (14.1).

It is immediate that the sequence with all terms equal to $0 \in R$ is an additive identity in $R[[t]]$, while the identity for our multiplication operation $\star$ is the sequence $\mathbf{1} = (1, 0, \ldots)$, that is $\mathbf{1}_n = 1$ if $n = 0$ and $\mathbf{1}_n = 0$ if $n > 0$. The fact that $\star$ distributes over addition is also straightforward to check, while the associativity of $\star$ follows because

$$((a_n) \star (b_n)) \star (c_n) = (\sum_{k+l+p=n} (a_k b_l) c_p) = (\sum_{k+l+p=n} a_k (b_l c_p)) = (a_n) \star ((b_n) \star (c_n))$$

Definition 14.2. Now let $R[t]$ be the subset of $R[[t]]$ consisting of sequences (a_n) such that there is some $N \in \mathbb{N}$ for which $a_n = 0$ for all $n > N$. To check this is a subring, notice that if $(a_n), (b_n) \in R[t]$ and $a_n = 0$ for all $n > N$ and $b_n = 0$ for all $n > M$, then the sequence $(c_n) = (a_n) \star (b_n)$ is zero for all $n > N + M$: indeed if $k + l = n > N + M$ we cannot have both $k \leq N$ and $l \leq M$ and so the product $a_k b_l$ will be zero, and hence $c_n = \sum_{k+l=n} a_k b_l = 0$ for all $n \geq N + M$.

Finally we want to relate our construction to the notation we are used to for polynomials. Let $t \in R[t]$ be the sequence $(0, 1, 0, \ldots)$, that is $t_n = 1$ for $n = 1$ and $t_n = 0$ for all other $n \in \mathbb{N}$. Then it is easy to check by induction that $t^k = t \star t \star \ldots \star t$ (k times) has $t^k_n = 1$ if $n = k$ and $t^k_n = 0$ for all other $n \in \mathbb{N}$. It follows that if (a_n) is a sequence in $R[t]$ for which $a_n = 0$ for all $n > N$ then $(a_n) = \sum_{k=0}^{N} a_k t^k$. Note that if (a_n) is any element of $R[[t]]$ it is the case that $(a_n) = \sum_{k \in \mathbb{N}} a_k t^k$, where the right-hand side gives a well-defined sequence in R despite the infinite sum, because for any integer k only finitely many (in fact exactly one) of the terms in the infinite sum are non-zero. This is why the ring $R[[t]]$ is known as the ring of *formal power series*.

Remark 14.3. This definition also allows us to define polynomial rings with many variables: given a ring R let $R[t_1, \ldots, t_k]$ be defined inductively by $R[t_1, \ldots, t_{k+1}] = (R[t_1, \ldots, t_k])[t_{k+1}]$. Thus for example, $R[t_1, t_2]$ is the ring of polynomials in t_2 with coefficients in $R[t_1]$.

Remark 14.4. The problem that a polynomial $p \in R[t]$ is not determined by the function it gives on the ring R can be resolved: recall the Evaluation Lemma which says that if S is a ring and we are given a ring homomorphism $i \colon R \to S$ and an element $s \in S$, then there is a unique ring homomorphism $\theta_s \colon R[t] \to S$ which restricts to i on R and has $\theta_s(t) = s$. This allows us to produce, for every ring homomorphism $i \colon R \to S$ and polynomial $p \in R[t]$ a function $p_S : S \to S$: simply take $p_S(s) = \theta_s(p)$. In other words, given any homomorphism of rings $i \colon R \to S$ we can evaluate a polynomial in $R[t]$ on the elements of S, so the polynomial gives not just a function on R but a function on any ring we can relate R to. In particular, if R is a subring of S, the it makes sense to evaluate $p \in R[t]$ on every element of S.

The collection of all the functions we can associate to a polynomial in this way *does* completely determine the polynomial.

14.1. **Convolution algebras.** The key to defining polynomial rings was the multiplication formula given by (14.1). This uses the fact that $\mathbb{N}$ has an addition operation. We can generalise this to give another interesting and important construction of a ring.

Definition 14.5. Let G be a group and let R be a ring. If $f\colon G \to R$ is a function, we let $\text{supp}(f) = \{x \in G : f(x) \neq 0\}$. Let $R[G]$ be the set of R-valued functions on G which have finite support, that is functions f for which $\text{supp}(f)$ is a finite set. We define $\star$ to be the binary operation

$$(f \star g)(x) = \sum_{y_1 y_2 = x} f(y_1)g(y_2). \tag{14.2}$$

If $S_1 = \text{supp}(f)$ and $S_2 = \text{supp}(g)$, then the terms in the sum on the right-hand side are zero unless $(y_1, y_2) \in S_1 \times S_2$, hence this sum is finite since S_1 and S_2 are. Moreover, $\text{supp}(f \star g)$ is a subset of $\{xy \in G : x \in S_1, y \in S_2\}$, which is also clearly finite, thus $\star$ is indeed a binary operation. It is associative because

$$((f \star g) \star h)(x) = (f \star (g \star h))(x) = \sum_{y_1 y_2 y_3 = x} f(y_1)g(y_2)h(y_3).$$

Just as for polynomials, it is straight-forward to check that $(R[G], +, \star, 0, \delta_e)$ is a ring, where $+$ is pointwise addition, 0 is the function on G which is 0 on every element of G, and δ_e is the indication fucntion of the identity element e of G, that is $\delta_e(x) = 1$ if $x = e$, $\delta_e(x) = 0$ otherwise. This ring is known as the *group algebra* of G with coefficients in R. This ring (when $R = \mathbb{C}$) will be very important in the study of representations of finite groups in the Part B representation theory course.

Example 14.6. Let $G = \mathbb{Z}$. Show that the ring $R[\mathbb{Z}]$ is just the ring of Laurent polynomials $R[t, t^{-1}]$.

Remark 14.7. *i*) The natural number $\mathbb{N}$ are of course not a group under addition, but the convolution product still makes sense. This is because we only ever use the associativity of the product and the existence of an identity element in constructing the ring $R[G]$. Thus the construction actually works for any *monoid* $(\Gamma, \times, e)$, that is, a set Γ with an associative binary operation $\times$ and an identity element $e \in \Gamma$ for the binary operation.

ii) Some books, especially those on representations of finite groups, present the group algebra slightly differently: they define $R[G]$ to be the set of formal R-linear combinations on the elements of the group, and then define multiplication by extending the group product "linearly", that is the elements of $R[G]$ are of the form $\sum_{g \in G} a_g.g$ where $a_g \in R$ and the product is given by

$$(\sum_{g \in G} a_g.g)(\sum_{h \in G} b_h.h) = \sum_{g,h \in G} a_g b_h (g.h) = \sum_{x \in G}(\sum_{gh = x} a_g b_h).x$$

> This is readily seen to be isomorphic to the definition above via the map which sends a group element g to the function e_g which takes the value 1 on g and 0 on all other elements of G. The function-based approach is more important when studying infinite groups with additional structure (such as being a metric space say) when you can consider subrings of the ring of all functions, such as continuous functions.

Remark 14.8. The restriction on the support of the functions in $R[G]$ ensures that the formula (14.2) gives a well-defined operation. Products given by this formula are called *convolution products* and come up in many parts of mathematics. Note that the formula is sometimes written less symmetrically as:

$$(f \star g)(x) = \sum_{y \in G} f(xy^{-1})g(y),$$

If the group G is infinite, for example if $G = \mathbb{R}$, then instead of summing over elements of the group one can integrate (imposing some condition on functions which ensures the integral makes sense and is finite) and the convolution formula becomes:

$$(f \star g)(x) = \int_{\mathbb{R}} f(x-y)g(y)dy,$$

which may be familiar from the study of integral transforms.

15. Appendix B: Unique Factorization for $\mathbb{Z}$.

In this Appendix we establish unique factorization for the ring $\mathbb{Z}$. The strategy of proof will be what motivates the definition of a Euclidean Domain, so if you have a good understanding of the material in this note, it should make the part of the course on EDs, PIDs and UFDs easier to grasp.

Theorem 15.1. *If $n \in \mathbb{Z}\backslash\{0, \pm 1\}$, then we may write $n = p_1 \dots p_k$ where the p_i are primes and the factorization is unique up to sign and reordering of the factors.*

Remark 15.2.

i) Most of the work we will need to do to prove the theorem will be to understand what the right notion of a "prime" is. Once we establish that, the proof of unique factorization will be quite straight forward.

ii) We work with all integers, not just positive ones, so we will have positive and negative primes numbers.

iii) The uniqueness statement is slightly cumbersome to say, but in essence it says the factorization is as unique as it can possibly be: for example if $n = 6$ then we can write

$$n = 2.3 = 3.2 = (-2).(-3) = (-3).(-2),$$

and while each of these are prime factorizations they are clearly all "essentially the same". The ambiguity of signs would be removed if we insisted that n and all the primes were positive, but it is more natural to ask for a statement which holds for any element of $\mathbb{Z}$.

15.1. **Highest common factors.**

Let's begin with some (hopefully familiar enough) terminology:

Definition 15.3. If $n \in \mathbb{Z}$ we say that *a divides n*, or a is a *factor* of n is there is an integer b such that $n = a.b$. We will use the notation $a \mid n$ to denote the fact that a divides n.

Remark 15.4. Notice that if $a \mid b$ then any multiple of b is also a multiple of a and so $b\mathbb{Z} \subseteq a\mathbb{Z}$. Thus divisibility corresponds to containment of ideals. In particular, since $\{0\} \subseteq n\mathbb{Z}$, every integer divides 0 while 0 only divides itself.

The most basic observation about the integers and division is that, given *any* pair of integers a, b, provided b is not zero, we can do "division by b with remainders". The next lemma makes this precise. Since we are working with all integers rather than positive integers, it is convenient to use the absolute value function $|.|\colon \mathbb{Z} \to \mathbb{Z}_{\geq 0}$.

Lemma 15.5. *(Division algorithm.) For any $a, b \in \mathbb{Z}$ such that $b \neq 0$, there exist integers q and r such that*

i) $a = q.b + r$.

ii) $|r| < |b|$;

Proof. Note that if $a = q.b + r$ then $-a = (-q).b + (-r)$, and $|r| = |-r|$ so that the pair $(-q, -r)$ satisfy the conditions we require for the integers $(-a, b)$. Thus it is enough to prove the Lemma for $a \geq 0$. Similarly if $a = q.b + r$ then $a = (-q).(-b) + r$, so it is enough to prove the Lemma in the case $b > 0$ also. We will prove this by induction on a. If $a < b$ then the result is clear as we may take $q = 0$ and $r = a$. If $a \geq b$ then consider the set $S = \{q.b : q \in \mathbb{N}, q.b \leq a\}$. Since S is clearly finite (it is for example contained in the set $\{1, 2, \ldots, a\}$) it has a maximal element. Set q to be this element. Then $qb \leq a < (q+1)b$, and so if $r = a - q.b$ it follows $0 \leq r < b$, and we are done. □

Remark 15.6. Note that if we work with nonnegative integers a, b and insist that q and r are nonnegative also then, for given a, b the integers q and r are unique, but if we work with $\mathbb{Z}$ then they are not: Indeed $qb + r = (q + 1).b + (r - b)$, and if $0 < r < b$ then $-b < b - r < 0$, so $(q + 1, r - b)$ is an alternative solution. Concretely, if $(a, b) = (10, 7)$ say, then $10 = 1.7 + 3 = 2.7 - 4$.

Notice also that while it makes sense to say m divides n for m, n elements of any ring R, condition *ii*) of the Division Algorithm uses the absolute value function and the ordering on positive integers, thus it won't make sense for an arbitrary ring. (It will, however, motivate the definition of a class of rings called "Euclidean Domains".)

The first step in understanding how factorization works in $\mathbb{Z}$ is to understand the notion of a highest common factor. The crucial point is that the right condition for a common factor to be the "highest" is not just to ask for the largest of the common factors in the usual sense:

Definition 15.7. Let $n, m \in \mathbb{Z}$. We say that d is a *common factor* of m, n if $d \mid m$ and $d \mid n$. We say that c is the *highest common factor* if it is a common factor and whenever $d \in \mathbb{Z}$ is any other commmon factor then $d|c$. We will write $h.c.f(m, n)$ for a highest common factor of m and n.

The only downside of this definition is that it is not immediately clear that *h.c.f*s always exist! On the other hand, it does follow from the definition that the highest common factor, if it exists, is unique up to sign: If c_1 and c_2 are highest common factors, then because c_1 is a common factor and c_2 is a highest common factor we must have $c_1|c_2$, but symmetrically we also see that $c_2|c_1$. It is easy to see from this that $c_1 = \pm c_2$, and so if we require highest common factors to be non-negative, they are unique. (Indeed the argument essentially repeats the proof we saw in lectures that in an integral domain, the generators of a principal ideal are all associates.)

The existence of highest common factors relies on the division algorithm, as we now show. The argument also proves that the ideals in $\mathbb{Z}$ are exactly the principal ideals $n\mathbb{Z}$, so since that is of independent interest, we establish this first.

Lemma 15.8. *i*) *Let I be an ideal of $\mathbb{Z}$. Then there is an $n \in \mathbb{Z}$ such that $I = n\mathbb{Z}$, that is, I is principal.*

ii) *Let $m, n \in \mathbb{Z}$. The highest common factor $h.c.f(m, n)$ exists and moreover there are integers $r, s \in \mathbb{Z}$ such that $h.c.f(m, n) = am + bn$.*

Proof. For the first part, if $I = \{0\}$ then clearly I is the ideal generated by $0 \in \mathbb{Z}$ and we are done. If $I \neq \{0\}$, then the set $\{|k| : k \in I\backslash\{0\}\}$ is nonempty, and so we may take $n \in I$ with $|n|$ minimal among nonzero elements of I. But now if $a \in I$ is any element, we may write $a = qn + r$ for some $q, r \in \mathbb{Z}$ with $|r| < |n|$. But $r = a - q.n \in I$, so by the minimality of $|n|$ we must have $r = 0$ and so $a = qn$. It follows that $I \subseteq n\mathbb{Z}$. But since $n \in I$ we have by definition that $n\mathbb{Z}$ (the ideal generated by n) must lie in I, hence $I = n\mathbb{Z}$ as required.

For the second part, note that if m, n are integers then

$$I = n\mathbb{Z} + m\mathbb{Z} = \{r.n + s.m : r, s \in \mathbb{Z}\},$$

is the ideal generated by $\{m, n\}$. By the first part, I must be principal, and hence there is some $k \in \mathbb{Z}$ such that $I = k\mathbb{Z}$. But then since $n, m \in I$ we must have $k \mid n$ and $k \mid m$ so that k is a common factor. On the other hand, since $k \in I$, it follows immediately from the definition of I that there are integers a, b such that $k = am + sb$. Now if d is any common factor of m and n, then it is clear that d divides any integer of the form $r.m + s.n$, and so d divides every element of I and hence $d \mid k$. The second part of the Lemma follows immediately. □

Remark 15.9. The second part of the above Lemma is usually known as *Bézout's Lemma*. Its proof has the advantage that it can actually be made constructive. (This is not needed for the rest of this note, but is something you saw before in Constructive Mathematics.)

Suppose that m, n are integers and $0 < n < m$. Euclid's Algorithm gives a way to compute $h.c.f(n, m)$. Let $n_0 = m, n_1 = n$, and if $n_0 > n_1 > \ldots > n_k > 0$ have been defined, define n_{k+1} by setting $n_{k-1} = q_k n_k + n_{k+1}$ where $0 \leq n_{k+1} < n_k$ (since we are insisting everything is positive, the division algorithm ensures this uniquely defines the integers q_k and n_{k+1}). Clearly this process must terminate with $n_{l-1} > n_l = 0$ for some $l > 0$.

Lemma 15.10. *The integer n_{l-1} is the highest common factor of the pair (m, n)*

Proof. The equation $n_{k-1} = q_k n_k + n_{k+1}$ shows that any common factor of the pair (n_{k-1}, n_k) is also a common factor of the pair (n_k, n_{k+1}). Thus

$$h.c.f(m, n) = h.c.f(n_0, n_1) = h.c.f(n_1, n_2) = \ldots = h.c.f(n_{l-1}, 0) = n_{l-1}.$$

□

15.2. **Characterising prime numbers.** We are almost ready to prove unique factorization now. The last ingredient that we need is a better understanding of the properties of prime numbers.

Definition 15.11. An integer $n \in \mathbb{Z}$ is said to be *irreducible* if $n \notin \{\pm 1\}$ and its only factors are $\{\pm 1, \pm n\}$, that is, if $n = a.b$ for some $a, b \in \mathbb{Z}$ then either $a = \pm 1$ or $b = \pm 1$.

The notion of an irreducible integer is what people normally call a "prime", but there is another characterization of prime integers which is the key to the proof of unique factorization, and we reserve the term "prime" for this characterization. (For rings other than $\mathbb{Z}$, the two notions are *not* necesarily the same.)

Definition 15.12. Let $n \in \mathbb{Z}$. Then we say n is *prime* if $n \notin \{\pm 1\}$ and whenever $n \mid a.b$, either $n \mid a$ or $n \mid b$ (or both). (Using the terminology for ideals which we have developed, n is prime whenever $n\mathbb{Z}$ is a prime ideal in $\mathbb{Z}$).

Remark 15.13. Note that it follows easily by induction that if p is a prime number and $p \mid a_1 \dots a_k$ for $a_i \in \mathbb{Z}$, $(1 \leq i \leq k)$, then there is some i with $p \mid a_i$.

We now want to show that the irreducible and prime integers are the same thing. This is a consequence of Bézout's Lemma, as we now show:

Lemma 15.14. *If $n \in \mathbb{Z}\backslash\{0\}$ then n is irreducible if and only if n is prime.*

Proof. Suppose that n is a nonzero prime and write $n = a.b$ for integers a, b. Then clearly $n \mid a.b$ so by definition we must have $n \mid a$ or $n \mid b$. By symmetry we may assume that $n \mid a$. Then $a = n.p$ for some integer p and so $n = (np)b = n(pb)$ and hence $n(1 - pb) = 0$. Since n is nonzero, it follows that $p.b = 1$ so that $b = \pm 1$, and thus n is irreducible.

Conversely, suppose that n is irreducible. Then if $n \mid a.b$, suppose that n does not divide a. Then by irreducibility, we must have $h.c.f(a, n) = 1$, and so by part *ii*) of Lemma 15.8 (Bézout's Lemma) we may write $1 = ra + sn$ for some $r, s \in \mathbb{Z}$. But then $b = r(a.b) + n(sb)$, and hence n divides b as required. □

15.3. **Unique factorization.** We are now ready to prove unique factorization for $\mathbb{Z}$.

Theorem 15.15. *Any integer $n \in \mathbb{Z}\backslash\{0, \pm 1\}$ can be written as a product $n = p_1 p_2 \dots p_k$ of primes, uniquely up to reordering and sign.*

Proof. We first show that any such integer n is a product of primes using induction on $|n|$. Since $n \notin \{0, \pm 1\}$, the smallest value for $|n|$ is 2, and in that case $n = \pm 2$ and so n itself is prime. If $|n| > 2$, then there are two cases: either n is prime, in which we are done, or it can be written as a product $n = a.b$ where $|a| > 0$ and $|b| > 0$. But then certainly $|a|, |b| < |n|$, so by induction we can write $a = r_1 \dots r_p$ and $b = s_1 \dots s_q$ where the r_i and s_j are primes. Thus we see that

$$n = a.b = r_1 \dots r_p s_1 \dots s_q,$$

showing that n is a product of primes.

Next we must show the product is unique. For this we again use induction, but now on the number of prime factors of n. Since we don't know uniqueness yet, we have to be slightly careful here: it might be that n can be written a product of primes in two different ways where the number of factors is different in the two factorizations. Thus we define, for $n \in \mathbb{Z}\backslash\{0, \pm 1\}$, the number $P(n)$ to be the *minimum* number of prime factors occuring in a prime factorization of n, and use induction on $P(n)$. If $P(n) = 1$ then n is prime. But then the only factors of n are $\{\pm 1, \pm n\}$ and so n can be written uniquely as a product of one prime (itself) (or as $(-1).(-n)$).

Now if $P(n) = k > 1$, assume uniqueness holds for any $m \in \mathbb{Z}$ with $P(m) < k$. Write $n = p_1 \dots p_k$ and suppose that we have another prime factorization $n = q_1 \dots q_t$ (where by definition $t \geq k$). Now p_1 is prime and $p_1 \mid n = q_1 \dots q_t$, so as above we must have $p_1 \mid q_i$ for some i $(1 \leq i \leq t)$ and reordering the q_js we can assume that $i = 1$. Then since q_1 is prime we find $p_1 = \pm q_1$. Now if $m = p_2 \dots p_k$ we must have $P(m) \leq k - 1$, and so by induction unique factorization holds for $m = p_2 \dots p_k = (\pm q_2) \dots q_l$, and so it follows that the p_is and q_js for $i, j > 1$ are also equal up to signs and reordering. □

Remark 15.16. As mentioned before, the ambiguity about signs disappears if we only consider factorization for positive integers. If we were only interested in $\mathbb{Z}$ that might well be the best thing to do, but since we are interested in generalising to other rings where there may not be a convenient analogue of the notion of a positive integer, it is better to find a statement valid for all integers. Note that the reason signs appear in the uniqueness statement is because $\{\pm 1\}$ is the group of units in the ring $\mathbb{Z}$. Thus one could rephrase the statement by saying that the factorization is unique "up to reordering and units".

We finish with a brief discussion of a ring only slightly bigger than $\mathbb{Z}$ where the notion of an irreducible element and a prime element are different, and where unique factorization fails.

Example 15.17. Let $R = \mathbb{Z}[\sqrt{-17}] = \{a + b\sqrt{-17} : a, b \in \mathbb{Z}\}$. It is straight-forward to check that R is a subring of $\mathbb{C}$. The function $N \colon R \to \mathbb{Z}$ given by $N(a + b\sqrt{-17}) = a^2 + 17b^2$ is *multiplicative*, in that $N(z.w) = N(z).N(w)$ (indeed it is just the restriction to R of the function $z \mapsto |z|^2 = z\bar{z}$ on $\mathbb{C}$). Using this, one can show that $2, 3$, and $1 \pm \sqrt{-17}$ are all irreducible element of R (in the same sense as we used for ordinary integers). It then follows that none of these elements are prime – indeed it is easy to see that there is an element of R which is a product of both two irreducibles and three irreducibles.

16. Appendix C: A PID which is not a ED

In this appendix, following the article[62] of Cámploi which you can read through JSTOR, we outline a proof that the ring $R = \mathbb{Z}[\frac{1}{2}(1+\sqrt{-19})]$ is a PID but not a Euclidean domain. (*This is for curiosity only, it is not examinable.*)

For convenience we write $\theta = \frac{1}{2}(1+\sqrt{-19})$. We need to do two things, first to show that R is not a Euclidean domain, and then to show that nevertheless it is a PID. We will need the restriction of the square of the modulus function on complex numbers, which we will write as N, that is $N(a+ib) = a^2+b^2$ (and crucially $N(z.w) = N(z).N(w)$). It is easy to check that $\theta^2 = \theta - 5$, so that in particular $R = \{a+b\theta : a, b \in \mathbb{Z}\}$. Moreover, we have $N(a+b\theta) = a^2+ab+5b^2$, and so if $z \in R\backslash\{0\}$ then $a^2+ab+5b^2 = N(z)$ is a positive integer.

R is not a Euclidean Domain:

Using the function N it is not too hard to see that the units in R are exactly $\{\pm 1\}$, because if $r \in R$ is a unit $N(r)$ must be 1. But if $a^2+ab+5b^2 = 1$ and $a.b \geq 0$ then clearly we must have $b = 0$ and $a = \pm 1$. If $a.b \leq 0$ then writing $N(a+b\theta) = (a+b)^2 - ab + 4b^2$ we again see that $b = 0$, and so the only units are ± 1. Similar considerations using N allow you to show that $2, 3$ are irreducible elements in R. Using these facts we can show that R is not a Euclidean domain: Suppose for the sake of a contradiction that $d \colon R\backslash\{0\} \to \mathbb{N}$ is a Euclidean function on R. Then consider the minimal value k of d on the elements of $R\backslash\{0\}$ which are not units, and pick some $m \in R$ with $d(m) = k$. If we divide m into 2 we see that we have $2 = q.m + r$, where $r = 0$ or $d(r) < d(m)$, and hence either m divides 2 or r is a unit. Since 2 is irreducible this would imply either $m = \pm 2$ (since the only units in R are ± 1) or $2 = q.m \pm 1$, which since m is not a unit, forces $q.m = 3$ and hence again since 3 is irreducible $m = \pm 3$. Thus we see that m is ± 2 or ± 3. But now consider dividing m into θ, say $\theta = q.m + r$. Certainly θ cannot be divisible by 2 or 3, but then as before we must have $r = \pm 1$, which would imply $\theta \pm 1 = \pm 2q$ or $\pm 3q$ for some $q \in R$, which is clearly impossible. It follows that R is not a Euclidean Domain.

R is a PID:

To see that R is a PID the key is to show that it is not too far from being a Euclidean Domain. Recall that the proof that a Euclidean Domain is a PID takes a element n of minimal Euclidean norm in an ideal I and shows this must be a generator because if $m \in I$ then $m = q.n + r$ where $d(r) < d(n)$ or $r = 0$, and $r = m - q.n \in I$ forces $r = 0$. This argument works for r any linear combination of m and n, so we can prove a ring is a PID if we can find a function d which satisfies the following weaker version of the condition for a Euclidean norm: Say that $d \colon R\backslash\{0\} \to \mathbb{N}$ is a *weak norm* if given any $n \in R\backslash\{0\}$ and $m \in R$ there exist $\alpha, \beta \in R$ such that $d(\alpha.m + \beta.n) < d(n)$. If a ring has a weak norm then the above argument shows it is a PID. Hence to see that $R = \mathbb{Z}[\theta]$ is a PID, it is enough to check that $N(a+b\theta) = a^2+ab+5b^2$, the squared modulus, is a weak norm. The proof is similar but somewhat more involved to how

[62] "A Principal Ideal Domain that is not a Euclidean Domain", Oscar A. Cámpoli, American Mathematical Monthly, vol. 95, no. 9, 868-871.

one shows that $\mathbb{Z}[i]$ are a Euclidean Domain with the same function. Let $m, n \in R$, and consider $m/n \in \mathbb{C}$. We want to find $\alpha, \beta \in R$ so that $N(\beta(m/n) - \alpha) < 1$, so that $N(\beta.n - \alpha.n) < N(n)$. But it is easy to check that any ratio of elements of R lies in $\{a + b\theta : a, b \in \mathbb{Q}\}$ (just clear denominators using the complex conjugate). Hence for any such m/n we can subtract from it an element of R to ensure that the imaginary part of the result lies between $\pm\sqrt{19}/4$. Now[63] if the imaginary part of the result, q say, is less that $\sqrt{3}/2$ in modulus, then q will be within 1 or an element of $\mathbb{Z} \subset R$, so it suffices to consider the case $\sqrt{3}/2 < Im(q) < \sqrt{19}/4$ (the case when $Im(q) < 0$ being similar). But then $2q-\theta$ has imaginary part between $\sqrt{3} - \sqrt{19}/2$ and 0, which you can check is less that $\sqrt{3}/2$, and so we are done.

MATHEMATICAL INSTITUTE, OXFORD.

[63] This analysis follows Rob Wilson's note which you can read at www.maths.qmul.ac.uk/ raw/MTH5100/PIDnotED.pdf.